Pontificia Universidad JAVERIANA Colombia

PREDICCIÓN DE FRAUDES Y ANOMALÍAS EN EL SUMINISTRO DE AGUA POTABLE CON TÉCNICAS DE ANALÍTICA DE DATOS

MAESTRÍA EN CIENCIA DE DATOS

Cristian Fabián Rodríguez Rodríguez

Dany Alexander Enríquez Sánchez

José David Zamudio Rojas

Directora

Dra. Sandra Milena Ramírez Buelvas

Pontificia Universidad Javeriana (Cali, Colombia)

Codirector

Dr. Fredy Humberto Troncoso Espinosa

Universidad del Bío-Bío (Concepción, Chile)

Índice general

Resumen

El estudio desarrolló una metodología integral para la detección de fraudes y anomalías en el consumo de agua potable en Chile, utilizando una base de datos anonimizada con series de consumo mensuales, características metrológicas y localización geográfica de medidores mecánicos. En la detección de fraude sin enfoque de negocio, los modelos supervisados *Random Forest* (RF) y *XGBoost* (XGB) mostraron desempeños técnicos similares, con diferencias leves en métricas como el F1-Score. Al incorporar el enfoque de negocio mediante matrices de costos y métricas económicas, las métricas estadísticas se mantuvieron estables, pero el *Random Forest* resultó con mejor desempeño operativamente, alcanzando una ganancia promedio cercana a 2,31 millones de dólares y un mROI del 64,3 %, superior al 59,9 % obtenido por *XGBoost*. En el análisis de anomalías técnicas, los modelos supervisados también se evaluaron, pero el desempeño fue limitado debido a la baja frecuencia de muestreo, mientras que el enfoque no supervisado basado en DBSCAN no logró separar de forma efectiva los casos de interés de fraude y anomalías. Adicional, se empleó la regresión logística para el problema de fraude como modelo base interpretable: permitió identificar el efecto de variables de consumo, clase metrológica, año de instalación y zona geográfica, alcanzó un AUC–ROC de 0,78 y un *Brier Score* de 0,18, aunque la prueba de Hosmer–Lemeshow ($p \approx 0$) indicó un ajuste limitado frente a la complejidad del fenómeno. Por su parte, en el caso de anomalías el modelo de regresión logística no cumplió los supuestos fundamentales de linealidad en el logit ni de adecuación del ajuste, por lo que no resultó apropiado para predecir anomalías y fue descartado como alternativa metodológica en este componente. A su vez, enfoque no supervisado basado en DBSCAN tampoco logró separar de forma efectiva los casos de interés, presentando sensibilidades cercanas a cero para la detección de fraude y utilidad acotada para anomalías. Finalmente, pese a las limitaciones inherentes del muestreo mensual y el desbalance extremo, los resultados obtenidos muestran que es posible construir modelos predictivos útiles para la priorización de inspecciones y el fortalecimiento de estrategias de control de fraude.

Lista de figuras

Lista de tablas

FICHA RESUMEN

Título: PREDICCIÓN DE FRAUDES Y ANOMALÍAS EN EL SUMINISTRO DE AGUA POTABLE CON TÉCNICAS DE ANALÍTICA DE DATOS.

1. **Área de trabajo:** Gestión de recursos hídricos y servicios públicos.
2. **Tipo de proyecto:** Aplicado.
3. **Estudiantes y contacto:**

   - **Cristian Fabián Rodríguez Rodríguez**
     cristianrodriguez@javerianacali.edu.co
     +57 3212659968.
     Bogotá D.C.
   - **Dany Alexander Enríquez Sánchez**
     danyenriquez@javerianacali.edu.co
     +57 315 2203135.
     Popayán, Cauca.
   - **José David Zamudio Rojas**
     josezamundio@javerianacali.edu.co
     +57 3144254589.
     Bogotá D.C

4. **Directora y codirector:**

   - **Directora:**
     Dra. Sandra Milena Ramírez Buelvas.
     Pontificia Universidad Javeriana (Cali, Colombia).
     Profesora Asistente | Departamento de Ciencias Naturales y Matemáticas.
     smramirez@javerianacali.edu.co
   - **Codirector:**
     Dr. Fredy Humberto Troncoso Espinosa.
     Universidad del Bío-Bío (Concepción, Chile).
     Profesor Asistente | Departamento de Ingeniería Industrial.
     ftroncos@ubiobio.cl

5. **Palabras clave:** anomalías en consumo de agua, fraude en el consumo de agua, gestión de recursos, machine learning, monitoreo de consumo, regresión logística, sostenibilidad.
6. **Fecha de inicio:** Enero del 2025.
7. **Duración estimada:** 11 meses.

8. **Resumen:** El estudio desarrolló una metodología integral para la detección de fraudes y anomalías en el consumo de agua potable en Chile, utilizando una base de datos anonimizada con series de consumo mensuales, características metrológicas y localización geográfica de medidores mecánicos. En la detección de fraude sin enfoque de negocio, los modelos supervisados *Random Forest* (RF) y *XGBoost* (XGB) mostraron desempeños técnicos similares, con diferencias leves en métricas como el F1-Score. Al incorporar el enfoque de negocio mediante matrices de costos y métricas económicas, las métricas estadísticas se mantuvieron estables, pero el *Random Forest* resultó con mejor desempeño operativamente, alcanzando una ganancia promedio cercana a 2,31 millones de dólares y un mROI del 64,3 %, superior al 59,9 % obtenido por *XGBoost*. En el análisis de anomalías técnicas, los modelos supervisados también se evaluaron, pero el desempeño fue limitado debido a la baja frecuencia de muestreo, mientras que el enfoque no supervisado basado en DBSCAN no logró separar de forma efectiva los casos de interés de fraude y anomalías. Adicional, se empleó la regresión logística para el problema de fraude como modelo base interpretable: permitió identificar el efecto de variables de consumo, clase metrológica, año de instalación y zona geográfica, alcanzó un AUC–ROC de 0,78 y un *Brier Score* de 0,18, aunque la prueba de Hosmer–Lemeshow ($p \approx 0$) indicó un ajuste limitado frente a la complejidad del fenómeno. Por su parte, en el caso de anomalías el modelo de regresión logística no cumplió los supuestos fundamentales de linealidad en el logit ni de adecuación del ajuste, por lo que no resultó apropiado para predecir anomalías y fue descartado como alternativa metodológica en este componente. A su vez, enfoque no supervisado basado en DBSCAN tampoco logró separar de forma efectiva los casos de interés, presentando sensibilidades cercanas a cero para la detección de fraude y utilidad acotada para anomalías. Finalmente, pese a las limitaciones inherentes del muestreo mensual y el desbalance extremo, los resultados obtenidos muestran que es posible construir modelos predictivos útiles para la priorización de inspecciones y el fortalecimiento de estrategias de control de fraude.

1 INTRODUCCIÓN

El agua es un recurso esencial para la vida, el desarrollo humano y la sostenibilidad ambiental. Su papel en la salud pública, la producción de alimentos, la energía y la industria la convierte en un elemento central para el bienestar y la prosperidad de las sociedades modernas. No obstante, la creciente presión sobre los recursos hídricos ha dado lugar a una crisis global que se manifiesta en diversas escalas: escasez, sobreexplotación, contaminación, desigualdad en el acceso y deficiencias en la gestión. Estos fenómenos son el resultado de múltiples factores interrelacionados, entre los que destacan el cambio climático, el crecimiento demográfico, la expansión urbana, el deterioro de los ecosistemas y las limitaciones tecnológicas para medir y controlar el uso del agua.

En este contexto, la gestión eficiente del recurso hídrico se ha convertido en un desafío prioritario para los gobiernos, las empresas prestadoras del servicio y la sociedad en general. La pérdida de agua no facturada, los fraudes en los sistemas de medición y las fallas técnicas en los medidores son problemas que impactan directamente la sostenibilidad económica y operativa de las empresas sanitarias, al tiempo que limitan la equidad en la distribución del recurso. En países como Chile, donde la disponibilidad de agua ha disminuido de manera sostenida en las últimas décadas, se han identificado pérdidas significativas que superan los márgenes aceptables para una gestión eficiente, lo que refuerza la necesidad de estrategias innovadoras de control y monitoreo [1].

Frente a esta problemática, la Ciencia de Datos ofrece un marco analítico y metodológico capaz de transformar grandes volúmenes de información en conocimiento útil para la toma de decisiones. Su aplicación en el sector hídrico permite aprovechar los datos de consumo y de operación registrados por las empresas para identificar patrones anómalos, predecir eventos de fraude o fallas, y optimizar los procesos de mantenimiento e inspección. De este modo, la analítica se convierte en una herramienta estratégica no solo para mejorar la eficiencia operativa, sino también para fortalecer la transparencia, la sostenibilidad y la confianza de los usuarios en la gestión del agua.

El presente trabajo propone una metodología integral basada en Ciencia de Datos para la detección y predicción de fraudes y anomalías en el consumo de agua potable. La propuesta se fundamenta en el análisis de datos históricos provenientes de medidores mecánicos, aplicando técnicas de aprendizaje automático y métodos estadísticos de detección de comportamientos inusuales. A través de este enfoque se busca diferenciar entre irregularidades derivadas de intervenciones humanas y aquellas asociadas a fallas técnicas o errores de medición, aportando así un soporte analítico para la toma de decisiones informadas en las empresas prestadoras del servicio.

La metodología desarrollada contempla distintas fases. En primer lugar, se realizó un proceso de preprocesamiento de datos, que incluyó la depuración, imputación de valores faltantes, georreferenciación y normalización de variables. Posteriormente, se llevó a cabo un análisis exploratorio y la extracción de características estadísticas y temporales que permitieron describir el comportamiento del consumo. A continuación, se implementaron modelos de aprendizaje supervisado, modelos no supervisados y regresiones logísticas, que facilitaron la detección de patrones atípicos diferentes en contextos. Finalmente, se incorporó un enfoque económico-operativo para evaluar la utilidad práctica de los modelos, relacionando su desempeño con la priorización de inspecciones y la eficiencia en la asignación de recursos.

El desarrollo de esta investigación representa un aporte tanto metodológico como aplicado. Desde el punto de vista técnico, contribuye a la adaptación de herramientas de analítica avanzada a un problema de interés público, demostrando la capacidad de la Ciencia de Datos para abordar desafíos complejos en sectores tradicionales. Desde el punto de vista operativo, ofrece una metodología que puede integrarse en los procesos de gestión de las empresas sanitarias, permitiendo reducir pérdidas, optimizar costos y mejorar la sostenibilidad del servicio.

Más allá del ámbito empresarial, el impacto de esta propuesta trasciende hacia dimensiones sociales y ambientales. En términos sociales, promueve una cultura de gestión transparente y basada en evidencia, fortaleciendo la relación entre las empresas prestadoras y los usuarios. En el plano ambiental, contribuye al uso racional del agua, a la reducción de desperdicios y al cumplimiento de los Objetivos de Desarrollo Sostenible relacionados con el acceso equitativo y sostenible al recurso. Finalmente, desde una perspectiva académica, este trabajo abre la posibilidad de seguir explorando la integración de la analítica de datos en la gestión hídrica, fomentando investigaciones futuras orientadas a la sostenibilidad, la eficiencia y la innovación tecnológica.

2 DEFINICIÓN DEL PROBLEMA

2.1. Planteamiento del problema

Los servicios públicos, incluyendo el suministro de agua potable, son esenciales para la calidad de vida. Su adecuada gestión y facturación resultan fundamentales para asegurar la eficiencia y sostenibilidad del sistema [2]. En Chile, el sistema de suministro de agua potable se organiza en dos sectores: urbano y rural, alcanzando casi el 100 % de cobertura. En las áreas urbanas, el servicio es gestionado por empresas privadas bajo la supervisión de la Superintendencia de Servicios Sanitarios (SISS), mientras que, en las zonas rurales, la distribución del agua está a cargo de Comités de Agua Potable Rural (APR) administrados por las propias comunidades [3].

Según la Asociación Nacional de Empresas de Servicios Sanitarios (Andess), al cierre de 2023 Chile registró uno de los niveles de consumo de agua más bajos de los últimos 25 años, con un consumo promedio mensual por cliente de 16 m^3 [4]. Aunque el consumo ha disminuido, las pérdidas de agua no facturada (ANF) continúan siendo un desafío estructural, representando aproximadamente un 33 % de diferencia entre el agua producida y la que se factura a los clientes [1]. Estas pérdidas se deben tanto a fugas e ineficiencias operativas como a irregularidades intencionadas en el consumo. En este trabajo definimos *fraude* como la manipulación deliberada del medidor o la conexión ilegal destinada a reducir la lectura facturada, y *anomalía* como una falla técnica o comportamiento atípico del sistema de medición que no responde a una acción intencionada. Las manipulaciones representan cerca del 25 % del agua no contabilizada en algunos estudios, cifra que excede el rango óptimo de ANF (15–25 %) para una operación eficiente [1].

La problemática del fraude y de las anomalías adquiere mayor gravedad en un país vulnerable al cambio climático y a la sequía prolongada: en 2023 se registraron descensos de entre el 30 % y el 90 % en niveles de precipitación respecto a los promedios históricos [5], lo que intensifica la presión sobre los recursos hídricos y exige una gestión más eficiente basada en evidencia. No obstante, pese a la magnitud del problema, las empresas sanitarias carecen en general de protocolos automatizados y sistemas predictivos consolidados para distinguir y detectar en tiempo real tanto las manipulaciones del medidor (fraude) como las fallas técnicas (anomalías) a partir de series de consumo y metadatos de medidores. Esta brecha dificulta la priorización de inspecciones y la optimización de recursos en un contexto de escasez creciente.

2.2. Formulación del problema

¿Cómo pueden los modelos de Ciencia de Datos, aplicados a series de consumo de agua y características físicas de los medidores, predecir fraudes y anomalías, considerando las limitaciones de acceso a la información de los usuarios y la necesidad de gestionar eficazmente los recursos en un contexto operativo?

Para abordar esta pregunta, se analizarán y responderán los siguientes interrogantes:

- ¿Qué modelos estadísticos o de Machine Learning presentan el mejor desempeño para predecir fraudes y mediciones anómalas en una empresa distribuidora de agua potable, considerando las series de consumo de agua y las características físicas y funcionales de los medidores?
- ¿Qué características son clave para predecir fraudes y mediciones anómalas en una empresa distribuidora de agua potable, considerando las series de consumo de agua, las características físicas y funcionales de los medidores, las limitaciones en el acceso a información de los usuarios y la necesidad de gestionar eficientemente los recursos en un contexto operativo?
- ¿Cuáles son y cómo se interpretan las predicciones de fraudes y mediciones anómalas en una muestra de medidores de la red de agua de una empresa distribuidora de agua potable?

3 OBJETIVOS DEL PROYECTO

3.1. Objetivo general

Desarrollar una metodología basada en Ciencia de Datos para predecir fraudes y mediciones anómalas, identificando las características relevantes de las series de consumo de agua y las características físicas y de funcionamiento de los medidores, teniendo en cuenta las limitaciones en el acceso a información de los usuarios y la necesidad de gestionar eficazmente los recursos de la empresa en un contexto operativo.

3.2. Objetivos específicos

- Desarrollar una metodología basada en modelos estadísticos o de Machine Learning para predecir fraudes y mediciones anómalas en una empresa distribuidora de agua potable, teniendo en cuenta las series de consumo de agua y las características físicas y de funcionamiento de los medidores.
- Determinar las variables clave de las series de consumo de agua y las características físicas y funcionales de los medidores para predecir fraudes y mediciones anómalas en una empresa distribuidora de agua potable, considerando las limitaciones de acceso a la información de los usuarios y la necesidad de gestionar eficientemente los recursos de la empresa en un contexto operativo.
- Evaluar las mediciones fraudulentas y anómalas en una muestra de medidores de la red de una empresa distribuidora de agua potable con la metodología propuesta.

4 MARCO DE REFERENCIA

4.1. MARCO TEÓRICO

4.1.1. SUMINISTRO DE AGUA POTABLE

El suministro de agua potable es un componente esencial para garantizar el acceso equitativo y sostenible a este recurso. Esta sección aborda aspectos clave como los medidores de agua y sus tipos, el análisis de comportamientos fraudulentos y anómalos en el consumo, el concepto de agua no facturada (ANF), el uso de tecnologías avanzadas para medición y monitoreo, y los marcos legales que regulan su gestión. Estos elementos son fundamentales para comprender y mejorar la eficiencia en la distribución de agua potable.

Medidores de agua

De acuerdo con la International Organization of Legal Metrology (OIML), un medidor de agua se define como un instrumento destinado a medir, memorizar y mostrar el volumen de agua potable que pasa a través del sensor de medición [6]. Los medidores de agua se clasifican en dos grupos:

- **Medidores mecánicos:** El medidor opera registrando el flujo de agua a través de una turbina o hélice. La medición se realiza al observar el impacto del agua sobre dicha turbina o hélice, lo que permite calcular el volumen de agua que fluye a través del sistema [7].
- **Medidores no mecánicos:** El medidor opera midiendo el flujo de agua mediante el uso de sensores electrónicos, ultrasónicos o magnéticos [7].

Teniendo en cuenta que la base de datos suministrada para este proyecto se centra en *medidores mecánicos*, es fundamental analizar su funcionamiento y método de registro del consumo de agua, ya que esto asegura la precisión y confiabilidad en la interpretación de los datos recopilados. El medidor mecánico está equipado con una ventanilla que muestra 7 dígitos: 5 en color negro, que representan los metros cúbicos consumidos, y 2 en color rojo, que indican las centenas y decenas de litros, respectivamente. Además, cuenta con dos agujas adicionales: una que marca los litros, señalada con 0.001, y otra que indica las décimas de litro, representada por 0.0001. Este diseño permite una lectura precisa del consumo de agua en diferentes unidades [8].

Figura 4.1: Medidor de agua mecánico marca CICASA [8].

Comportamientos normales o anormales en el consumo de agua

El análisis del consumo de agua de un usuario se fundamenta en una secuencia de observaciones registradas en momentos específicos y ordenados cronológicamente [9]. Estas observaciones, denominadas series temporales, pueden medirse en intervalos regulares o irregulares, y permiten identificar patrones de comportamiento, estacionalidades y variaciones en el uso del recurso. A partir de dichas series es posible derivar variables representativas que sirven como insumos para la elaboración de modelos destinados a la detección de comportamientos anómalos o fraudulentos.

En este contexto, el *fraude* en el consumo de agua se define como la manipulación intencionada del sistema de medición con el propósito de reducir el volumen registrado y, por ende, el valor facturado. Este fenómeno puede manifestarse a través de prácticas como la derivación ilegal de acometidas, la inversión del sentido de flujo, la alteración física del medidor o el uso de imanes que interfieren en el registro del caudal [10]. Los patrones asociados al fraude tienden a presentar comportamientos sistemáticos y persistentes en el tiempo, caracterizados por disminuciones abruptas e inusuales del consumo o secuencias prolongadas de valores nulos que no corresponden con el historial del usuario. En consecuencia, la detección de fraude requiere de metodologías capaces de identificar desviaciones estructurales y recurrentes, diferenciándolas de fluctuaciones naturales o eventuales [11].

Por su parte, las *anomalías* en los datos de consumo se entienden como desviaciones significativas respecto al patrón esperado de comportamiento, sin que necesariamente exista una causa deliberada. Estas pueden originarse por fallas técnicas en los medidores, errores de comunicación, fugas internas o variaciones transitorias en los hábitos de uso. A diferencia del fraude, las anomalías suelen presentarse de manera puntual o aislada, y su interpretación requiere distinguir entre causas técnicas, operativas o de comportamiento. Desde la perspectiva analítica, la detección de anomalías busca identificar registros atípicos que se alejan del patrón general de consumo, aportando información valiosa para la mejora de la calidad de los datos y la gestión operativa del sistema [12].

Uno de los principales desafíos en la detección de fraudes y anomalías radica en la definición precisa de lo que constituye un *comportamiento normal de consumo*, ya que este puede variar según el tipo de cliente, la estacionalidad o las condiciones socioeconómicas y geográficas del entorno [13]. Además, los sistemas de detección deben ser capaces de procesar datos incompletos, ruidosos o heterogéneos, y adaptarse a patrones de consumo que evolucionan con el tiempo. En este sentido, las técnicas de analítica de datos y aprendizaje automático resultan esenciales para modelar el comportamiento y distinguir entre irregularidades técnicas, variaciones legítimas y conductas potencialmente fraudulentas en el suministro de agua.

Agua No Facturada (ANF)

Hace referencia a la diferencia entre toda el agua que un sistema de distribución produce para ser suministrada y la cantidad de agua que realmente se registra como utilizada por los usuarios y se cobra. Este concepto incluye pérdidas reales, como fugas, y pérdidas aparentes, causadas por imprecisiones de medidores y fraudes, siendo estas últimas las que generan importantes pérdidas financieras. Estas pueden mitigarse mejorando la precisión de los medidores y aplicando técnicas de detección de anomalías [10], [14]. Matemáticamente se describe como:

$$ANF = \frac{Volumen\,producido - Volumen\,facturado}{Volumen\,producido} \cdot 100$$

Medición mecánica vs. Medición inteligente

Históricamente, la medición del consumo de agua se ha realizado mediante medidores mecánicos, los cuales operan con principios volumétricos o de desplazamiento positivo y requieren lecturas manuales periódicas. Este tipo de medición, aunque robusta y de bajo costo, presenta limitaciones importantes en términos de resolución temporal, exactitud y disponibilidad de información en tiempo real. La necesidad de registrar los datos de forma manual introduce un margen de error asociado a la lectura, digitación y frecuencia de medición, lo cual dificulta la detección temprana de anomalías como fugas, fraudes o fallas en la infraestructura [15].

En contraste, los sistemas de medición inteligente, que integran tecnologías de lectura automática de medidores (AMR, por sus siglas en inglés) y de infraestructura avanzada de medición (AMI), permiten la captura y transmisión continua de datos de consumo en intervalos cortos de tiempo. Estos dispositivos recopilan información de manera remota y automatizada, eliminando la necesidad de intervención humana en la toma de lecturas. Su adopción ha transformado los procesos de monitoreo y análisis, al posibilitar la aplicación de técnicas avanzadas de analítica de datos y aprendizaje automático para la detección de patrones de consumo inusuales o fraudulentos [10], [13].

La medición inteligente ofrece ventajas sustanciales frente a la medición mecánica: permite una mayor granularidad temporal, reduce los errores humanos y proporciona datos más precisos para la modelación predictiva. Además, favorece la implementación de estrategias de mantenimiento preventivo y la gestión eficiente de recursos, al detectar variaciones anómalas casi en tiempo real. Sin embargo, su adopción también implica desafíos técnicos y económicos, como la interoperabilidad de sistemas, la gestión del gran volumen de datos generados y la protección de la privacidad de los usuarios [16].

En conjunto, la transición de la medición mecánica a la medición inteligente representa un cambio estructural en la gestión del recurso hídrico. Mientras los medidores tradicionales proporcionan una visión estática y limitada del consumo, los medidores inteligentes ofrecen una perspectiva dinámica y continua, esencial para el desarrollo de modelos de predicción, la detección temprana de fraudes y la optimización operativa de las empresas de agua.

Aspectos legales en Chile

El marco legal sobre el uso y gestión de aguas en Chile establece sanciones y regulaciones para garantizar su sostenibilidad. El *Código Penal (Artículo 489° bis)* sanciona con presidio, multas de 500 a 5.000 unidades tributarias mensuales e indemnización a quienes, sin título legítimo, usen, contaminen o dañen aguas superficiales o subterráneas [17]. Por su parte, la Ley de Servicios Sanitarios (Artículo 36) regula derechos y obligaciones de prestadores y usuarios, permitiendo, entre otros, suspender servicios por morosidad y cobrar daños causados por usuarios [18]. Además, el *Artículo 47A* obliga a concesionarias a compartir sus redes con otras empresas para grandes consumidores, promoviendo eficiencia y acceso equitativo [19].

4.1.2. MÉTODOS EN CIENCIA DE DATOS

La detección de fraudes y anomalías en el consumo de agua es fundamental para optimizar la gestión de las empresas de suministro y reducir pérdidas técnicas y comerciales. Estos fraudes pueden originarse por manipulaciones, mientras que las anomalías por fallas en los medidores o problemas en la infraestructura [12]. Los métodos tradicionales de monitoreo, basados en inspecciones manuales y

revisiones periódicas, resultan costosos e insuficientes frente al volumen de datos generado por los medidores, lo que resalta la necesidad de soluciones automatizadas [20]. La automatización en este ámbito proporciona beneficios importantes, como la optimización de recursos, que prioriza inspecciones en áreas con mayor probabilidad de fraudes o anomalías [20], la reducción de pérdidas comerciales, mediante la detección oportuna de fugas y fraudes para disminuir el Agua No Facturada (ANF) [14], y la de la eficiencia operativa, al facilitar una respuesta rápida y precisa a incidentes, optimizando flujos de trabajo y reduciendo costos operativos.

La *Minería de Datos* es el uso de técnicas analíticas avanzadas para descubrir patrones, relaciones y fraudes o anomalías en grandes volúmenes de datos. En el consumo de agua, permite identificar comportamientos inusuales o indicadores de manipulación en los medidores [21]. Los métodos como redes neuronales, árboles de decisión y regresión logística son herramientas clave para analizar y clasificar estos datos complejos. Su impacto en la detección de fraudes y anomalías es significativo. Estas técnicas reducen el tiempo necesario para identificar irregularidades y ayudan a focalizar inspecciones en casos de alta probabilidad de fraude. Esto mejora la eficiencia operativa de las empresas de suministro, optimizando la asignación de recursos y reduciendo pérdidas económicas. Además, los modelos predictivos generados apoyan decisiones estratégicas en la gestión de recursos hídricos [21].

En la analítica de datos, los métodos supervisados y no supervisados son enfoques fundamentales en el desarrollo de modelos. Los *métodos supervisados* utilizan datos etiquetados para entrenar modelos, como los de clasificación, que asignan categorías [22], mientras que los métodos *no supervisados* trabajan con datos no etiquetados para encontrar patrones, como el *clustering*, que agrupa datos similares [22]. Los datos utilizados para ajustar el modelo se denominan *conjunto de entrenamiento*, y los empleados para evaluar su desempeño en nuevas situaciones conforman el *conjunto de prueba*. Estos conceptos son fundamentales para garantizar que el modelo aprenda patrones relevantes y pueda generalizar correctamente a datos no vistos, asegurando precisión y confiabilidad en su aplicación [23].

Datos atípicos, erróneos y faltantes: Mecanismos de generación

En todo proceso de análisis de datos, la calidad de la información constituye un requisito esencial para la validez de los resultados. Sin embargo, los conjuntos de datos reales suelen contener valores atípicos, erróneos y faltantes, originados por diferentes mecanismos de generación asociados con el proceso de medición, captura, transmisión o incluso con comportamientos humanos. Comprender dichos mecanismos permite distinguir entre irregularidades aleatorias, errores sistemáticos y ausencias informativas, facilitando la selección de estrategias adecuadas de limpieza, imputación o modelado [24], [25].

Los *datos atípicos* son observaciones que se desvían significativamente del patrón general de los datos. Su aparición puede deberse a causas instrumentales (como fallas en sensores o errores de calibración), a procesos naturales excepcionales o a comportamientos anómalos en los sistemas observados [26]. En la literatura, se reconoce que los mecanismos que generan datos atípicos pueden clasificarse como aleatorios, cuando surgen por fluctuaciones o ruido no sistemático, o sistemáticos, cuando reflejan cambios estructurales o eventos relevantes del fenómeno estudiado. En el contexto de consumo de agua, estos valores pueden representar errores de lectura o indicadores de fraude, siendo necesario analizarlos antes de su eliminación, ya que podrían contener señales significativas sobre alteraciones en el medidor o patrones irregulares de consumo [27]. Desde la perspectiva del análisis exploratorio de datos, Tukey propuso un criterio robusto para caracterizar valores extremos sin asumir normalidad en

la distribución. Este enfoque se basa en el rango intercuartílico (IQR), definido como [28]:

$$\text{IQR} = Q_3 - Q_1,$$

y considera como observaciones inusuales aquellos valores que se encuentran por fuera del siguiente intervalo:

$$Q_1 - 1,5 \cdot \text{IQR} \quad \text{a} \quad Q_3 + 1,5 \cdot \text{IQR}.$$

Dado su carácter descriptivo, robusto y ampliamente aceptado en la literatura, este criterio constituye uno de los marcos más utilizados para interpretar valores extremos. En coherencia con ello, fue también el referente adoptado en el proyecto como base conceptual para la identificación de valores atípicos [28], [29].

Los *datos erróneos*, por su parte, corresponden a registros que violan las restricciones lógicas, físicas o semánticas del sistema. Su mecanismo de generación se asocia generalmente con fallas humanas o tecnológicas, tales como errores de digitación, inconsistencias de unidades, duplicación de registros o pérdida de precisión en la transmisión [24]. En bases de datos provenientes de medidores, estos errores suelen manifestarse como valores imposibles (por ejemplo, consumos negativos o lecturas abruptamente elevadas) o discordancias entre variables correlacionadas, como volumen medido y tiempo de lectura. La identificación de estos casos requiere aplicar reglas de validación y análisis de consistencia.

En coherencia con esta distinción, es importante señalar que, entre los valores de consumo señalados como atípicos bajo el criterio de Tukey, se revisó conceptualmente cuáles podían corresponder a variaciones reales del fenómeno y cuáles, en cambio, debían interpretarse como datos erróneos. Este análisis resulta fundamental porque no todo valor extremo constituye necesariamente un error: algunos consumos elevados pueden ser viables dentro del comportamiento del usuario o reflejar circunstancias particulares del servicio, mientras que otros sí representan inconsistencias que deben corregirse o excluirse del análisis [30].

Finalmente, los *datos faltantes* constituyen un tipo particular de irregularidad que ocurre cuando la información de una o más variables está ausente. El estudio de su mecanismo de generación es fundamental, ya que de ello depende la validez del análisis posterior. Los datos faltantes pueden generarse bajo tres mecanismos principales: faltantes completamente al azar (MCAR), cuando la ausencia no depende de ningún valor observado o no observado; faltantes al azar (MAR), cuando la probabilidad de ausencia está relacionada con otras variables observadas; y faltantes no al azar (MNAR), cuando la ausencia depende del propio valor faltante [25], [31], [32]. En el dominio del consumo de agua, estos mecanismos pueden reflejar problemas operativos (fallas en la transmisión o en los dispositivos de lectura), errores humanos (omisiones en el registro) o incluso acciones intencionales de manipulación para ocultar el consumo. En este último caso, la falta de datos puede adquirir relevancia analítica, pues podría actuar como una señal indirecta de fraude o comportamiento anómalo.

Finalmente, los *datos faltantes* constituyen un tipo particular de irregularidad que ocurre cuando la información de una o más variables está ausente. El estudio de su mecanismo de generación es fundamental, ya que de ello depende la validez del análisis posterior. En términos estadísticos, los datos faltantes pueden originarse bajo tres esquemas principales: faltantes completamente al azar, cuando la ausencia no depende de ningún valor observado o no observado; faltantes al azar, cuando

la probabilidad de ausencia está relacionada con otras variables observadas; y faltantes no al azar, cuando la ausencia depende del propio valor que falta [25], [31], [32]. En el ámbito del consumo de agua, estos mecanismos pueden reflejar problemas operativos (como fallas en la transmisión o en los dispositivos de lectura), errores humanos (omisiones involuntarias en el registro) o incluso acciones intencionales orientadas a ocultar información relevante sobre el consumo.

En el contexto de este proyecto, es importante señalar que no existen datos faltantes asociados a fallas de transmisión, dado que los registros provienen de medidores mecánicos. Por ello, los valores ausentes en las series de consumo se interpretan como errores de lectura o de reporte y, en consecuencia, se consideran casos en los que la ausencia ocurre completamente al azar. Bajo esta interpretación, su tratamiento puede apoyarse en técnicas de imputación diseñadas para series de tiempo. Para el resto de las variables, los datos faltantes también ocurren completamente al azar, dado que los análisis descriptivos y las comparaciones entre variables no evidenciaron patrones sistemáticos que indicaran una ausencia dependiente de otras variables o del propio valor faltante.

En conjunto, los datos atípicos, erróneos y faltantes no solo representan fuentes potenciales de distorsión estadística, sino también manifestaciones del proceso subyacente que genera los datos. Su estudio desde una perspectiva teórica y aplicada permite diferenciar entre ruido y señal, mejorar la integridad del conjunto de datos y fortalecer la capacidad del modelo para capturar patrones reales del fenómeno observado. En proyectos de detección de fraude y anomalías, estas consideraciones resultan esenciales para garantizar que la información procesada refleje adecuadamente tanto las variaciones naturales del sistema como las alteraciones inducidas por intervenciones humanas.

Georreferenciación de datos

La georreferenciación es el proceso mediante el cual se asocia un dato o registro al espacio geográfico, otorgándole coordenadas que permiten su representación y análisis dentro de un sistema de referencia espacial. En el caso de los datos alfanuméricos que contienen información textual sobre ubicaciones (como direcciones postales, nombres de calles, barrios o municipios), el proceso específico de conversión de dichas descripciones en coordenadas se denomina *geocodificación* (*geocoding*). Este procedimiento constituye una de las herramientas fundamentales para la integración de datos espaciales en los sistemas de información geográfica (SIG) y en aplicaciones de analítica espacial [33], [34].

La georreferenciación de direcciones permite vincular la información de registros administrativos, comerciales o técnicos con su localización física, lo que amplía las posibilidades de análisis, visualización y toma de decisiones basadas en la dimensión territorial. Desde una perspectiva teórica, la geocodificación parte del principio de que toda entidad o fenómeno social, económico o ambiental se manifiesta en un lugar específico del espacio geográfico, y que su estudio adquiere mayor valor cuando puede analizarse en relación con su contexto espacial [35].

En el ámbito de los sistemas de datos urbanos, la georreferenciación cumple un papel esencial para comprender la distribución espacial de fenómenos, identificar patrones de concentración o dispersión, y establecer relaciones entre variables espaciales y no espaciales. En el contexto del análisis de consumo de agua, por ejemplo, la asignación de coordenadas a los registros de medidores permite representar espacialmente los patrones de consumo, detectar áreas con comportamientos fraudulentos o anómalos, estimar la densidad de consumo por zona y correlacionar dichos valores con variables geográficas como altitud, densidad poblacional o infraestructura de red.

Conceptualmente, la precisión y confiabilidad de la georreferenciación dependen de la calidad de la información original (formato, completitud y normalización de las direcciones) y de la exactitud de las bases cartográficas o de referencia utilizadas. Las bases de direcciones pueden estructurarse en niveles jerárquicos (país, departamento, municipio, localidad, vía y número) que, una vez codificados, permiten vincular los datos a sistemas de proyección y coordenadas definidos. En términos teóricos, la georreferenciación se concibe como una operación de correspondencia entre un sistema de información semántico (el registro textual) y un sistema geométrico (el espacio físico), mediante un proceso de traducción apoyado en topologías y relaciones espaciales [36].

Así, la georreferenciación de registros constituye un componente esencial de la infraestructura de datos espaciales moderna, al permitir la integración de información heterogénea bajo un marco común de referencia geográfica. Su aplicación trasciende la cartografía y alcanza campos como la analítica de datos, la planeación territorial, la gestión de servicios públicos y la modelación de fenómenos urbanos y ambientales.

Kernel de densidad

La estimación de densidad mediante kernel, o *Kernel Density Estimation (KDE)*, es una técnica *no paramétrica* utilizada para aproximar la función de densidad de probabilidad de una variable aleatoria a partir de una muestra finita de datos. A diferencia de los métodos paramétricos, que suponen una forma funcional predefinida (por ejemplo, normal o exponencial), la KDE permite estimar la distribución subyacente sin imponer un modelo específico, ofreciendo así una representación más flexible de la estructura de los datos [37].

Formalmente, la estimación de densidad para un conjunto de observaciones $x_1, x_2, \ldots, x_n$ se define como:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4.1}$$

donde $K(\cdot)$ es la *función kernel*, y $h > 0$ es el *parámetro de suavizamiento* o *bandwidth*, que controla el grado de suavidad de la estimación.

La función kernel K actúa como un ponderador que asigna mayor influencia a las observaciones cercanas al punto x, decreciendo progresivamente con la distancia. Comúnmente, se emplean funciones simétricas y con soporte compacto, tales como el *kernel* gaussiano, epanechnikov o uniforme.

El parámetro de suavizamiento h es crítico para el desempeño del estimador: valores pequeños de h generan una densidad altamente variable (sobreajuste), mientras que valores grandes producen una estimación excesivamente lisa (subajuste). En términos teóricos, la KDE busca un equilibrio entre *sesgo* y *varianza*, reflejando la clásica tensión del principio de parsimonia estadística [38].

En el ámbito de la analítica espacial y geográfica, la estimación de densidad por kernel se ha convertido en una herramienta fundamental para *identificar concentraciones o patrones de intensidad* en el territorio. A partir de coordenadas geográficas, la KDE genera una superficie continua de probabilidad o densidad de eventos, permitiendo visualizar zonas de alta concentración o puntos calientes (*hotspots*). Esta propiedad la hace especialmente útil en campos como la criminología, la epidemiología o el análisis

de fraudes, donde los fenómenos tienden a presentar *dependencia espacial* y *heterogeneidad local* [39].

En síntesis, la estimación de densidad mediante kernel constituye un enfoque robusto y versátil para la exploración de distribuciones de datos sin necesidad de supuestos paramétricos estrictos. Su capacidad para adaptarse a distintas estructuras y escalas de los datos la convierte en una técnica esencial tanto en el análisis estadístico clásico como en la modelación espacial moderna.

Series de tiempo: imputación y extracción de características

Las series de tiempo son conjuntos de observaciones registradas en intervalos de tiempo sucesivos y equidistantes, en los que el orden cronológico de los datos desempeña un papel fundamental. A diferencia de los datos transversales, las series temporales permiten analizar la evolución de un fenómeno a lo largo del tiempo, identificando patrones sistemáticos y estructuras de dependencia entre observaciones adyacentes [40]. En el contexto de la analítica de datos aplicada al consumo de agua, una serie de tiempo representa el registro histórico del consumo asociado a un medidor, lo que permite examinar su comportamiento, detectar anomalías o predecir su evolución futura.

En el análisis de series de tiempo reales, es frecuente encontrar valores faltantes o inconsistentes que pueden alterar las estimaciones de tendencia, estacionalidad o complejidad. Por ello, antes de la modelación es necesario aplicar técnicas de imputación que preserven la estructura temporal de los datos. Entre los métodos más utilizados se encuentran la *interpolación lineal* y los *splines cúbicos*, que permiten estimar los valores ausentes a partir de la información disponible en los puntos vecinos [40].

La *interpolación lineal* asume una variación uniforme entre dos observaciones consecutivas y calcula el valor faltante como:

$$\hat{y}_t = y_{t-1} + \frac{(y_{t+1} - y_{t-1})}{(t_{+1} - t_{-1})}(t - t_{-1})$$

Este método es adecuado cuando los datos presentan cambios graduales y sin oscilaciones abruptas. En el contexto de este trabajo, los términos de la expresión anterior corresponden directamente a los valores de la serie de tiempo de consumo mensual de cada cliente. Así, y_t representa el consumo registrado en el mes t, mientras que y_{t-1} y y_{t+1} corresponden, respectivamente, a los consumos observados en el mes anterior y en el mes posterior al valor faltante. Es decir, la interpolación lineal utiliza los consumos adyacentes de la misma unidad de medición (el mismo cliente y el mismo medidor) para estimar el valor ausente dentro de su trayectoria temporal.

Por su parte, la *interpolación mediante splines cúbicos* ajusta una serie de funciones polinómicas por tramos, garantizando continuidad en el valor y en las dos primeras derivadas de la serie. Matemáticamente, un spline cúbico $S(t)$ se define como una combinación de polinomios cúbicos locales:

$$S(t) = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3, \quad t_i \leq t \leq t_{i+1}$$

donde los coeficientes a_i, b_i, c_i, d_i se determinan de forma que la función sea suave en los puntos de unión. En el caso del presente trabajo, la interpolación mediante splines cúbicos, $S(t)$ denota la función suave que aproxima la trayectoria temporal del consumo para un cliente. Cada intervalo $[t_i, t_{i+1}]$ corresponde a dos meses consecutivos con lecturas válidas, y la función $S(t)$ permite reconstruir el comportamiento intermedio. En este trabajo, los puntos t_i representan los meses en los cuales existe una lectura real de consumo, mientras que la evaluación de $S(t)$ en los valores faltantes proporciona

una estimación continua y coherente con la forma global de la serie. Este enfoque permite reconstruir trayectorias más realistas y continuas, especialmente en series con comportamiento no lineal o fluctuaciones estacionales. La elección entre interpolación lineal o spline depende del nivel de variabilidad de la serie y del impacto que los valores faltantes tengan en los análisis posteriores, como el cálculo de métricas simbólicas o de complejidad [41].

Series de tiempo: descomposición

Una vez imputados los datos faltantes o datos erróneos, el valor observado y_t se define como el consumo mensual efectivamente registrado por el medidor de un cliente en el mes t. A partir de este valor, la serie de consumo puede descomponerse en varios componentes fundamentales:

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

y se interpreta del siguiente modo: T_t refleja la tendencia o evolución de largo plazo del consumo en el mes t; S_t representa la estacionalidad propia de patrones que se repiten periódicamente a lo largo de los meses del año; C_t captura posibles oscilaciones cíclicas no estrictamente periódicas; y ε_t recoge el ruido aleatorio o la variación no explicada en ese mismo mes. De esta manera, cada componente se entiende siempre en relación con el valor mensual del consumo para un cliente en el instante temporal t.

La descomposición puede asumirse *aditiva*, cuando las variaciones son independientes de la magnitud del valor medio, o *multiplicativa*, cuando las amplitudes varían proporcionalmente con el nivel de la serie [42].

Series de tiempo: normalización

En el contexto del análisis de series de tiempo, la reducción de la dimensionalidad y la extracción de características permiten representar la estructura esencial de los datos sin perder su información temporal relevante. Estas técnicas buscan condensar las variaciones significativas y facilitar la identificación de patrones característicos, irregularidades y comportamientos atípicos [42]. Entre los métodos más utilizados se encuentran el escalamiento de los datos, las representaciones simbólicas y las métricas de complejidad temporal. La normalización mín–máx constituye una técnica de escalamiento lineal ampliamente utilizada para estandarizar los valores de una serie dentro de un rango fijo, generalmente [0,1], lo que permite comparar diferentes perfiles de comportamiento independientemente de su magnitud o unidad de medida. En este trabajo, la serie de tiempo se ha representado de forma consistente mediante la notación y_t, donde cada y_t corresponde al consumo mensual registrado por un cliente en el mes t. Para referirnos a la serie completa, utilizamos la notación $Y = \{y_t\}_{t=1}^{T}$, entendida como la secuencia ordenada de todos los valores mensuales del periodo de observación.

Bajo esta notación, la normalización min–max se aplica directamente sobre la serie Y, generando una versión escalada Y^*. Así, cada valor transformado y_t^* se obtiene mediante:

$$y_t^* = \frac{y_t - \text{mín}(Y)}{\text{máx}(Y) - \text{mín}(Y)}.$$

Este proceso preserva las proporciones relativas entre los valores mensuales de consumo y evita que observaciones extremas dominen el análisis posterior. En el ámbito de la complejidad temporal, esta normalización resulta indispensable para el cálculo de métricas como la *Lempel–Ziv Complexity* (LZC) y el *Time Series Length Factor* (TSLF), que requieren trabajar con series escaladas para caracterizar la irregularidad y la estabilidad de los patrones de consumo. Este procedimiento preserva las proporciones relativas entre observaciones y evita que las series con valores extremos dominen el análisis posterior.

En el ámbito de la complejidad temporal, la normalización resulta indispensable para el cálculo de métricas como la *Lempel–Ziv Complexity* (LZC) y el *Time Series Length Factor* (TSLF). En el contexto de este trabajo, estas métricas se utilizan en la etapa de ingeniería de características descrita en la metodología, específicamente dentro del proceso de extracción de atributos derivados de las series mensuales de consumo.

La *LZC* y el *TSLF* forman parte del conjunto de variables utilizadas como insumo de los modelos predictivos de fraude y anomalías. Su propósito es capturar propiedades estructurales de las series temporales que no pueden ser detectadas mediante estadísticas convencionales, tales como media, varianza o curtosis. La *LZC* cuantifica el grado de irregularidad o complejidad en la secuencia de consumos mensuales, permitiendo identificar clientes cuyo comportamiento presenta patrones altamente impredecibles o fragmentados. Por su parte, el TSLF evalúa la estabilidad y la variación relativa en los tramos de la serie, proporcionando información sobre cambios abruptos o alteraciones en la continuidad del consumo [43]. Estas métricas fueron incorporadas directamente en los conjuntos de características utilizados para entrenar los modelos *Random Forest* y *XGBoost*, y demostraron aportar valor predictivo.

Ahondando en la formulación matemática de las variables mencionadas anteriormente, la complejidad de Lempel–Ziv (*Lempel–Ziv Complexity*, LZC) se define como:

$$\mathrm{LZC}_{abs} = \frac{c(n)}{n/\log_2 n}$$

donde $c(n)$ corresponde al conteo de subcadenas únicas identificadas en una secuencia de longitud n. Con el fin de facilitar la comparación entre series de diferente longitud, se emplea una versión normalizada dada por:

$$\mathrm{LZC}_{norm} = \frac{\mathrm{LZC}_{abs}}{\mathrm{LZC}_{abs}^{max}}$$

donde LZC_{abs}^{max} representa la complejidad teórica de una secuencia completamente aleatoria. Valores cercanos a 1 indican una dinámica compleja y poco predecible, mientras que valores bajos evidencian patrones regulares o deterministas. De manera complementaria, el *Time Series Length Factor* (TSLF) mide la estabilidad relativa de una serie temporal y se expresa como:

$$\mathrm{TSLF} = \frac{N}{\sum_{i=1}^{N-1} |y_{i+1} - y_i| + 1}$$

En este trabajo, N es el número de meses observados en la serie de consumo de un cliente y y_i representa el consumo mensual del cliente en el mes i. Un TSLF alto indica estabilidad y variaciones suaves, mientras que valores bajos reflejan cambios abruptos o inestabilidad temporal.

En cuanto a la representación simbólica, el método *Symbolic Aggregate Approximation* (SAX) [44] ha sido ampliamente adoptado como técnica de reducción de dimensionalidad. SAX transforma una serie temporal continua en una secuencia discreta de símbolos que conservan las tendencias principales del comportamiento original. Este proceso se realiza en dos etapas: primero, la serie se divide en segmentos de igual longitud y se calcula el promedio de cada uno mediante la aproximación por segmentos (*Piecewise Aggregate Approximation*, PAA):

$$\bar{y}_i = \frac{\beta}{N} \sum_{j=\frac{N}{\beta}(i-1)+1}^{\frac{N}{\beta}i} y_j$$

En el contexto de este trabajo, y_j corresponde al consumo mensual del cliente en el mes j, mientras que N representa la longitud total de la serie de consumo del cliente, es decir, el número de meses observados en su historial. El parámetro β indica el tamaño de cada segmento en que se divide la serie para aplicar la aproximación PAA, de modo que cada bloque de β meses consecutivos se resume mediante su promedio. Así, $\bar{y}_i$ denota el valor medio del consumo en el segmento i, el cual sirve como base para la representación simbólica SAX empleada posteriormente en la ingeniería de características para la detección de fraude y anomalías.

Posteriormente, los valores promedio se discretizan de acuerdo con los umbrales de una distribución normal estándar, asignando un símbolo a cada intervalo. De esta forma, la serie original $Y = (y_1, y_2, \ldots, y_N)$ se convierte en una secuencia simbólica $S = \{s_1, s_2, \ldots, s_\beta\}$, que mantiene las fluctuaciones esenciales del proceso y permite aplicar métricas como la entropía o los cambios de símbolo.

La *Entropía SAX*, mide la diversidad de símbolos en la secuencia y se define como:

$$H = -\sum_{i=1}^{\alpha} p_i \log_2 p_i$$

donde p_i representa la probabilidad de aparición de cada símbolo dentro del alfabeto de tamaño α. Valores altos de H reflejan mayor irregularidad o variabilidad en la secuencia temporal, mientras que valores bajos indican uniformidad o repetición de patrones [45]. De manera complementaria, el número de cambios de símbolo entre posiciones consecutivas puede expresarse como:

$$C_s = \sum_{i=1}^{n-1} I(s_i \neq s_{i+1})$$

En esta métrica, s_i representa el símbolo SAX asociado al consumo mensual del medidor en el mes i, mientras que s_{i+1} corresponde al símbolo del mes siguiente. El parámetro n denota la longitud de la serie simbólica (es decir, el número de meses observados). La función indicadora $\mathbf{I}(s_i \neq s_{i+1})$ toma el valor 1 cuando hay un cambio de símbolo entre dos meses consecutivos y 0 en caso contrario. En el contexto del proyecto, C_s resume la frecuencia de fluctuaciones relevantes en el consumo, lo que permite identificar patrones inestables asociados a anomalías técnicas y ciertas irregularidades vinculadas al fraude.

Junto con las métricas de complejidad y representación simbólica, es posible incorporar variables estadísticas complementarias que describen la estructura general y la estabilidad de una serie temporal. Estas métricas permiten analizar la tendencia central, la dispersión, la forma y la presencia de valores atípicos, proporcionando una descripción cuantitativa del comportamiento global del sistema

[40], [46]. Entre las más utilizadas se encuentran la *media* ($\bar{x}$), la *mediana* ($\tilde{x}$) y la *desviación estándar* (SD), que cuantifican el valor central y la variabilidad de los consumos. La *media* se define como:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

y la *desviación estándar* como:

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

mientras que el *coeficiente de variación* (CV) expresa la variabilidad relativa respecto a la *media* y se calcula mediante:

$$CV = \frac{SD}{\bar{x}} \cdot 100$$

La forma de la distribución puede describirse mediante la *asimetría* (Sk) y la *curtosis* (K), las cuales permiten identificar sesgos o concentraciones inusuales de valores. La *asimetría* mide el grado de simetría de la distribución en torno a la media, y se calcula como

$$Sk = \frac{N\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)(N-2)\,SD^3}$$

mientras que la *curtosis* cuantifica el peso de las colas de la distribución y se define como:

$$K = \frac{N(N+1)\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)(N-2)(N-3)\,SD^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Una distribución simétrica presenta $Sk \approx 0$, mientras que valores de curtosis superiores a 3 indican colas más pesadas que la distribución normal.

En cuanto a las medidas de posición, los cuartiles dividen los datos en cuatro partes iguales, y el rango intercuartílico (IQR) mide la dispersión del 50 % central de la distribución, definido como

$$IQR = Q_3 - Q_1$$

A partir del IQR es posible identificar valores atípicos. Los *valores moderadamente atípicos* son aquellos que se encuentran fuera del rango determinado por 1.5 veces el IQR:

$$x_i < Q_1 - 1{,}5 \cdot IQR \quad \text{o} \quad x_i > Q_3 + 1{,}5 \cdot IQR$$

mientras que los *valores extremadamente atípicos* son aquellos que exceden tres veces dicho rango:

$$x_i < Q_1 - 3 \cdot IQR \quad \text{o} \quad x_i > Q_3 + 3 \cdot IQR$$

Otra medida complementaria para la detección de valores atípicos es el *Z-score*, que mide cuántas desviaciones estándar se encuentra un valor respecto a la media, y se expresa como:

$$z_i = \frac{x_i - \bar{x}}{SD}$$

Un valor se considera atípico si $|z_i| > 2$, y el número *total de valores atípicos* puede estimarse mediante

$$N_{\text{outliers}} = \sum_{i=1}^{N} I(|z_i| > 2)$$

donde $I(\cdot)$ es la función indicadora que vale 1 si la condición es verdadera y 0 en caso contrario.

La estabilidad temporal también puede analizarse mediante el rango ($R = x_{\text{máx}} - x_{\text{mín}}$) y por la detección de cambios bruscos (*Change Points*), definidos cuando la diferencia entre valores consecutivos excede un múltiplo del desvío estándar. El número de cambios bruscos se puede estimar mediante:

$$\Delta_{\text{brusco}} = \sum_{i=1}^{N-1} I(|x_{i+1} - x_i| > 3 \cdot SD)$$

Finalmente, la recurrencia o continuidad de la serie puede representarse por el conteo de *secuencias de ceros*, que indican periodos sin consumo o sin registro, y se expresa como:

$$N_{\text{ceros}} = \sum_{i=1}^{N} I(x_i = 0)$$

así como por la detección de secuencias consecutivas de baja variación, que permiten identificar estabilidad o interrupciones en la dinámica temporal:

$$N_{\text{secuencias}} = \sum_{i=1}^{N} I(\text{rlen}(\text{consumo} < 1) \geq 3)$$

En conjunto, las métricas estadísticas, simbólicas y de complejidad conforman un marco integral para la caracterización de series temporales. Su aplicación permite representar la estabilidad, irregularidad y estructura simbólica de los datos, favoreciendo la identificación de patrones anómalos, comportamientos atípicos y posibles irregularidades en procesos de medición o consumo [47].

Datos de entrenamiento y prueba y validación cruzada estratificada

El proceso de evaluación de modelos predictivos requiere dividir el conjunto de datos en subconjuntos que permitan estimar la capacidad de generalización del modelo. En el contexto del aprendizaje supervisado, es habitual separar la base en datos de entrenamiento y de prueba. El conjunto de entrenamiento (D_{train}) se utiliza para ajustar los parámetros del modelo, mientras que el conjunto de prueba (D_{test}) permite evaluar su desempeño sobre observaciones no vistas. Si D representa la base total con N observaciones, el proceso puede expresarse como:

$$D = D_{train} \cup D_{test}, \quad D_{train} \cap D_{test} = \varnothing$$

donde típicamente el conjunto de entrenamiento corresponde a una proporción entre el 70 % y el 80 % del total, y el conjunto de prueba al 20–30 % restante [48]. Este procedimiento busca evitar el sobreajuste (*overfitting*), es decir, que el modelo aprenda patrones específicos del conjunto de entrenamiento que no se generalizan a nuevos datos. En el caso del presente proyecto, se usó el 80 % del conjunto de datos para entrenamiento y el 20 % restante en prueba.

Para estimar de manera más robusta el desempeño del modelo, se recurre a la técnica de *validación cruzada* (*cross-validation*). En su forma más común, la validación cruzada k-fold consiste en

dividir el conjunto de datos en k subconjuntos (*folds*) de igual tamaño. En cada iteración, uno de los subconjuntos se utiliza como conjunto de prueba, mientras que los $k-1$ restantes se emplean para el entrenamiento. El proceso se repite k veces, de manera que cada subconjunto actúa una vez como conjunto de prueba, y el desempeño promedio se calcula como:

$$\text{CV}_k = \frac{1}{k}\sum_{i=1}^{k} \text{Métrica}_i$$

donde Métrica_i puede representar indicadores como la exactitud, la precisión, la sensibilidad o el área bajo la curva ROC (AUC) en la i-ésima iteración [49].

Cuando las clases del conjunto de datos se encuentran desbalanceadas, es decir, una categoría tiene una frecuencia significativamente menor que las demás, se recomienda aplicar una *validación cruzada estratificada*. En este método, cada subconjunto preserva aproximadamente la misma proporción de instancias de cada clase que la base original. Formalmente, si la clase minoritaria corresponde a una proporción p del conjunto total, entonces:

$$p = \frac{|C_{min}|}{|D|} \approx \frac{|C_{min}^{(i)}|}{|D^{(i)}|} \quad \text{para } i = 1, 2, \ldots, k$$

donde $C_{min}^{(i)}$ representa las instancias de la clase minoritaria en el i-ésimo subconjunto $D^{(i)}$. Este tipo de validación es especialmente útil en problemas de detección de anomalías o fraudes, donde la clase positiva suele ser escasa. La estratificación garantiza que cada partición contenga ejemplos de ambas clases, permitiendo que el modelo aprenda y se evalúe de manera consistente en todos los ciclos de validación [50], [51].

De esta forma, la combinación de un particionamiento adecuado de los datos y la validación cruzada estratificada constituye una práctica esencial para evaluar de manera confiable el rendimiento de los modelos predictivos, mitigando el sesgo de muestreo y proporcionando estimaciones más estables de las métricas de desempeño.

Optimización de hiperparámetros con GridSearchCV

La selección adecuada de *hiperparámetros* es esencial para garantizar el equilibrio entre ajuste, generalización y rendimiento de los modelos de aprendizaje automático. Estos parámetros, definidos antes del entrenamiento, determinan el comportamiento y la complejidad del modelo [52].

La *búsqueda en rejilla* o *Grid Search* constituye un método sistemático para optimizar hiperparámetros mediante la evaluación exhaustiva de todas las combinaciones posibles dentro de un espacio predefinido. Cada configuración se entrena y valida con base en una métrica de desempeño (por ejemplo, F1-Score o AUC), seleccionando aquella que maximiza el rendimiento promedio [53].

Su implementación mediante la función `GridSearchCV` en `scikit-learn` incorpora la técnica de *validación cruzada*, lo que permite estimar la capacidad de generalización del modelo. No obstante, su naturaleza exhaustiva conlleva un alto costo computacional, especialmente cuando el número de parámetros o rangos de búsqueda es amplio. Por esta razón, se han propuesto alternativas más eficientes, como la *Random Search* o la optimización bayesiana, que exploran el espacio de búsqueda de manera más selectiva [54].

En conclusión, el *Grid Search* ofrece una evaluación estructurada y reproducible del espacio de hiperparámetros, combinando simplicidad conceptual con rigor estadístico, y se mantiene como una referencia central en la optimización de modelos supervisados.

Calibración de probabilidades

La calibración de probabilidades busca ajustar las salidas de un modelo de clasificación para que las probabilidades estimadas reflejen de manera coherente la frecuencia real de ocurrencia del evento positivo. Los dos métodos más ampliamente empleados en la literatura son la *Regresión Isotónica* y el *Platt Scaling*, los cuales difieren en su naturaleza funcional y en los supuestos que imponen sobre la relación entre las puntuaciones del modelo y las probabilidades verdaderas [55].

En el contexto de este trabajo, la calibración de probabilidades se aplicó específicamente a los modelos supervisados de detección de fraude (*Random Forest* y *XGBoost*). Su propósito fue garantizar que las probabilidades generadas por los modelos fueran comparables y reflejaran adecuadamente el riesgo real de fraude, lo que resultó esencial para la construcción del rankeo de inspecciones y para la posterior evaluación económica basada en umbrales de decisión. Sin este ajuste, las probabilidades tenderían a estar sesgadas debido al fuerte desbalance de clases, afectando tanto la interpretación como la priorización operativa de los casos.

Dentro de los métodos de calibración tenemos la *Regresión Isotónica*, el cual es un método *no paramétrico* que asume una relación *monótona no decreciente* entre las salidas del modelo y las probabilidades observadas. Su objetivo es encontrar una función f que minimice el error cuadrático sujeto a la restricción de monotonía:

$$\min_{f} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \quad \text{sujeto a } f(x_i) \leq f(x_j) \text{ si } x_i < x_j$$

donde x_i son las probabilidades estimadas por el modelo, mientras que y_i son las etiquetas observadas. En el contexto de este trabajo, y_i corresponde a la variable binaria de referencia: *Fraude* en el modelo de fraude y *Anomalías* en el modelo de anomalías.

Por su parte, el *Platt Scaling*, también conocido como *regresión sigmoidal*, corresponde a un método *paramétrico* propuesto inicialmente para calibrar las salidas de los clasificadores SVM [55]. Parte del supuesto de que la relación entre la puntuación del modelo $f(x)$ y la probabilidad real puede aproximarse mediante una función logística:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(Af(x) + B)} \tag{4.2}$$

donde $P(y = 1 \mid x)$ denota la probabilidad de que un registro sea clasificado como un caso de fraude o anomalía (según el caso), condicionada a la información contenida en las variables predictoras x.

Los parámetros A y B se estiman mediante una regresión logística aplicada sobre un conjunto independiente destinado exclusivamente a la calibración. Este método proporciona un ajuste más estable y menos sensible al ruido, aunque suponer una forma sigmoidal puede limitar la flexibilidad frente a distribuciones no lineales complejas.

En síntesis, la regresión isotónica y el Platt Scaling representan enfoques *complementarios*: la primera prioriza la *flexibilidad* al no imponer una estructura funcional, mientras que la segunda privilegia la *estabilidad* mediante una formulación paramétrica y suavizada. La elección entre ambos depende de la cantidad de datos disponibles para calibración y del grado de no linealidad en las salidas del modelo base [56].

Matriz de confusión, métricas y tipo de error

En el análisis de modelos de clasificación, la evaluación del desempeño es un componente esencial para determinar la calidad de las predicciones. Una de las herramientas más utilizadas con este propósito es la matriz de confusión, la cual resume las coincidencias y discrepancias entre las etiquetas reales y las predichas por el modelo. En el contexto del presente trabajo, tanto el modelo de detección de fraude como el de anomalías se formulan como problemas de clasificación binaria, donde la variable respuesta y toma dos posibles valores. En el caso del modelo de detección de fraudes, *1* representa la existencia de fraude, mientras que 0 señala la ausencia del mismo. De manera análoga, en el modelo de anomalías, *1* indica la presencia de una anomalía en el consumo, mientras que *0* corresponde a un comportamiento normal. De acuerdo con lo anterior (problema binario), la matriz se organiza en cuatro categorías fundamentales: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Su estructura general se expresa como:

	Predicho Positivo	Predicho Negativo
Real Positivo	TP	FN
Real Negativo	FP	TN

A partir de estos elementos se derivan las métricas más utilizadas para cuantificar el rendimiento del modelo. El *Accuracy* mide la proporción de observaciones correctamente clasificadas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sin embargo, cuando las clases se encuentran desbalanceadas, la exactitud puede ser engañosa, ya que un modelo podría alcanzar valores altos simplemente favoreciendo la clase mayoritaria. Por esta razón, se emplean métricas adicionales que evalúan el comportamiento del modelo en cada clase. Entre ellas, la *Precision* y el *Recall* son las más relevantes y se definen como:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

La precisión mide el porcentaje de predicciones positivas que son realmente correctas, mientras que la sensibilidad refleja la capacidad del modelo para identificar correctamente los casos positivos. Ambas medidas pueden combinarse en la métrica $F1$, definida como la media armónica entre ambas:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

El valor de $F1$ es especialmente útil cuando se busca un equilibrio entre la precisión y la sensibilidad, como en los problemas de detección de fraudes o anomalías, donde los falsos negativos y los falsos positivos tienen costos distintos. En estos contextos, se analiza también la *especificidad*, que mide la proporción de verdaderos negativos correctamente identificados:

$$Specificity = \frac{TN}{TN + FP}$$

A partir de la sensibilidad y la especificidad se construye la curva ROC (*Receiver Operating Characteristic*), que representa la relación entre la tasa de verdaderos positivos ($TPR = Recall$) y la tasa de falsos positivos ($FPR = 1 - Specificity$). El área bajo la curva ROC (*AUC-ROC*) constituye una medida integral del desempeño del modelo, independiente de los umbrales de decisión, y es especialmente robusta ante el desbalance de clases [57].

El análisis de errores también permite identificar el tipo de fallo cometido por el modelo. Los falsos positivos (FP) ocurren cuando una observación negativa se clasifica incorrectamente como positiva, mientras que los falsos negativos (FN) representan casos positivos no detectados. En tareas sensibles, como la detección de fugas, fraudes o diagnósticos médicos, los errores de tipo II (FN) pueden tener consecuencias más graves que los de tipo I (FP), lo que justifica la priorización de métricas sensibles a la clase minoritaria [58], [59].

De manera complementaria, se emplean métricas más equilibradas para evaluar modelos en contextos con clases desbalanceadas, donde la proporción entre instancias positivas y negativas difiere significativamente. En estos casos, la exactitud tradicional puede sobreestimar el rendimiento real del modelo al favorecer la clase mayoritaria. Por ello, el *Balanced Accuracy*, que considera de manera equitativa la tasa de verdaderos positivos y la tasa de verdaderos negativos. Su formulación se expresa como:

$$Balanced\,Accuracy = \frac{Recall + Specificity}{2}$$

Esta métrica proporciona una visión más justa del desempeño del clasificador, ya que pondera por igual la capacidad del modelo para identificar correctamente tanto la clase minoritaria como la mayoritaria. De este modo, se reduce el sesgo introducido por el desequilibrio de clases y se obtiene una medida más representativa de la habilidad global del modelo [59].

En conjunto, la matriz de confusión y las métricas derivadas (como las mencionadas anteriormente) conforman un marco teórico robusto para la evaluación de modelos de clasificación. Estas permiten no solo cuantificar la proporción de aciertos, sino también analizar los tipos de error cometidos, su distribución entre clases y su relevancia según el contexto de aplicación, proporcionando una base sólida para la comparación e interpretación de resultados en escenarios complejos y desbalanceados.

Métodos supervisados

En el contexto del presente trabajo, los métodos supervisados constituyen el núcleo de la estrategia predictiva diseñada para la detección de fraude en el consumo de agua y la identificación de anomalías técnicas en los medidores. Dado que la empresa dispone de un historial de eventos confirmados de fraude y de fallas técnicas registradas en campo, fue posible emplear algoritmos supervisados para aprender patrones en los perfiles de consumo mensual. Estos modelos utilizan como variables predictoras un conjunto amplio de características derivadas de las series de tiempo de consumo, incluyendo estadísticas robustas, medidas de complejidad (LZC, TSLF), representaciones simbólicas (SAX), atributos espaciales (latitud, longitud, localidad) y factores metrológicos (clase del medidor, año de instalación), las cuales permiten capturar señales irregulares vinculadas tanto a manipulación intencional del medidor como a deterioros o fallas no intencionales.

La inclusión de estos métodos no responde únicamente a su uso extendido en la literatura, sino a su adecuación a los desafíos propios del dominio: fuerte desbalance de clases, patrones no lineales de

consumo y coexistencia de señales locales, temporales y técnicas. A continuación, se presentan los métodos supervisados considerados en esta investigación, haciendo énfasis en su fundamento teórico y en su pertinencia para el problema analizado.

- **Random Forest:** En el contexto de este trabajo, el *Random Forest* se incorporó debido a su capacidad para modelar interacciones complejas entre variables y manejar patrones de consumo altamente irregulares, algo recurrente tanto en casos de fraude como en fallas técnicas de los medidores . A diferencia de otros algoritmos supervisados, el Random Forest resulta especialmente adecuado cuando las señales relevantes no provienen de un único predictor, sino de combinaciones no lineales entre características temporales (como LZC, TSLF y estadísticas robustas), atributos espaciales y variables metrológicas [60]. Esta propiedad lo convierte en un candidato natural para capturar patrones sutiles en los perfiles mensuales de consumo, incluso cuando las irregularidades se manifiestan como fluctuaciones inestables, cambios abruptos o tendencias atípicas distribuidas a lo largo del tiempo.

  El método *Random Forest*, propuesto por Breiman [60], pertenece a la familia de los modelos de ensamblaje y representa una extensión del enfoque de *bagging* (*bootstrap aggregating*) aplicado a árboles de decisión. Su fundamento teórico radica en la generación de múltiples modelos base (árboles de decisión), cada uno entrenado sobre un subconjunto distinto del conjunto de datos original. Dichos subconjuntos se obtienen mediante muestreo aleatorio con reemplazo (*bootstrap*), lo que introduce variabilidad en el proceso de aprendizaje y reduce la dependencia entre árboles.

  Una vez dividido el conjunto de datos en particiones de entrenamiento y prueba, el algoritmo define un conjunto de hiperparámetros que controlan el número de árboles (B), la profundidad máxima de cada uno y la cantidad de variables consideradas en cada división (m_{try}). Durante la construcción del bosque, cada árbol se entrena con una muestra aleatoria del conjunto de entrenamiento, y en cada nodo se selecciona de forma aleatoria un subconjunto de variables predictoras. Esta doble fuente de aleatoriedad (tanto en las observaciones como en las variables) contribuye a disminuir la varianza y a aumentar la estabilidad de las predicciones finales.

  El proceso de agregación, eje central de la metodología, consiste en combinar los resultados individuales de los árboles entrenados. En problemas de clasificación, la predicción final se obtiene por votación mayoritaria, mientras que en regresión se emplea el promedio aritmético de las predicciones. Este principio de consenso puede expresarse como:

$$\hat{y} = \begin{cases} \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B), & \text{si es clasificación,} \\ \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b, & \text{si es regresión.} \end{cases}$$

  En este trabajo, el *Random Forest* se emplea exclusivamente como un modelo de clasificación, ya que las variables objetivo (fraude y anomalías técnicas) son de naturaleza binaria. Por ello, el modelo predice la pertenencia a una de dos clases (evento presente o ausente), y la decisión final se obtiene mediante votación mayoritaria entre los árboles que conforman el bosque.

  Dado que cada árbol debe generar estas predicciones de manera independiente, cada uno define las divisiones óptimas de sus nodos utilizando una medida de impureza. En tareas de clasificación, el índice de Gini es el criterio más común, definido como:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

donde p_i representa la proporción de observaciones de la clase i dentro del nodo y C el número total de clases. En regresión, el objetivo equivalente es minimizar la varianza intranodo, buscando la homogeneidad dentro de las particiones resultantes.

Una de las propiedades distintivas del algoritmo es la estimación interna del error mediante las observaciones no incluidas en cada muestra de entrenamiento, conocidas como *Out-of-Bag* (OOB). Estas observaciones se utilizan como subconjunto de validación para calcular una medida de desempeño sin requerir un conjunto de prueba adicional. Este mecanismo permite evaluar la generalización del modelo durante su construcción, fortaleciendo su fiabilidad estadística.

Asimismo, *Random Forest* ofrece una medida inherente de la importancia de las variables (*Feature Importance*), obtenida a partir de dos criterios principales: la reducción promedio de la impureza (Gini) y la disminución de la precisión cuando se permuta aleatoriamente una variable (*Permutation Importance*). Estas métricas proporcionan información valiosa para la interpretación del modelo y la identificación de los predictores con mayor influencia en la respuesta.

En conjunto, *Random Forest* integra la simplicidad de los árboles de decisión con el poder de generalización de los métodos de ensamblaje. Su estructura permite combinar resultados parciales de forma robusta, reduciendo el sobreajuste y mejorando la estabilidad del modelo ante datos ruidosos o de alta dimensionalidad. Por su equilibrio entre precisión, interpretabilidad y facilidad de implementación, se ha convertido en una técnica ampliamente utilizada en la predicción, la detección de anomalías y el análisis exploratorio de datos.

La Figura 4.2 sintetiza visualmente este proceso, mostrando cómo el algoritmo articula las etapas de muestreo, entrenamiento de árboles, validación OOB, cálculo de importancia de variables y combinación de resultados para generar la predicción final del modelo.

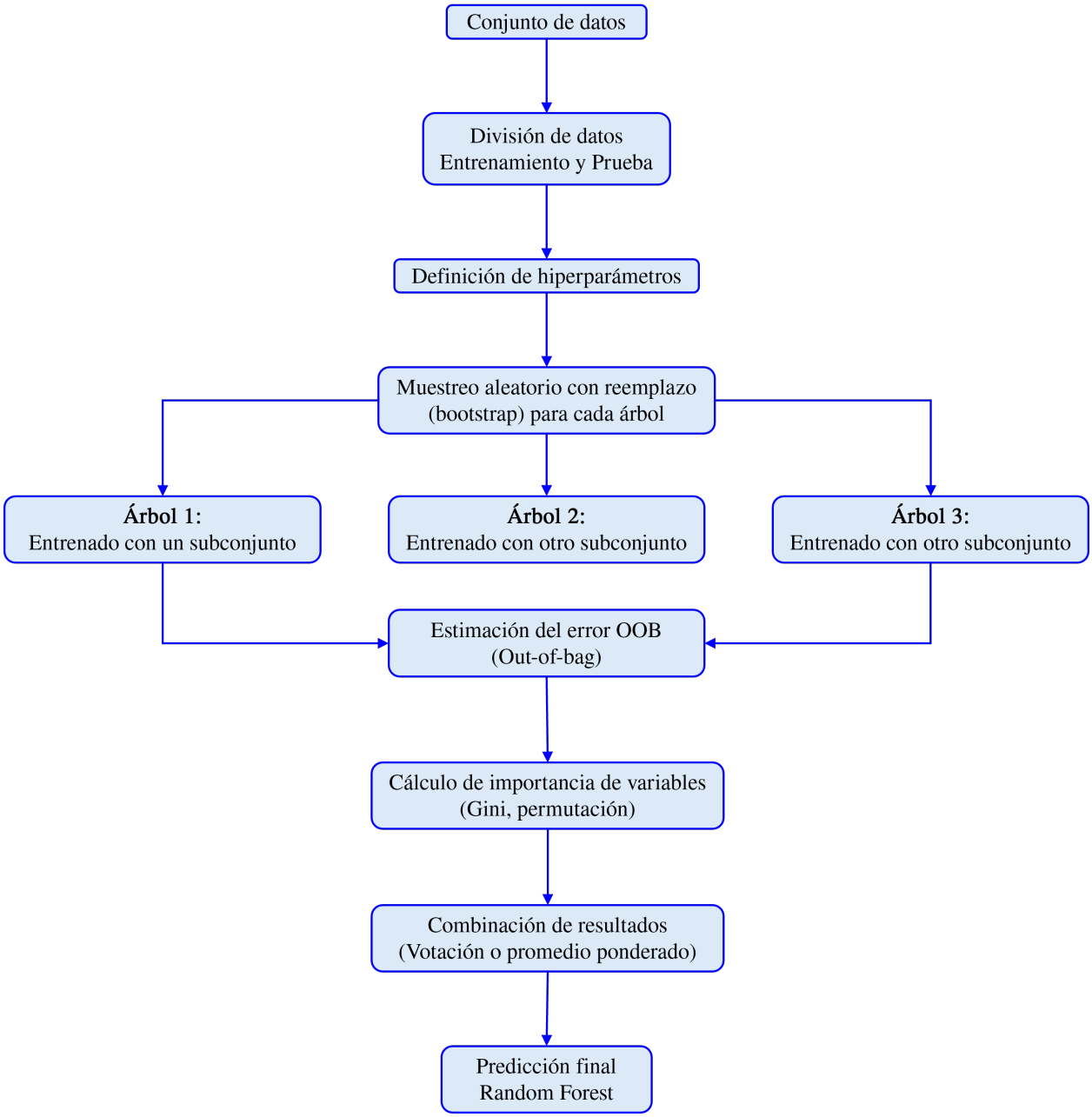

Figura 4.2: Flujograma del modelo *Random Forest*.
Fuente: Elaboración propia.

- **XGBoost:** En el contexto de este trabajo, *XGBoost* se empleó como alternativa supervisada para modelar patrones de consumo particularmente complejos, ya que su esquema secuencial permite refinar la detección de irregularidades mes a mes. Este enfoque resulta especialmente útil cuando las señales asociadas al fraude o a fallas técnicas son sutiles y requieren un modelo capaz de corregir sistemáticamente los errores previos durante el entrenamiento [61]. El algoritmo *XGBoost* (*Extreme Gradient Boosting*) es una de las implementaciones más eficientes del enfoque de *Gradient Boosting*, propuesto por Friedman (2001) [61]. Este método se ha consolidado como una herramienta esencial en la ciencia de datos por su equilibrio entre precisión, velocidad y control de la complejidad del modelo mediante regularización explícita [62]. *XGBoost* construye árboles de decisión de forma secuencial, donde cada nuevo árbol corrige los errores de los anteriores. A diferencia del *Gradient Boosting* tradicional, introduce una función objetivo regularizada que equilibra el ajuste a los datos y la simplicidad del modelo, expresada como:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

donde $l(\hat{y}_i, y_i)$ es la función de pérdida y $\Omega(f_k)$ el término de regularización:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

Aquí, T representa el número de hojas y w los pesos asociados. Los hiperparámetros γ y λ controlan la penalización por complejidad y la magnitud de los pesos, ayudando a evitar el sobreajuste. Esta estructura hace de *XGBoost* un modelo más robusto y generalizable que las versiones previas de *Gradient Boosting*.

El aprendizaje se basa en una expansión de segundo orden de la función de pérdida mediante los gradientes (g_i) y hessianos (h_i) de las predicciones previas. Dicha aproximación, expresada como:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

optimiza la función objetivo de forma estable y rápida. Con estos valores, el algoritmo identifica las divisiones que maximizan la ganancia de información dentro de cada árbol, determinando la contribución de cada partición al ajuste global del modelo.

Durante el entrenamiento, *XGBoost* aplica estrategias que fortalecen su rendimiento, como el submuestreo de datos y variables, la ponderación de observaciones y el manejo de valores faltantes. Estos mecanismos, junto con la paralelización en múltiples núcleos o GPU, lo hacen altamente eficiente frente a grandes volúmenes de datos. También incorpora evaluación continua y detención anticipada (*early stopping*) para prevenir sobreentrenamiento. Conceptualmente, *XGBoost* combina la potencia predictiva de los métodos de ensamblaje con una base estadística sólida. Cada árbol introduce una mejora incremental sobre la predicción anterior:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

En esta expresión, $\hat{y}_i^{(t)}$ representa la predicción actualizada para el registro i en la iteración t del proceso de entrenamiento del modelo. Por su parte, $\hat{y}_i^{(t-1)}$ corresponde a la predicción acumulada hasta la iteración anterior, es decir, antes de incorporar el nuevo árbol. El término $f_t(x_i)$ denota la contribución del árbol añadido en la iteración t, el cual se entrena específicamente para corregir los errores residuales que el modelo mantiene hasta ese momento. Finalmente, η es la tasa de aprendizaje que controla cuánta influencia tiene dicho árbol en la actualización final.

En conjunto, el algoritmo logra un equilibrio entre flexibilidad, interpretabilidad y rendimiento, y permite analizar la importancia de las variables mediante métricas de ganancia, frecuencia o cobertura. La Figura 4.3 ilustra este proceso, mostrando cómo *XGBoost* integra las fases de configuración, optimización y predicción.

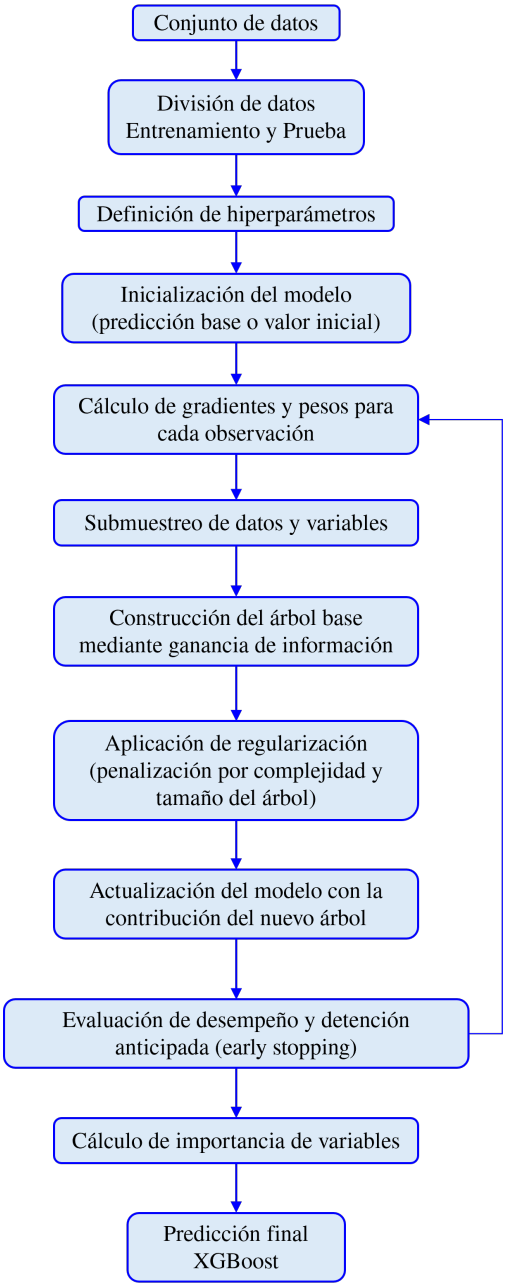

Figura 4.3: Flujograma del modelo *XGBoost*.
Fuente: Elaboración propia.

Métodos no supervisados

En este trabajo, los métodos no supervisados se utilizaron como una estrategia complementaria para identificar patrones irregulares en los consumos de agua. En particular, este enfoque permitió explorar la estructura interna de los datos y detectar agrupamientos naturales en los perfiles de consumo mensual, diferenciando comportamientos típicos de aquellos que se desvían significativamente de la dinámica esperada. Para ello se emplearon características derivadas de las series de tiempo de cada medidor mecánico (como estadísticas robustas, medidas de variabilidad, complejidad temporal y transformaciones simbólicas).

- **DBSCAN:**

  En este trabajo, *DBSCAN* se utilizó como método no supervisado para detectar consumos anómalos sin necesidad de etiquetas, aprovechando su capacidad para identificar puntos aislados o patrones de baja densidad asociados a fallas técnicas o lecturas inusuales. Para ello se aplicó sobre variables derivadas exclusivamente de la serie temporal de cada medidor, como medidas de variabilidad, estadísticos robustos y métricas de complejidad.

  El algoritmo *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*), propuesto por Ester et al. [63], es un método de agrupamiento no supervisado basado en densidad que identifica regiones densas de puntos como grupos (*clusters*) y separa regiones menos densas como ruido o anomalías. A diferencia de los métodos de partición o jerárquicos, como *k-means*, no requiere especificar a priori el número de clústeres, sino que determina las agrupaciones según la densidad local de los datos, definida por dos parámetros: el radio de vecindad ε y el número mínimo de puntos $MinPts$ necesarios para formar un clúster.

  La idea central del algoritmo se fundamenta en la noción de densidad espacial. Dado un punto p, su ε-vecindad se define como el conjunto de puntos dentro de una distancia ε:

  $$N_\varepsilon(p) = \{q \in D \mid dist(p, q) \leq \varepsilon\}$$

  donde D representa el conjunto total de observaciones. Según la cantidad de vecinos dentro de esta región, cada punto se clasifica como núcleo (*core point*), borde (*border point*) o ruido (*noise point*). Un punto núcleo cumple que $|N_\varepsilon(p)| \geq MinPts$; un punto de borde no alcanza esa densidad, pero se encuentra en la vecindad de un punto núcleo; mientras que un punto de ruido no pertenece a ninguna región suficientemente densa.

  A partir de estos criterios, *DBSCAN* expande los clústeres de forma recursiva. El proceso inicia con un punto núcleo y agrega todos los puntos alcanzables por densidad, es decir, aquellos que pueden conectarse a través de una secuencia continua de puntos núcleo cuyas vecindades se solapan dentro del radio ε. De esta manera, dos puntos p y q se consideran densidad-conectados si existe una cadena $p_1, p_2, \ldots, p_n$ tal que $p_1 = p$, $p_n = q$, y $p_{i+1} \in N_\varepsilon(p_i)$ para todo i, cumpliendo además que cada p_i sea un punto núcleo.

  El algoritmo continúa expandiendo regiones densas hasta que todos los puntos hayan sido asignados a un clúster o clasificados como ruido. Este enfoque permite identificar agrupamientos de forma arbitraria, sin restricciones geométricas, y detectar estructuras no lineales dentro de los datos.

  Desde una perspectiva teórica, la fortaleza de *DBSCAN* radica en su capacidad para determinar automáticamente el número de clústeres y manejar datos con ruido o valores atípicos.

Además, no asume una distribución particular en los datos y es robusto frente a transformaciones espaciales o escalas heterogéneas, siempre que se utilice una métrica de distancia apropiada (habitualmente la euclidiana). No obstante, su desempeño depende críticamente de la elección de ε y $MinPts$, parámetros que determinan la resolución de la agrupación: valores demasiado pequeños fragmentan los clústeres, mientras que valores excesivos pueden fusionar regiones distintas.

En el contexto de la analítica de datos espaciales y temporales, *DBSCAN* ha demostrado ser especialmente útil para identificar patrones anómalos, segmentos de comportamiento irregular o agrupamientos con formas complejas. Su implementación eficiente mediante estructuras de indexación espacial como *k-d trees* o *R-trees* reduce la complejidad computacional a $O(n \log n)$, lo que favorece su uso en conjuntos de datos extensos. La Figura 4.4 sintetiza las etapas conceptuales del método, destacando la relación entre los parámetros de densidad, la expansión de clústeres y la detección de ruido.

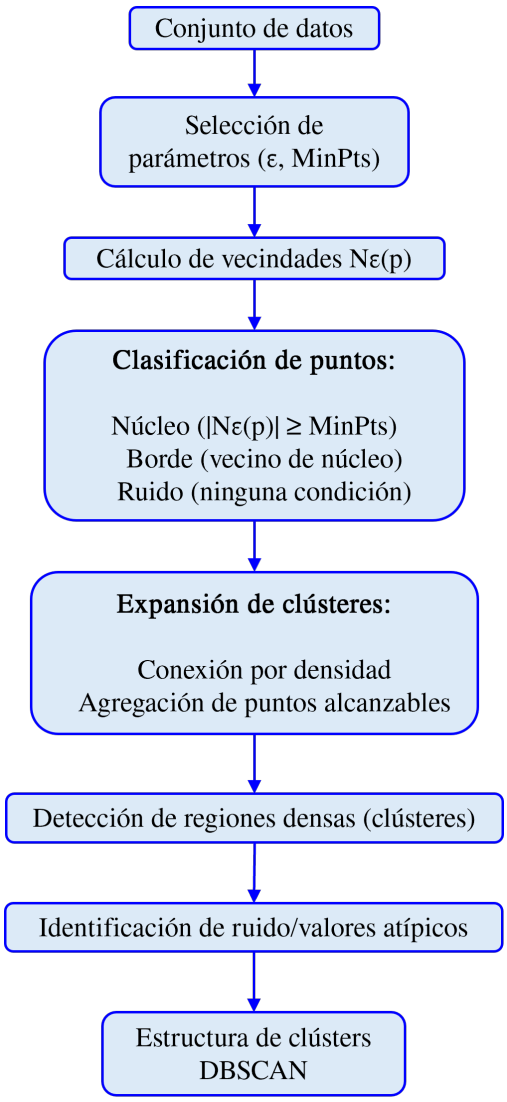

Figura 4.4: Flujograma del modelo *DBSCAN*.
Fuente: Elaboración propia.

Modelo de regresión logística

En el presente trabajo, la regresión logística se empleó como un modelo base para establecer un punto de comparación interpretativo frente a los métodos supervisados más complejos utilizados en la detección de fraude y anomalías técnicas. Su uso permitió analizar la dirección y magnitud de la asociación entre las características derivadas de los consumos mensuales (como medidas de variabilidad, complejidad temporal y atributos metrológicos) y la probabilidad de que un medidor presente un evento irregular. Aunque su capacidad para capturar relaciones no lineales es limitada, aporta un marco analítico valioso para comprender el aporte individual de cada predictor [64].

La regresión logística es una técnica estadística ampliamente utilizada para modelar la relación entre una variable dependiente dicotómica y un conjunto de variables independientes cuantitativas o cualitativas. A diferencia de la regresión lineal, que asume una relación lineal directa entre las variables y puede producir valores fuera del rango de probabilidades, la regresión logística transforma la variable respuesta mediante una función logística que restringe los valores de salida al intervalo $[0, 1]$, lo que permite interpretar el resultado como una probabilidad [64]. El modelo se expresa como:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

donde $P(Y = 1 \mid X)$ representa la probabilidad de que un medidor presente el evento de interés (fraude o anomalía técnica, dependiendo del modelo evaluado) dada la información contenida en el conjunto de predictores X. Aquí, Y es la variable objetivo binaria, donde $Y = 1$ indica la presencia del evento irregular y $Y = 0$ su ausencia. Por su parte, los términos $X_1, X_2, \ldots, X_k$ corresponden a las características utilizadas por el modelo, entre las que se incluyen métricas derivadas de las series temporales de consumo (como *LZC*, *TSLF*, estadísticas robustas y medidas de variabilidad), atributos metrológicos del medidor y variables espaciales. Los coeficientes $\beta_1, \beta_2, \ldots, \beta_k$ cuantifican el aporte de cada predictor al incremento o disminución de la probabilidad de que ocurra el evento. A través de la transformación logit, el modelo puede reescribirse en forma lineal como:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Esta expresión indica que la regresión logística modela el logaritmo de la razón de probabilidades (*odds*) como una combinación lineal de las variables independientes [65]. Los coeficientes estimados (β_i) pueden interpretarse en términos de *odds ratios*, que cuantifican cómo varía la probabilidad del evento ante un cambio unitario en la variable correspondiente, manteniendo constantes las demás.

Para garantizar la validez de los resultados, la regresión logística debe cumplir ciertos supuestos estadísticos:

- **Independencia de las observaciones:** cada observación debe ser independiente de las demás, lo que implica que no existan repeticiones ni relaciones jerárquicas entre los datos.
- **Linealidad en el logit:** aunque la relación entre los predictores y la variable dependiente no necesita ser lineal, se requiere que el logit (log de la razón de probabilidades) sea lineal respecto a las variables predictoras continuas. Esta condición puede verificarse mediante la prueba de Box–Tidwell o mediante el análisis gráfico de los residuos parciales.
- **Ausencia de multicolinealidad:** las variables independientes no deben estar altamente correlacionadas entre sí. Este supuesto se evalúa mediante el Variance Inflation Factor (VIF) o la matriz de correlaciones.

- **Tamaño muestral adecuado:** se recomienda contar con al menos 10 eventos por predictor para garantizar la estabilidad de los coeficientes estimados [66].

El incumplimiento de estos supuestos puede afectar la estabilidad del modelo, su capacidad predictiva o la interpretación de los coeficientes.

La validez del modelo de regresión logística se evalúa en varias etapas complementarias:

- **Bondad de ajuste global:** el estadístico $-2LL$ (log-verosimilitud negativa) mide el ajuste global del modelo; una disminución significativa al comparar el modelo completo con el modelo nulo indica un mejor desempeño.
- **Prueba de Hosmer–Lemeshow:** contrasta las probabilidades predichas por el modelo con las proporciones observadas en los datos reales, evaluando así la calidad de la calibración. En este trabajo se adopta un nivel de significancia de 0,05, criterio ampliamente utilizado en muestras pequeñas o moderadas debido a que proporciona un equilibrio adecuado entre el riesgo de rechazar incorrectamente un modelo bien calibrado (error tipo I) y la potencia estadística de la prueba. Bajo este umbral, un valor de $p > 0,05$ indica que no existen discrepancias significativas entre las probabilidades estimadas y las observadas, indicando que el modelo presenta un ajuste aceptable a los datos [64].
- **Detección de valores influyentes:** el análisis de medidas como *Cook's Distance* o *DFBETAs* identifica observaciones que tienen un impacto desproporcionado sobre los coeficientes estimados.

Una vez validado el ajuste del modelo, se evalúa su capacidad de discriminación entre clases. Para ello se utilizan las siguientes métricas:

- **Curva ROC (Receiver Operating Characteristic):** representa la sensibilidad frente a 1− especificidad para distintos umbrales de decisión.
- **Área bajo la curva (AUC):** resume el poder discriminativo del modelo; valores cercanos a 1 indican alta capacidad de clasificación.
- **Matriz de confusión:** permite calcular indicadores como precisión, recall, especificidad, F1-score y Balanced Accuracy, que reflejan la proporción de aciertos y errores en las predicciones.

En síntesis, la regresión logística constituye un modelo estadístico sólido, interpretable y ampliamente utilizado en contextos donde se requiere explicar o predecir la ocurrencia de un evento binario. En el presente trabajo, su uso resulta pertinente porque tanto la detección de fraude como la identificación de anomalías se formulan como problemas en los que en la variable objetivo 1 indica la presencia del evento de interés y 0 su ausencia. Aunque los modelos principales del estudio incluyen algoritmos de aprendizaje automático más complejos (como Random Forest, XGBoost, DBSCAN, entre otros), la regresión logística se emplea como modelo para contrastar los resultados, evaluar la coherencia de las estimaciones y analizar la dirección y estabilidad de las relaciones entre las variables predictoras y la probabilidad del evento [64], [65], [67].

Estas variables incorporan características derivadas de las series de tiempo de consumo (métricas estadísticas, representaciones simbólicas como SAX, medidas de complejidad como LZC y transformaciones logarítmicas), junto con atributos técnicos del medidor. En conjunto, la regresión logística aporta un marco de interpretación complementario que permite comparar los patrones identificados por los modelos más avanzados y valorar la consistencia global de los hallazgos.

Modelos espaciales y autocorrelación

En el análisis estadístico tradicional, se asume que las observaciones son independientes entre sí; sin embargo, en muchos fenómenos geográficos o territoriales esta suposición no se cumple, ya que las observaciones cercanas tienden a estar relacionadas. La *estadística espacial* surge precisamente para describir, modelar y cuantificar este tipo de dependencia espacial, conocida como *autocorrelación espacial*. Dicho concepto refleja el grado en que el valor de una variable en una ubicación geográfica está correlacionado con los valores en ubicaciones vecinas. Cuando las unidades espaciales exhiben agrupamientos de valores similares, se presenta autocorrelación positiva; mientras que la alternancia sistemática de valores altos y bajos indica autocorrelación negativa. Si los valores se distribuyen aleatoriamente en el territorio, no existe autocorrelación espacial [68].

Para cuantificar este comportamiento se emplean indicadores globales y locales de asociación espacial. Entre los primeros, los más utilizados son el índice de Moran (I) y el índice de Geary (G), los cuales tradicionalmente evalúan la dependencia espacial presente en todo el conjunto de datos. No obstante, en el contexto del presente trabajo de grado, estos indicadores no se aplicaron directamente sobre la variable de fraude o de anomalías, sino sobre los residuos del modelo de regresión logística, con el fin de verificar el supuesto de independencia espacial de los errores. Este procedimiento permite determinar si la estructura espacial del territorio influye en el desempeño del modelo y si los residuos muestran patrones espaciales que podrían indicar variables omitidas o especificaciones inadecuadas.

En el contexto de este trabajo, el índice de Moran permite evaluar si los valores asociados a una variable georreferenciada (como la prevalencia de fraude o la presencia de anomalías técnicas por zona) presentan patrones de agrupamiento espacial o si, por el contrario, se distribuyen de manera aleatoria en el territorio.
En la expresión:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2},$$

cada término se interpreta de la siguiente forma:

- N es el número total de unidades espaciales analizadas (por ejemplo, localidades o zonas donde se agregan los medidores).
- x_i es el valor observado de la variable de interés en la unidad espacial i; en este trabajo corresponde, según el análisis, a la prevalencia de fraude o a la prevalencia de anomalías en la localidad i.
- $\bar{x}$ es el valor promedio de esa variable en todo el conjunto de unidades espaciales.
- w_{ij} es el peso espacial que cuantifica el nivel de vecindad entre las unidades i y j. En este proyecto, estos pesos provienen de la matriz de contigüidad que indica si dos localidades son adyacentes (vecinas) o no.
- $W = \sum_i \sum_j w_{ij}$ es la suma total de los pesos espaciales y actúa como factor de normalización del índice.

En conjunto, el numerador captura la covarianza espacial entre localidades vecinas (es decir, si zonas cercanas presentan valores similares de fraude o anomalías) mientras que el denominador representa la variabilidad total de la variable en el territorio. De esta manera, el índice de Moran permite determinar si el fenómeno irregular analizado tiende a agruparse geográficamente, dispersarse o distribuirse al azar.

El índice I toma valores aproximados entre -1 y 1: valores positivos indican agrupamiento de valores similares (autocorrelación positiva), valores negativos indican alternancia de valores diferentes (autocorrelación negativa), y valores cercanos a 0 reflejan aleatoriedad espacial. El valor esperado bajo la hipótesis nula de aleatoriedad espacial es $E(I) = -1/(N-1)$, y su significancia estadística puede evaluarse mediante pruebas de permutación o mediante un estadístico z estandarizado.

En este trabajo, el índice de Geary se emplea como una medida complementaria al índice de Moran para evaluar la autocorrelación espacial de la prevalencia de fraude y de las anomalías técnicas entre localidades. A diferencia de Moran, que captura patrones globales de similitud, el índice de Geary enfatiza las diferencias locales entre unidades vecinas, permitiendo detectar zonas donde el comportamiento del fenómeno cambia bruscamente respecto a sus alrededores.
Su formulación es:

$$C = \frac{(N-1)}{2W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

y cada uno de sus componentes se interpreta de la siguiente manera:

- N es el número total de unidades espaciales analizadas, es decir, el total de localidades consideradas en el estudio.
- x_i es el valor de la variable de interés en la localidad i, correspondiente en este trabajo a la prevalencia de fraude o a la prevalencia de anomalías técnicas.
- x_j es el valor de esa misma variable en la localidad vecina j.
- $\bar{x}$ es el valor promedio de la variable en todas las localidades.
- w_{ij} es el peso espacial que indica el grado de vecindad entre las localidades i y j, derivado de la matriz de contigüidad empleada en el análisis.
- $W = \sum_i \sum_j w_{ij}$ es la suma total de los pesos espaciales, que actúa como factor de normalización.

El numerador cuantifica las diferencias locales entre localidades adyacentes, elevando al cuadrado la diferencia $(x_i - x_j)$ para resaltar discontinuidades espacialmente próximas. El denominador, por su parte, captura la variabilidad global de la variable. En conjunto, el índice de Geary permite identificar zonas donde la prevalencia de fraude o anomalías difiere marcadamente de sus vecinas, lo cual resulta útil para reconocer límites espaciales, transiciones abruptas o focos muy localizados del fenómeno.

Al igual que el índice de Moran, C se interpreta en relación con la independencia espacial. Un valor $C < 1$ indica autocorrelación positiva (valores similares en ubicaciones cercanas), $C > 1$ refleja autocorrelación negativa (diferencias entre vecinos), y $C \approx 1$ indica ausencia de patrón espacial. A diferencia del índice de Moran, que capta la asociación global, el índice de Geary es más sensible a las variaciones locales, por lo que ambos se consideran complementarios en la caracterización de la estructura espacial de los datos [69].

En conjunto, los índices de Moran y Geary proporcionan una base teórica sólida para evaluar la presencia de dependencias espaciales en los fenómenos estudiados en este trabajo, específicamente fraude y anomalías técnicas en el consumo de agua. En lugar de aplicarse directamente sobre los valores de consumo, en este estudio estos índices se emplearon sobre los *residuos estandarizados de los modelos de regresión logística*, con el fin de verificar si, una vez modelados los factores individuales de cada

cliente, persiste algún patrón espacial no explicado por el modelo.

La detección de autocorrelación espacial en los residuos permite identificar si las predicciones del modelo tienden a concentrarse en ciertas zonas geográficas, ya sea por dinámicas territoriales, condiciones socioeconómicas compartidas o posibles focos localizados de fraude o fallas técnicas. Si los residuos presentan agrupamiento espacial significativo, esto sugiere que la regresión logística no captura completamente la estructura espacial subyacente, lo que abriría la necesidad de incorporar métodos espaciales más complejos o modelos con términos espaciales explícitos. En consecuencia, el análisis mediante Moran y Geary constituye un insumo esencial para evaluar la validez del modelo predictivo y para interpretar adecuadamente la distribución territorial del fraude y las anomalías.

Modelos Aditivos Generalizados (GAM)

Los Modelos Aditivos Generalizados (GAM) constituyen una extensión flexible de los Modelos Lineales Generalizados (GLM), que permite capturar relaciones no lineales entre la variable dependiente y los predictores mediante la inclusión de funciones suavizadas [70]. Mientras que los modelos lineales tradicionales asumen que el efecto de cada variable explicativa sobre la respuesta es estrictamente lineal, los GAM relajan esta restricción al expresar la relación de manera aditiva y no paramétrica:

$$g(\mathbb{E}[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

donde $g(\cdot)$ es la función de enlace, β_0 es el intercepto y $f_i(X_i)$ son funciones suavizadas que representan los efectos potencialmente no lineales de las variables predictoras. Estas funciones se estiman habitualmente mediante *splines* o técnicas de suavizamiento local, lo que permite modelar patrones complejos sin imponer una forma funcional específica [71].

La principal ventaja de los GAM radica en su capacidad para equilibrar la flexibilidad y la interpretabilidad. Al mantener una estructura aditiva, el modelo permite examinar el efecto parcial de cada predictor de forma separada, lo que facilita la comprensión de su influencia sobre la variable de respuesta. Además, esta característica evita el sobreajuste que puede presentarse en modelos completamente no paramétricos, al mismo tiempo que mejora la capacidad de ajuste frente a modelos lineales rígidos.

En el contexto del presente trabajo, los GAM resultan relevantes como referencia metodológica, ya que permiten comprender cómo las relaciones no lineales entre las variables derivadas de la serie de consumo como las transformaciones estadísticas, medidas de complejidad (LZC), representaciones simbólicas (SAX) y métricas de variabilidad pueden influir en la probabilidad de ocurrencia de anomalías o de fraude. Aunque los modelos principales implementados en este estudio corresponden a algoritmos de Machine Learning (como Random Forest y XGBoost para fraude) y métodos no supervisados (como DBSCAN para anomalías), los GAM ofrecen un marco conceptual útil para interpretar la posible estructura no lineal que dichos modelos detectan de manera automática. Su capacidad para descomponer el efecto parcial de cada predictor permitiría, por ejemplo, caracterizar tendencias, estacionalidades o patrones irregulares en el consumo que son consistentes con los hallazgos obtenidos mediante los modelos empleados en este proyecto.

Así, los GAM complementan el análisis realizado en este trabajo al proporcionar un enfoque que combina flexibilidad y capacidad explicativa, y que resulta coherente con la naturaleza compleja y no lineal de los fenómenos de fraude y anomalías en el consumo de agua.

Criterios de información AIC - BIC

En el análisis estadístico y la selección de modelos, los criterios de información desempeñan un papel fundamental para evaluar el equilibrio entre la calidad del ajuste y la complejidad del modelo. Entre los más utilizados se encuentran el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC), los cuales permiten comparar modelos alternativos y seleccionar aquel que logra un compromiso óptimo entre precisión y parsimonia [72], [73].

En el contexto de este trabajo de grado, estos criterios se utilizaron para comparar los distintos modelos de regresión logística estimados tanto para la detección de fraude (modelos M1–M6) como para la detección de anomalías (modelos M1–M5), tal como se reporta en las Tabla 7.33 y Tabla 8.18 de los resultados. Su objetivo fue identificar cuál modelo ofrecía un mejor equilibrio entre buen ajuste y menor complejidad, permitiendo evaluar la parsimonia de cada alternativa. En todos los casos, valores más bajos de AIC y BIC indicaron un mejor desempeño relativo entre los modelos comparados acompañado también del cumplimiento de los supuestos que este modelo exige.

El Criterio de Información de Akaike (AIC) se define como:

$$AIC = -2\ln(\hat{L}) + 2k$$

donde $\hat{L}$ es el valor máximo de la función de verosimilitud del modelo y k representa el número de parámetros estimados. El primer término mide el grado de ajuste del modelo a los datos, mientras que el segundo introduce una penalización por complejidad para evitar el sobreajuste. Un menor valor de AIC indica un modelo con mejor equilibrio entre ajuste y simplicidad.

Por su parte, el Criterio de Información Bayesiano o de Schwarz se expresa como:

$$BIC = -2\ln(\hat{L}) + k\ln(n)$$

donde n es el tamaño de la muestra. A diferencia del AIC, el BIC impone una penalización más severa a los modelos con muchos parámetros, lo que lo hace más conservador en la selección. En ambos casos, los criterios son relativos: no tienen significado absoluto, sino que se utilizan para comparar un conjunto de modelos estimados bajo los mismos datos y supuestos, seleccionando aquel con el valor más bajo.

Tanto el AIC como el BIC son medidas utilizadas para comparar modelos estadísticos considerando simultáneamente su capacidad de ajuste y su complejidad. Ambos criterios penalizan la inclusión de parámetros adicionales, de modo que valores más pequeños indican un mejor equilibrio entre ajuste y parsimonia. En general, un modelo con AIC o BIC más bajo se interpreta como superior a uno con valores más altos, aunque el BIC aplica una penalización más estricta por complejidad, por lo que tiende a favorecer modelos más simples, especialmente cuando el tamaño de la muestra es grande. Así, la comparación relativa (y no los valores absolutos) permite determinar qué modelo ofrece la mejor explicación de los datos sin sobreajuste.

En el contexto de la analítica de datos y la modelación predictiva, estos criterios resultan esenciales para comparar modelos con diferentes combinaciones de variables o estructuras, como los Modelos Aditivos Generalizados (GAM), los modelos de regresión logística o los modelos de series temporales. Su aplicación contribuye a garantizar que el modelo final no sólo reproduzca adecuadamente los datos observados, sino que también mantenga una estructura interpretable y generalizable a nuevos escena-

rios. Finalmente, en el presente trabajo trabajo, el AIC y BIC fueron claves para diagnosticar la validez estructural de los modelos: en fraude permitieron seleccionar las mejores especificaciones preliminares (M2 y M3), mientras que en anomalías confirmaron que, aunque el modelo M5 presentaba los menores valores, dichos resultados debían interpretarse solo de manera descriptiva debido al incumplimiento del supuesto de linealidad en el logit.

4.2. ANTECEDENTES

4.2.1. Fraude

El uso de técnicas de minería de datos y aprendizaje automático para la detección de fraudes en el consumo de agua ha cobrado relevancia en los últimos años, impulsado por la necesidad de reducir pérdidas no técnicas (NRW) y optimizar la gestión de los sistemas de distribución [74], [75]. Estos enfoques se han consolidado como alternativas eficaces frente a las inspecciones manuales, al permitir el análisis masivo de registros históricos y la identificación temprana de comportamientos anómalos. A partir de diversos estudios desarrollados en contextos internacionales, se observa una evolución metodológica que combina el procesamiento estadístico de datos de consumo con estrategias de modelación predictiva y, más recientemente, con herramientas basadas en visión por computador [76].

En la literatura se destaca el avance de modelos estructurados bajo metodologías de descubrimiento de conocimiento, tales como Knowledge Discovery in Databases (KDD) o Cross-Industry Standard Process for Data Mining (CRIPS-DM), con fases de selección, transformación y modelado. Bajo este enfoque, se construyen variables derivadas del comportamiento de los usuarios (periodos sin consumo, variabilidad y promedios históricos) que actúan como predictores de patrones irregulares [77], [78]. Los resultados muestran que clasificadores como árboles de decisión, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y redes neuronales alcanzan buenos niveles de precisión y priorización de casos, además de permitir reglas interpretables o perfiles de riesgo útiles para la operación [77], [78].

Paralelamente, se han propuesto enfoques que amplían la perspectiva más allá de la facturación. Un caso representativo incorpora visión por computador para identificar manipulación física en medidores (p. ej., sellos de seguridad) a partir de imágenes de terreno, utilizando descriptores basados en Histogram of Oriented Gradients (HOG) y clasificadores como Optimum-Path Forest (OPF) [76]. Esta línea sugiere integrar fuentes visuales, técnicas y de consumo en un mismo sistema para robustecer el diagnóstico.

También se reportan estudios a gran escala donde SVM y KNN se aplican sobre bases extensas y desbalanceadas, con CRISP-DM como guía de preparación, modelado y evaluación, y técnicas de muestreo/normalización para lidiar con el desbalance [78]. En Brasil, aproximaciones con regresión logística y random forest han mostrado sensibilidades elevadas (mayores al 80 %) y utilidades operativas claras para priorizar inspecciones y caracterizar factores asociados a reconexiones no autorizadas, bypass y acometidas directas [79], [80]. Entre los factores más relevantes figuran antecedentes de infracciones, reincidencias, periodos prolongados de consumo nulo y caídas abruptas [79].

A pesar de los avances, gran parte de la evidencia se basa en datos históricos provenientes de medidores mecánicos con lectura manual, lo que limita la incorporación de información técnica y espacial del dispositivo [75]. En este sentido, el proyecto actual propone integrar atributos técnicos del medidor (marca, diámetro, clase metrológica, número de ruedas, tipo de transmisión) y georreferenciación,

ampliando la capacidad explicativa y diferenciando mejor entre fallas instrumentales y comportamientos fraudulentos, en línea con las recomendaciones más recientes para fortalecer la gestión de pérdidas no técnicas [74].

4.2.2. Anomalías

La detección de anomalías asociadas a fallas en los medidores de agua se ha convertido en un campo clave dentro de la gestión analítica de redes de distribución. Estos métodos buscan identificar desviaciones en los patrones de consumo que no obedecen al comportamiento del usuario, sino a errores instrumentales como atascamientos, lecturas planas, deriva metrológica o fallas en la transmisión de datos. Con el aumento de la digitalización y la disponibilidad de lecturas de alta frecuencia, los enfoques basados en datos han permitido reemplazar la inspección manual por modelos predictivos capaces de reconocer comportamientos anómalos en tiempo real.

En el contexto de la analítica aplicada al monitoreo de medidores, los algoritmos de detección no supervisada han demostrado gran utilidad al aprender las regularidades estadísticas del consumo y señalar los registros que se alejan de ellas. Modelos como el *Isolation Forest*, el *Local Outlier Factor* o los autoencoders identifican observaciones con densidad atípica o con errores de reconstrucción elevados, lo que permite aislar posibles fallas del medidor sin requerir etiquetas previas. Este enfoque ha mostrado un buen desempeño en la identificación de lecturas planas prolongadas, picos abruptos y caídas súbitas, que suelen indicar desgaste mecánico o interrupciones en el flujo de lectura [81].

Paralelamente, los métodos supervisados han incorporado arquitecturas de ensamble y técnicas de *gradient boosting* para combinar la capacidad predictiva de distintos modelos y aumentar la sensibilidad ante comportamientos inusuales. Estas estrategias procesan variables tanto de consumo como del propio dispositivo (antigüedad, tipo de transmisión, diámetro o clase metrológica) para diferenciar entre anomalías operativas y fallas instrumentales. Su integración en sistemas de medición inteligente (*smart metering*) permite la generación de alertas tempranas y la actualización continua de los modelos mediante flujos de datos en línea [82].

Además, los avances en aprendizaje estadístico han permitido implementar esquemas de monitoreo híbridos que combinan el análisis temporal con indicadores derivados del comportamiento histórico. El cálculo de residuos entre el consumo esperado y el observado, junto con el uso de ventanas móviles, facilita detectar desviaciones sostenidas que no responden a cambios de hábito del usuario. Estas técnicas, aplicadas sobre datos mensuales o diarios, han mostrado su utilidad para identificar medidores defectuosos o con lecturas inconsistentes en sistemas residenciales [83].

En conjunto, los enfoques de detección de anomalías orientados a fallas de medidores representan una convergencia entre la ingeniería metrológica y la ciencia de datos. La incorporación de algoritmos no supervisados, modelos de ensamble y análisis temporal permite no solo detectar lecturas anómalas, sino también diagnosticar el origen de la desviación, priorizar intervenciones técnicas y mejorar la confiabilidad de los sistemas de medición en redes de agua potable.

5 METODOLOGÍA

DATOS: PREPARACIÓN Y ANÁLISIS

5.1. BASES SUMINISTRADAS

Se trabajó con dos conjuntos principales de información provistos por la Empresa de Servicios Sanitarios del Bío-Bío (Essbio). En primer lugar, se consolidó la Base A, que sirvió como fuente principal para el modelado orientado a la detección de fraude. Sobre esta base se ejecutaron los procedimientos de limpieza, depuración, imputación y selección de variables descritos en las secciones siguientes.

A partir de la Base A se generó la Base B, conformada únicamente por los registros sin presencia de fraude. Este subconjunto se empleó en la etapa de modelación destinada a la detección de anomalías. Dado su origen común, la Base B heredó los mismos procesos de tratamiento aplicados sobre la base principal, garantizando la coherencia y comparabilidad entre ambas.

La preparación incluyó la identificación de las variables relevantes para los objetivos del estudio, priorizando aquellas relacionadas con el consumo de agua, las características físicas del medidor y los indicadores de fraude y anomalías. Este proceso permitió reducir la dimensionalidad inicial de las bases y estructurar un conjunto de datos limpio y consistente para las etapas posteriores de análisis y modelado.

5.2. TRATAMIENTO DE DATOS

El tratamiento de datos comprendió las etapas de limpieza, depuración y validación de la información utilizada para el análisis. Inicialmente, se integraron las distintas bases de datos suministradas por la empresa, asegurando la correspondencia entre identificadores de cliente, periodos de medición y variables técnicas del medidor. Posteriormente, se eliminaron registros duplicados, inconsistentes o sin relevancia analítica, así como aquellos con valores sin sentido físico. Adicionalmente, se realizó la georreferenciación de los registros mediante la transformación de las direcciones en coordenadas geográficas (latitud y longitud), proceso que permitió incorporar información espacial a cada observación y que resultó fundamental para el análisis de densidad de kernel y para los modelos de detección de fraude y anomalías desarrollados en etapas posteriores.

5.2.1. Limpieza de datos

Se aplicaron procedimientos de identificación y depuración de registros inconsistentes con el objetivo de garantizar la calidad del conjunto de análisis. En primer lugar, se implementó una rutina de verificación de duplicados a partir del identificador único de cliente, con la cual se detectaron y eliminaron los registros repetidos, conservando únicamente una observación por ID. Posteriormente, se realizó una revisión de los tipos de medidor presentes en la base de datos. De acuerdo con las especificaciones técnicas de la empresa proveedora, se excluyeron los registros correspondientes a medidores clasificados como ZM-MED FICTICIO, dado que no representan mediciones reales de consumo.

5.2.2. Datos faltantes

Durante el tratamiento de datos se aplicaron procedimientos para identificar y corregir registros con información incompleta, tanto en las series de consumo como en las variables categóricas asociadas a los medidores. La detección de valores faltantes se realizó mediante un control de completitud por usuario y variable, verificando la continuidad de las observaciones temporales.

Para los casos de series de consumo, se aplicaron métodos de imputación que conservaron la estructura y tendencia de los datos. Se utilizó *interpolación lineal* cuando la cantidad de valores ausentes consecutivos fue reducida y existía suficiente información adyacente, mientras que se empleó ajuste por *spline* cuando la ausencia de datos fue más extensa pero mantenía coherencia temporal.

En el caso de la marca (variable categórica), los valores faltantes fueron completados mediante imputación por moda. Esta elección se sustenta en que los análisis preliminares mostraron que la ausencia de dicha información ocurría de manera aleatoria, sin evidenciar patrones asociados a otras variables (MCAR) [84]. Bajo este supuesto, imputar con la categoría más frecuente resulta apropiado para conservar la coherencia de las distribuciones y la consistencia de las categorías observadas. Estas acciones permitieron mantener la integridad y continuidad del conjunto de datos antes del análisis exploratorio y del modelado.

5.2.3. Datos erróneos

Se implementaron procedimientos para la detección y corrección de errores en las variables técnicas y en las series de consumo. En primer lugar, se revisaron las clasificaciones de los diámetros de los medidores para asegurar su coherencia con la normativa técnica de referencia. Los registros con valores no estandarizados fueron ajustados a la medida equivalente aceptada con el fin de mantener la consistencia en la base de datos.

Adicionalmente, se identificaron valores negativos de consumo, considerados físicamente imposibles y atribuibles a errores de captura o procesamiento. Estos registros fueron corregidos mediante *interpolación lineal*, técnica que permitió reemplazar los valores erróneos por estimaciones coherentes con la tendencia temporal de cada serie, evitando la generación de discontinuidades o distorsiones abruptas en los datos.

5.2.4. Georreferenciación de registros

Como parte del tratamiento de datos, se incluyó la transformación de las direcciones en coordenadas geográficas (latitud y longitud), siguiendo la metodología propuesta por Troncoso et al. (2021) [85].

Este proceso permitió georreferenciar los registros y disponer de información espacial asociada a cada observación. Para ello, se implementó un procedimiento de geocodificación utilizando la biblioteca *geopy* y el geocodificador GoogleV3, configurado con una clave de API de Google Maps. En cada registro del *DataFrame* se extrajo la dirección completa y se ejecutó el método `geocode` con un `timeout` de 60 segundos, lo que permitió mitigar interrupciones derivadas de latencia o sobrecarga del servicio.

Las coordenadas obtenidas fueron incorporadas al conjunto de datos y se utilizaron posteriormente en el análisis de densidad de kernel del estudio multivariado y en los modelos de detección de fraude y anomalías. Por este motivo, su cálculo se incluyó dentro de las etapas de tratamiento y preparación de datos, garantizando la consistencia espacial de la información desde las primeras fases del análisis.

5.3. ANÁLISIS EXPLORATORIO

Una vez completadas las etapas de limpieza, imputación y validación del conjunto de datos, se realizó un análisis exploratorio con el propósito de examinar la estructura y el comportamiento de las variables resultantes. Esta fase permitió verificar la efectividad del tratamiento aplicado, identificar patrones generales, distribuciones y relaciones entre variables, así como detectar posibles valores atípicos o inconsistencias residuales antes de proceder con las etapas de modelado.

5.3.1. Estadísticas descriptivas antes y después del tratamiento de datos

En esta fase se aplicaron estadísticas descriptivas comparativas para las variables numéricas y categóricas, con el propósito de evaluar el impacto del proceso de depuración y tratamiento de datos sobre la estructura original de la base. Para las variables numéricas, se calcularon medidas de tendencia central, dispersión y posición (como la media, mediana, desviación estándar, valores mínimos y máximos, y percentiles 25 %, 50 % y 75 %) antes y después del tratamiento. En el caso de las variables categóricas, se determinaron la moda, la frecuencia absoluta y relativa, el número de categorías únicas y la entropía estandarizada, también en ambos momentos.

5.3.2. Análisis univariado

En el análisis univariado de las variables numéricas, se calcularon medidas de tendencia central (media, mediana y moda), de dispersión (desviación estándar, rango y coeficiente de variación) y de posición (cuartiles). Estas métricas facilitaron la descripción del comportamiento general de los datos y la detección de posibles valores atípicos o concentraciones extremas. En el caso de las variables categóricas, se elaboraron distribuciones de frecuencia absoluta y relativa para cada categoría, junto con el cálculo de la entropía estandarizada, utilizada como indicador del nivel de diversidad o uniformidad en la distribución.

El procedimiento incluyó la comparación de estas distribuciones entre la base de datos de fraude y la base de anomalías, con el fin de establecer similitudes o diferencias estructurales entre ambas. Las tablas generadas presentan los resultados organizados en dos columnas (a) y (b), correspondientes a cada base de análisis, lo que permitió una lectura comparativa.

5.3.3. Análisis multivariado

Para las variables numéricas y categóricas, se efectuó un cruce de frecuencias entre cada una y las variables de respuesta (fraude o anomalías según el caso), determinando el número y porcentaje de registros de cada categoría asociados a casos de fraude y anomalías. Esta estrategia permitió analizar la distribución relativa de los casos en cada grupo, así como visualizar de forma comparativa la proporción de registros con y sin incidencia.

En el caso de la variable *Localidad*, el análisis se amplió mediante el cálculo de la prevalencia directa de fraude y anomalías. En el contexto de este trabajo, esta prevalencia corresponde a la proporción de medidores ubicados en cada localidad que presentan una clasificación positiva según el modelo (ya sea fraude o anomalía), en relación con el total de medidores instalados en esa misma zona. Es decir, la prevalencia es relativa al conjunto total de medidores de la zona y no al número absoluto de casos. De este modo, dos localidades con cantidades muy distintas de medidores pueden compararse de manera estandarizada, ya que la medida refleja el riesgo relativo dentro de cada área geográfica.

De forma complementaria, para esta misma variable se aplicó un análisis espacial de densidad kernel, con el propósito de representar la distribución geográfica de los registros y estimar la intensidad espacial de los casos de fraude y anomalías. A partir de las coordenadas geográficas de los medidores, se generaron mapas de densidad continua mediante un estimador de densidad kernel. En este trabajo, la densidad corresponde al número de casos detectados (fraude o anomalía) suavizado espacialmente según la proximidad entre puntos: cada registro positivo aporta mayor intensidad en su vecindad inmediata, y la contribución disminuye a medida que aumenta la distancia. De esta manera, áreas con más casos producen valores de densidad más altos.

Los colores del mapa representan esta intensidad derivada del conteo de casos, donde los tonos rojos indican zonas con mayor concentración relativa de eventos y los tonos verdes señalan áreas con poca presencia. El cálculo se realizó sobre la superficie urbana cubierta por los medidores, aplicando un parámetro de suavizamiento (bandwidth) que define el radio de influencia (en este proyecto 1 km) con el cual cada caso contribuye al mapa de densidad.

5.3.4. Análisis de consumos

Se construyeron series de tiempo mensuales a partir de los registros de consumo agregados, abarcando desde febrero de 2008 hasta septiembre de 2013. Estas series fueron sometidas a un proceso de suavizado y descomposición aditiva, que permitió aislar y representar tres componentes fundamentales: la tendencia, la estacionalidad y los residuos. La tendencia reflejó el comportamiento general de largo plazo; la estacionalidad describió fluctuaciones periódicas recurrentes, y los residuos capturaron las variaciones no explicadas por los componentes anteriores.

Luego, se aplicaron procedimientos de detección de valores anómalos sobre la serie total y por cliente, con el fin de excluir registros extremos que distorsionaran el comportamiento global del consumo, los cuales se identificaron mediante el test de Tukey [28]. Adicionalmente, se estimaron medidas descriptivas mensuales (media, mediana, cuartiles, coeficiente de variación, porcentaje de valores atípicos y porcentaje de registros anómalos), que permitieron caracterizar la estabilidad y variabilidad del consumo en el tiempo. Dichos indicadores se organizaron en tablas cronológicas para facilitar la evaluación de patrones y la identificación de meses con comportamiento irregular.

5.4. EXTRACCIÓN DE CARACTERÍSTICAS A PARTIR DE LAS SERIES DE CONSUMO

Con el fin de capturar patrones relevantes, reducir la dimensionalidad y facilitar la interpretación de los comportamientos asociados al uso del servicio, se realizó un proceso de extracción de características a partir de las series de consumo analizadas en la sección anterior. Este paso permitió transformar la información temporal en variables representativas que servirán como base para las etapas posteriores de modelado de fraude y anomalías.

5.4.1. Cambios de símbolo y Entropía SAX

Las variables *Cambios de símbolo* y *Entropía SAX* se calcularon a partir de la representación simbólica de las series de consumo generada mediante el método Symbolic Aggregate Approximation (SAX), siguiendo la metodología propuesta por Yan et al. (2022) [45]. Para este proceso se utilizó la configuración `n bins = 5` y `strategy = 'quantile'`, la cual permitió transformar cada serie en una secuencia simbólica compacta, asegurando una distribución equitativa de los símbolos y preservando las variaciones relevantes del consumo.

La variable *Cambios de símbolo* corresponde al número total de transiciones entre símbolos distintos en la secuencia SAX y permitió cuantificar la estabilidad o irregularidad del consumo mensual. Por su parte, la *Entropía SAX* mide la dispersión de los símbolos en la secuencia y refleja el nivel de aleatoriedad en los patrones de consumo. Ambas métricas se emplearon para caracterizar la variabilidad temporal de las series y facilitar la detección de comportamientos atípicos.

5.4.2. Lempel–Ziv Complexity y Time Series Length Factor

Siguiendo la metodología propuesta por Ghamkhar et al. (2023) [43], se calcularon las métricas *Time Series Length Factor (TSLF)* y *Lempel–Ziv Complexity (LZC)* como características para representar la estructura temporal de las series de consumo. En primer lugar, se estimó el *TSLF* a partir de las series originales para cuantificar la longitud efectiva y la estabilidad del consumo de cada cliente. Posteriormente, se aplicó una normalización mín–máx que transformó los valores de consumo mensual al rango [0,1], con el fin de garantizar la comparabilidad entre usuarios y preparar los datos para el cálculo de la *LZC*. Esta métrica se obtuvo en dos versiones: absoluta y normalizada; sin embargo, para el modelado se empleó únicamente la versión normalizada, conforme a la metodología original. La LZC se calculó en cuatro configuraciones independientes (`2, 4, 9` y `99 bins`) para capturar la complejidad del consumo a distintos niveles de granularidad.

5.4.3. Variables estadísticas complementarias

Se calcularon diversas variables estadísticas a partir de las series de consumo con el objetivo de describir su comportamiento general y complementar las métricas de complejidad temporal. Estas variables cuantificaron aspectos de tendencia central, dispersión, forma de la distribución y presencia de valores atípicos. Se incluyeron la media (*Media consumo*), desviación estándar (*SD consumo*), coeficiente de variación (*CV consumo*), mediana (*Mediana consumo*), valor máximo (Máx. *consumo*), valor mínimo (*Mín. consumo*) y rango (*Rango consumo*). También se calcularon medidas de forma como la asimetría (*Asimetría consumo*) y la curtosis (*Curtosis consumo*), junto con indicadores asociados al comportamiento del consumo, como la proporción de registros nulos (*Consumo ceros*), el número de valores atípicos moderados (*N° atípicos moderados*) y número de atípicos extremos (*N° atípicos*

moderados), los valores z normalizados (*Z máx.*, *Z mín.*, *Z atípicos*), el cambio brusco entre periodos (*Delta brusco*) y las secuencias de consumo cero (*Secuencias cero*).

En conjunto, las variables derivadas del SAX (*Cambios de símbolo* y la *Entropía SAX*), las métricas de complejidad y longitud temporal (*TSLF* y *LZC*) y las variables estadísticas calculadas a partir de las series de consumo conformaron un conjunto de características diseñado para representar de manera integral el comportamiento de los usuarios. Estas variables capturaron distintos niveles de información, desde la estructura simbólica y la variabilidad temporal hasta los patrones agregados de consumo. Su integración permitió disponer de una base sólida para el entrenamiento de los modelos de detección, facilitando la identificación de anomalías y posibles fraudes a partir de las particularidades de cada perfil de consumo.

5.5. TRANSFORMACIÓN DE VARIABLES

La transformación de variables se llevó a cabo con el propósito de optimizar la estructura de la base de datos y garantizar la validez estadística de las relaciones entre las variables categóricas y numéricas. Este proceso permitió reducir la dimensionalidad, eliminar redundancias y asegurar que las variables conservaran su significado analítico sin comprometer la representatividad de la información original.

5.5.1. Colapso de variables

El colapso de variables se implementó como una estrategia metodológica orientada a reducir la dispersión de categorías, mejorar la consistencia estadística y garantizar la aplicabilidad de las pruebas de independencia en las variables categóricas del conjunto de datos. Este proceso permitió reagrupar valores con baja frecuencia y unificar categorías con comportamientos equivalentes o cercanos, manteniendo la coherencia semántica y técnica de la información original. Las transformaciones se aplicaron de la siguiente manera:

- **Año a Año simplificado (colapso):**

La variable *Año* se agrupó en intervalos de décadas, dando origen a la variable *Año simplificado*, con las categorías: Antes 1990, Años 90 (intervalo entre 1990 y 1999), Años 2000 (intervalo entre 2000 y 2009) y Años 2010 (invervalo entre 2010 y 2013).

- **Clase metrológica a Clase metrológica simplificada (colapso):**

A partir de la variable original *Clase metrológica*, se construyó la nueva variable *Clase metrológica simplificada*, agrupando los valores conforme a su equivalencia técnica y frecuencia de aparición. La clase 100 se mantuvo como una categoría independiente bajo la misma denominación (*Clase 100*), mientras que las clases 102 y 103 se integraron en una categoría conjunta denominada *Otras clases*.

- **Diámetro a Diámetro simplificado (colapso):**

La variable *Diámetro* se reestructuró mediante la creación de la variable *Diámetro simplificado*, agrupando los valores numéricos en dos rangos técnicos: *Diámetro bajo* (13 y 19 mm) y *Diámetro medio alto* (25, 38, 50, 75 y 100 mm).

- **Localidad a Localidad simplificada (colapso):**

A partir de la variable original *Localidad*, se construyó la variable *Localidad simplificada*, con el propósito de reducir la dispersión de categorías y facilitar la interpretación geográfica de los resultados. Para ello, las comunas fueron agrupadas según su proximidad territorial.

En este proceso, las comunas de Los Ángeles, Nacimiento y Santa Bárbara se agruparon en la *Zona Bio Bío interior*; San Fernando y Chimbarongo en la Zona Colchagua; las localidades del Gran Concepción y la Provincia de Arauco en la *Zona Concepción Arauco*; Cauquenes, Pelluhue y Constitución en la *Zona Litoral Maule*; Curicó, Molina, Linares, Talca y San Clemente en la *Zona Maule*; Chillán, Quillón, San Carlos y Yungay en la *Zona Ñuble*; y Rancagua, Requínoa, San Vicente y Santa Cruz en la *Zona O'Higgins*. Finalmente, aquellas localidades no incluidas en las categorías anteriores fueron agrupadas en la *Zona Otras*.

- **Marca a Marca simplificada (colapso):**

La variable *Marca* se unificó bajo la nueva estructura *Marca simplificada*. Las categorías principales se reorganizaron de la siguiente forma:

En primer lugar, se agrupó en la categoría *CCM-Maipo-Actaris* con sus combinaciones asociadas: CCM-Maipo-Actaris y CCM-Maipo-Actaris (TMI-TMII). De forma análoga, la denominación Lautaro-Invensys-Sensus fue renombrada como *Lautaro-Sensus*, mientras que la marca Elster (Ex Tavira) se ajustó a la forma *Elster Ex Tavira*. Asimismo, Tavira-Iberconta-abb se renombró como *Tavira-Iberconta*, con el fin de estandarizar la nomenclatura de los registros.

Las marcas que incluían la expresión NO VIGENTE, tales como AA-NO VIGENTE, BM-NO VIGENTE, HM-NO VIGENTE, IB-NO VIGENTE, LM-NO VIGENTE, MM-NO VIGENTE, MN-NO VIGENTE, MONTROUGE-NO VIGENTE, MR-NO VIGENTE, OTROS-NO VIGENTE, PL-NO VIGENTE, SC-NO VIGENTE, STTUGAR-NO VIGENTE, TA-NO VIGENTE, WA-NO VIGENTE y WO-NO VIGENTE, fueron consolidadas en una única categoría denominada *Marca no vigente*, con el propósito de homogenizar la representación de medidores fuera de uso o retirados del mercado.

Finalmente, las marcas restantes, como CCM-Actaris (Flostar M), CCM-Schlumberger-Woltman, CCM-Itron-Flodis TU1M25, CCM-Itron-Flodis Cyble, MD-Siemens, Medidor de Descarga-Aquamaster, MV-U-CCM-Itron-Flodis TU1M2 y Sappel, fueron clasificadas en la categoría *Otras marcas*, la cual agrupa los casos residuales que no pertenecen a los grupos principales y presentan una baja frecuencia de aparición en la base de datos.

- **Ruedas a Ruedas simplificada (colapso):**

La variable *Ruedas* se simplificó en dos categorías principales, dando origen a la variable *Ruedas simplificada*. Los valores 4 y 5 se agruparon como 4-5 ruedas, mientras que 6 y 7 conformaron la categoría 6-7 ruedas.

- **Transmisión a Transmisión simplificada (colapso):**

Finalmente, la variable *Transmisión* fue reestructurada como *Transmisión simplificada*, unificando los valores originales en cuatro categorías comprensibles: T1 (Transmisión 1), T2 (Transmisión 2), T511 (Transmisión 511) y Otro tipo, que integra los valores de Transmisión 507 y 512, considerados residuales por su baja frecuencia de aparición.

5.5.2. Validación de independencia y asociación

Con el objetivo de analizar la relación estadística entre las variables categóricas simplificadas y la variable objetivo, se aplicó la prueba de independencia chi-cuadrado (χ^2). Este procedimiento permitió evaluar si existía dependencia significativa entre ambas, contrastando la hipótesis nula de independencia. Se estableció un nivel de significancia de $\alpha = 0{,}02$ para el caso de fraude, debido al tamaño grande de la muestra, y de $\alpha = 0{,}05$ para anomalías, dada la menor cantidad de observaciones. Los valores p inferiores al umbral correspondiente se interpretaron como evidencia suficiente para rechazar la independencia y, por tanto, confirmar la existencia de asociación entre las variables analizadas y la presencia de fraude o de comportamientos anómalos.

Dado que la prueba chi-cuadrado (χ^2) únicamente permite establecer la existencia de asociación pero no su intensidad, se calculó el coeficiente V de Cramer como medida complementaria de la fuerza de la relación entre las variables. Este coeficiente, cuyo valor oscila entre 0 y 1, permitió clasificar la magnitud de la asociación en tres niveles: baja ($V < 0{,}3$), moderada ($0{,}3 \leq V < 0{,}5$) y alta ($V \geq 0{,}5$).

La combinación de ambas medidas (significancia estadística mediante chi-cuadrado y fuerza de asociación mediante V de Cramer) proporcionó un marco integral para identificar las variables categóricas con mayor relación con el fraude, garantizando la validez estadística y la relevancia analítica de los resultados incluidos en las etapas multivariadas posteriores.

5.5.3. Análisis de correlación

Para identificar relaciones entre las variables numéricas del conjunto de datos, se aplicó el coeficiente de correlación de Spearman (ρ), una medida no paramétrica que permite evaluar la fuerza y dirección de las asociaciones monótonas entre variables sin requerir supuestos de normalidad. Este método se basó en los rangos de los valores, resultando apropiado ante la presencia de distribuciones asimétricas, valores atípicos o relaciones no lineales. Se calculó la matriz de correlaciones para todas las variables numéricas relacionadas con el consumo y el comportamiento de los medidores, considerando como de alta correlación aquellas con un valor absoluto de $|\rho| > 0{,}7$.

5.6. VARIABLES FINALES PARA LOS MODELOS

La preparación y selección de variables se realizó considerando las características y requerimientos de cada tipo de modelo, aplicándose de manera consistente tanto en los análisis de detección de fraudes como en los de identificación de anomalías.

En los modelos supervisados, se utilizaron los algoritmos *Random Forest* y *XGBoost*, los cuales presentan alta robustez frente a la escala y a la presencia de valores extremos. Por esta razón, las variables se mantuvieron en su versión original y se conservaron los valores atípicos, dado que estos algoritmos son capaces de manejar su efecto sin comprometer la estabilidad del modelo. En contraste, para las regresiones logísticas se aplicó un proceso de transformación y estandarización de las variables numéricas, combinando las técnicas de *winsor*, *shift* y *log1p* con el fin de reducir la asimetría de las distribuciones y mitigar la influencia de valores extremos. Posteriormente, las variables fueron estandarizadas para homogeneizar la escala y permitir una comparación adecuada entre coeficientes, garantizando la estabilidad y la interpretación coherente de los parámetros.

En el caso de la modelación no supervisada, se implementó el algoritmo *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*), adoptando el conjunto de variables propuesto por Ghamkhar et al. [47] y siguiendo las recomendaciones metodológicas de su estudio para el agrupamiento y la detección de patrones de consumo atípico en redes de distribución de agua. Las variables del estudio original incluyeron los indicadores *LZC 2 bins*, *LZC 9 bins*, *LZC 4 bins*, *LZC 99 bins* y *TSLF*, asociados a la complejidad y estabilidad temporal de las series de consumo. Además de dichas variables, se incorporaron atributos contextuales y técnicos relevantes para el dominio de estudio, específicamente: *Marca simplificada*, *Localidad simplificada*, *Año simplificado*, *Transmisión simplificada*, *Ruedas simplificada*, *Diámetro simplificado*, *Clase metrológica simplificada*, *Latitud* y *Longitud*. Este conjunto ampliado permitió capturar tanto la variabilidad temporal del consumo como las características físicas y geográficas de los medidores, fortaleciendo la capacidad del modelo para identificar patrones anómalos sin requerir etiquetas previas, y priorizando la densidad y la distancia como criterios de similitud.

Las variables categóricas se conservaron sin modificaciones, manteniendo su codificación original en todos los enfoques. En conjunto, la estructura final de variables permitió disponer de un conjunto equilibrado y coherente para el modelado de fraude y anomalías, asegurando la consistencia entre los distintos métodos aplicados y la comparabilidad de resultados entre los enfoques supervisados y no supervisados.

5.7. ELECCIÓN DE LOS MODELOS SUPERVISADOS

La elección de los modelos supervisados se llevó a cabo a partir de criterios metodológicos relacionados con la estructura de los datos y la naturaleza del problema. Dado que el conjunto de observaciones presentaba un marcado desbalance y alta complejidad, se optó por emplear algoritmos de ensamble basados en árboles de decisión. En particular, se seleccionaron *Random Forest* y *XGBoost* por su capacidad para manejar datos estructurados, capturar interacciones no lineales y permitir el ajuste de la importancia relativa de las clases minoritarias mediante parámetros internos. La evaluación de los modelos se planificó con base en métricas adecuadas para contextos de clases desbalanceadas, priorizando el área bajo la curva de precisión vs. exhaustividad (AUC-PR) como criterio principal de comparación. Adicionalmente, en el caso del modelo con enfoque de negocio, se consideraron los errores tipo I y tipo II con el fin de equilibrar el costo asociado a las falsas alarmas y a la omisión de casos de fraude, garantizando una evaluación más alineada con el impacto operativo real.

METODOLOGÍA: MODELOS FRAUDE

5.8. MODELACIÓN SUPERVISADA FRAUDE

En esta sección se presentan dos enfoques complementarios de modelación supervisada orientados a la detección de fraude en el consumo de agua. El primero corresponde a la modelación sin enfoque de negocio, cuyo objetivo principal es comparar el desempeño técnico de los algoritmos *Random Forest (RF)* y *XGBoost (XGB)* en términos de precisión, estabilidad y capacidad de generalización. El segundo enfoque incorpora una perspectiva operativa y económica, al integrar los costos asociados a los errores de clasificación y la priorización de recursos de inspección.

5.8.1. Modelación Random Forest vs. XGBoost sin enfoque de negocio (fraude)

La modelación supervisada sin enfoque de negocio se desarrolló con el propósito de analizar comparativamente el desempeño predictivo de los algoritmos *Random Forest* y *XGBoost* en la detección de patrones de fraude, sin considerar restricciones operativas ni criterios económicos. Esta etapa se centró en la evaluación técnica de ambos modelos, atendiendo a su capacidad de generalización, estabilidad y precisión en la clasificación de observaciones pertenecientes a una clase minoritaria.

Los algoritmos seleccionados fueron *Random Forest* y *XGBoost*, elegidos por su robustez frente al ruido y su capacidad para capturar relaciones no lineales. Adicionalmente, se incorporaron dos métodos de calibración probabilística: la *Regresión Isotónica*, de carácter no paramétrico, flexible y monotónico, y el *Platt Scaling* o regresión sigmoidal, de naturaleza paramétrica, que ajusta una regresión logística sobre las probabilidades estimadas. La inclusión de estos calibradores permitió evaluar la capacidad de cada modelo para producir estimaciones probabilísticas coherentes con la frecuencia real del fenómeno.

El conjunto de datos se dividió en dos subconjuntos: un 80 % destinado al entrenamiento y validación cruzada, y un 20 % reservado para la prueba final. Esta partición se realizó de forma aleatoria y estratificada, respetando la prevalencia de la clase fraude ($\approx$ 1,65 %), con el fin de conservar la representatividad estadística del fenómeno bajo estudio. La optimización de hiperparámetros se llevó a cabo mediante una búsqueda exhaustiva con `GridSearchCV`, utilizando validación cruzada estratificada de cinco *folds*. El espacio de búsqueda incluyó los parámetros característicos de cada algoritmo, los métodos de calibración probabilística y un rango amplio de umbrales de clasificación comprendido entre 0,1 y 0,9. La métrica de optimización en este escenario fue el *F1-Score* y los niveles evaluados fueron:

Tabla 5.1: Hiperparámetros evaluados en el *Grid Search* de los modelos *Random Forest* y *XGBoost*. Fuente: Elaboración propia.

Hiperparámetro	Random Forest	XGBoost
`n estimators`	150 y 250	150 y 250
`Max depth`	Ninguno, 10 y 20	3, 5 y 7
`Min samples split`	5 y 10	–
`Min samples leaf`	2 y 4	–
`Max features`	*sqrt*	–
`Learning rate`	–	0,05 y 0,10
`Subsample`	–	0,8
`Colsample bytree`	–	0,8

Para el modelado, como métrica principal de selección se empleó el *F1-Score*, por su capacidad para equilibrar la precisión y la exhaustividad en contextos de clases desbalanceadas. De forma complementaria, se consideró el *Brier Score* como criterio secundario, dado que mide la calidad de las probabilidades predichas y permite evaluar la coherencia probabilística de las salidas del modelo. Durante la validación cruzada se determinó también el umbral de clasificación óptimo, definido como aquel que maximizó el *F1-Score* promedio en los pliegues.

Una vez identificadas las configuraciones óptimas de hiperparámetros y calibración, el modelo final (correspondiente a la combinación que mejor equilibró desempeño y consistencia probabilística) fue

reentrenado sobre la totalidad del conjunto de entrenamiento y posteriormente evaluado en el subconjunto de prueba (*hold-out*), sin aplicar ajustes adicionales de balance de clases. Esta etapa permitió verificar la estabilidad de los resultados y la capacidad de generalización del modelo en datos no observados. En la Figura 5.1 se presenta el flujograma de dicha metodología.

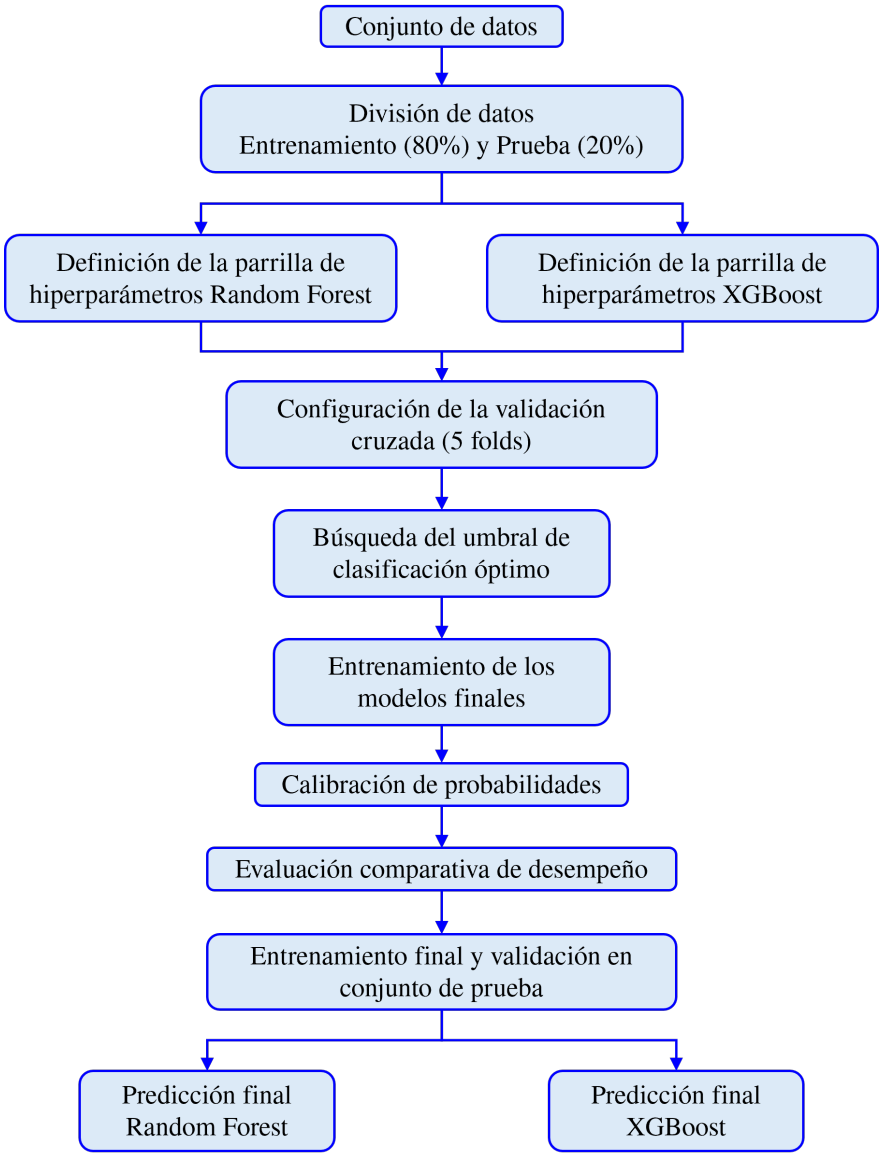

Figura 5.1: Flujograma de la modelación supervisada sin enfoque de negocio.
Fuente: Elaboración propia.

5.8.2. Modelación Random Forest vs. XGBoost con enfoque de negocio (fraude)

La modelación supervisada con enfoque de negocio se desarrolló utilizando los algoritmos *Random Forest* y *XGBoost*, seleccionados por su capacidad para manejar grandes volúmenes de datos estructurados, capturar relaciones no lineales y mantener un equilibrio adecuado entre precisión e interpretabilidad. Este proceso integró consideraciones técnicas y operativas con el propósito de construir un modelo predictivo sólido, capaz de apoyar la detección de fraude en el consumo de agua desde una perspectiva orientada a la toma de decisiones empresariales y a la optimización de recursos de inspección.

El conjunto de datos se dividió en dos subconjuntos: un 80 % destinado al entrenamiento y validación cruzada, y un 20 % reservado para la prueba final. Esta división se realizó de manera aleatoria y estratificada, conservando la proporción original de casos de fraude y no fraude para mantener la representatividad de la variable objetivo en ambas particiones (1,65 % no fraude vs. 98,35 % fraude).

Dado el fuerte desbalance entre clases, se implementó un manejo diferenciado de los costos de clasificación errónea a través de ponderaciones proporcionales al costo relativo del escenario típico de falsos negativos y falsos positivos (CFN/CFP $= \frac{200.000}{35.000} \approx 5{,}7$, redondeado a 6). En el caso de *Random Forest*, se utilizó el parámetro `class weight = {0:1, 1:6}`, mientras que para *XGBoost* se aplicó `scale pos weight = 6`. Estas ponderaciones permitieron aumentar la sensibilidad del modelo hacia la clase minoritaria sin sacrificar de manera significativa la estabilidad del desempeño general.

La optimización de hiperparámetros se realizó mediante una búsqueda exhaustiva con `GridSearchCV`, bajo un esquema de validación cruzada estratificada de cinco pliegues. Este proceso permitió explorar distintas combinaciones de parámetros estructurales, tales como el número de árboles, la profundidad máxima, la tasa de aprendizaje y los términos de regularización, buscando un equilibrio entre complejidad y capacidad predictiva. La métrica de optimización en este escenario fue el *AUC-PR* y los niveles evaluados fueron los mismos de la Tabla 5.1.

Una vez identificadas las configuraciones de mejor rendimiento promedio, se procedió a la calibración de las probabilidades estimadas con el objetivo de obtener valores interpretables y coherentes con la frecuencia real del fenómeno. Se evaluaron dos métodos de ajuste: el *Platt Scaling*, basado en regresión logística paramétrica, y la *Regresión Isotónica*, de naturaleza no paramétrica y monotónica. La selección final del calibrador se basó en el *Brier Score*, métrica que mide la calidad de las predicciones probabilísticas al evaluar la correspondencia entre las probabilidades estimadas y los valores observados.

Los modelos calibrados fueron posteriormente comparados de acuerdo con tres criterios teóricos: desempeño estadístico (medido por el *AUC-PR*), adecuación probabilística (evaluada mediante el *Brier Score*) y utilidad operativa (reflejada en métricas económicas que vinculan la predicción del modelo con el beneficio potencial de las inspecciones). Este enfoque permitió integrar la evaluación técnica con la perspectiva empresarial, garantizando que el modelo seleccionado optimizara tanto la precisión estadística como el valor operativo de las decisiones derivadas.

El modelo final fue reentrenado sobre la totalidad del conjunto de entrenamiento, incorporando las ponderaciones de clase y los hiperparámetros óptimos determinados en la búsqueda exhaustiva. Finalmente, fue evaluado sobre el conjunto de prueba (*hold-out*) para verificar su capacidad de generalización y estabilidad sobre datos no observados. En la Figura 5.2 se presenta el flujograma de dicha metodología.

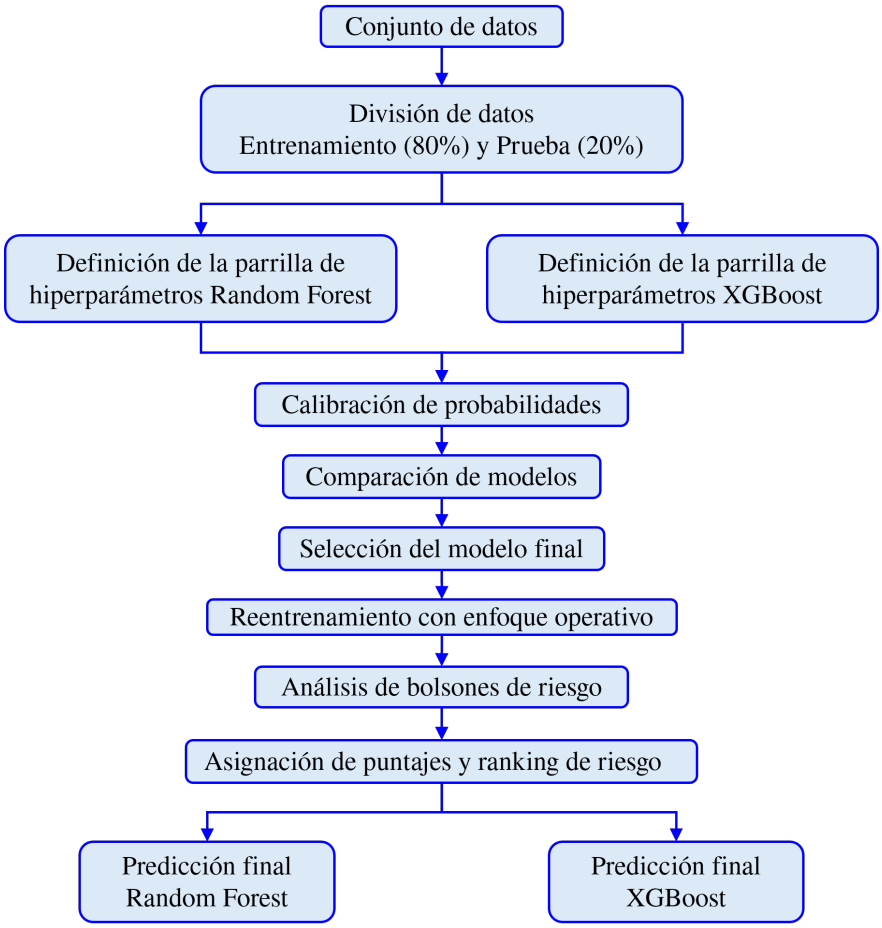

Figura 5.2: Flujograma de la modelación supervisada con enfoque de negocio.
Fuente: Elaboración propia.

5.8.3. Marco de costos, ganancia y umbral de rentabilidad

El modelo supervisado con enfoque de negocio parte de una concepción en la que la predicción se entiende no solo como un ejercicio estadístico, sino como un proceso orientado a la toma de decisiones con implicaciones económicas. Desde esta perspectiva, la literatura ha destacado la importancia de vincular el desempeño de los modelos con el valor económico esperado y la rentabilidad operativa, de modo que la precisión y el error trasciendan su interpretación técnica y adquieran un sentido financiero [86], [87].

Definición de costos y variables

- C_{FP} (costo de falso positivo): Gasto de realizar una inspección sin encontrar fraude.
- C_{FN} (costo de falso negativo): Pérdida asociada a un fraude no detectado.
- TP (verdaderos positivos): Fraudes correctamente detectados.
- FP (falsos positivos): inspecciones realizadas a clientes sin fraude.

Con estas definiciones, la *ganancia total* para un lote de inspecciones de tamaño k (donde $k = TP + FP$) se expresa como:

$$G = TP\,(C_{FN} - C_{FP}) - FP\,C_{FP}$$

Umbral de rentabilidad: precisión mínima

Sea la precisión $Prec = \frac{TP}{k}$. Entonces:

$$TP = k\,\text{Prec}, \qquad FP = k\,(1 - \text{Prec})$$

Sustituyendo en G y resolviendo para el punto de equilibrio ($G = 0$), el *umbral mínimo de precisión* requerido para que el lote sea rentable es:

$$\textit{Precisión mínima} = \frac{C_{FP}}{C_{FN}}$$

Cualquier precisión por debajo de este umbral implica pérdidas netas; valores por encima lo superan y generan beneficios.

Escenarios de costos y precisión requerida

Tabla 5.2: Escenarios de costos y precisión mínima requerida.
Fuente: [88], [89], [90], [91], [92]

Escenario	Costo Inspección (C_{FP})	Pérdida Evitada (C_{FN})	Precisión Mínima
Típico	35.000 CLP	200.000 CLP	18 %
Optimista	20.000 CLP	200.000 CLP	10 %
Adverso	50.000 CLP	150.000 CLP	33 %

5.9. MODELACIÓN NO SUPERVISADA FRAUDE

En esta sección se presenta la modelación no supervisada aplicada al conjunto de datos de consumo de agua, utilizada como complemento a la modelación supervisada. El propósito fue identificar patrones de comportamientos irregulares sin depender de etiquetas previas, con el fin de contrastar y validar los resultados obtenidos en los modelos de clasificación.

5.9.1. Modelación DBSCAN sin enfoque de negocio con optimización del F1-Score (fraude)

Esta metodología se inspiró directamente en la propuesta de Ghamkhar et al. [47], quienes plantearon un enfoque no supervisado para la detección de consumos anómalos en medidores de agua de baja resolución. En su trabajo, los autores combinaron el algoritmo *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) con métricas de complejidad de series temporales (*Lempel-Ziv Complexity*) y un proceso de calibración semi-supervisado. Aunque el estudio original se centró en la detección de anomalías, en el presente trabajo dicho enfoque se extrapoló al contexto de fraude.

El carácter innovador radicó en la implementación de un proceso híbrido: aunque DBSCAN se ejecuta de manera no supervisada sobre el conjunto de prueba, sus hiperparámetros se calibraron utilizando un subconjunto etiquetado del entrenamiento, optimizando el *F1-Score* como métrica guía. De esta manera, se replicó el enfoque semi-supervisado de Ghamkhar et al. [47], adaptándolo a la naturaleza y características del conjunto de datos empleado en este estudio.

Para la ejecución del modelo se utilizaron cinco variables derivadas de las series temporales mediante la técnica *Symbolic Aggregate Approximation* (SAX): *LZC norm 2 bins*, *LZC norm 9 bins*, *LZC norm 4 bins*, *LZC norm 99 bins* y *TSLF*. Adicionalmente, se incorporaron características de los medidores (*Marca simplificada*, *Clase metrológica simplificada*, *Transmisión simplificada*, *Ruedas simplificada*, *Diámetro simplificado* y *Año simplificado*) así como características geográficas y espaciales (*Localidad simplificada*, *Latitud* y *Longitud*). El uso conjunto de variables temporales, técnicas y espaciales permitió representar de manera integral tanto la dinámica de consumo como las propiedades físicas y de localización de los medidores, facilitando la detección de patrones atípicos en distintas dimensiones de análisis.

Con el fin de optimizar los hiperparámetros del algoritmo de manera eficiente, se construyó un subconjunto de calibración a partir del conjunto de entrenamiento (80 %), incluyendo la totalidad de los casos de fraude disponibles y una muestra aleatoria de no fraude equivalente a seis veces la cantidad de fraudes. Esta proporción se fundamenta en la recomendación de Ghamkhar et al. [47], quienes propusieron la construcción de subconjuntos calibrados que mantengan una relación acotada entre clases, y en la coherencia con la metodología supervisada orientada al retorno de inversión (ROI), en la que la relación CFN/CFP fue aproximadamente 6. Este ajuste aseguró consistencia metodológica entre los enfoques supervisados y no supervisados.

El proceso de optimización de hiperparámetros se desarrolló en tres etapas. En primer lugar, se estimó un valor inicial del parámetro ε mediante el método del *codo* aplicado a la gráfica de distancias al vecino más cercano (*KneeLocator*). Posteriormente, se realizó una búsqueda exhaustiva en rejilla (*Grid Search*) evaluando distintas combinaciones de ε y `min samples` en torno a los valores iniciales. Finalmente, se seleccionó la combinación que maximizó el *F1-Score* promedio en el subconjunto de calibración, asegurando un balance adecuado entre sensibilidad y precisión en la identificación de fraudes.

El cálculo de la distancia al k-ésimo vecino se empleó únicamente como criterio auxiliar de optimización, manteniendo el mecanismo clásico de DBSCAN para la detección de fraudes. En este esquema, los puntos que tuvieron al menos el número de vecinos definido por el parámetro `min samples` dentro del radio ε se consideraron puntos núcleo, alrededor de los cuales se expandieron los *clusters*. Los puntos que no pertenecen a ningún *cluster* se etiquetaron como ruido y, en el contexto del presente estudio, se interpretaron como potenciales fraudes.

Una vez calibrados los hiperparámetros, el modelo *DBSCAN* se aplicó sobre el conjunto de prueba (20 %) para la detección final de fraudes. Las observaciones clasificadas como ruido (*cluster* = -1) fueron interpretadas como fraudes, en coherencia con la hipótesis de que los comportamientos fraudulentos representan desviaciones significativas respecto a los patrones regulares de consumo. En la Figura 5.3 se presenta el flujograma general de esta metodología.

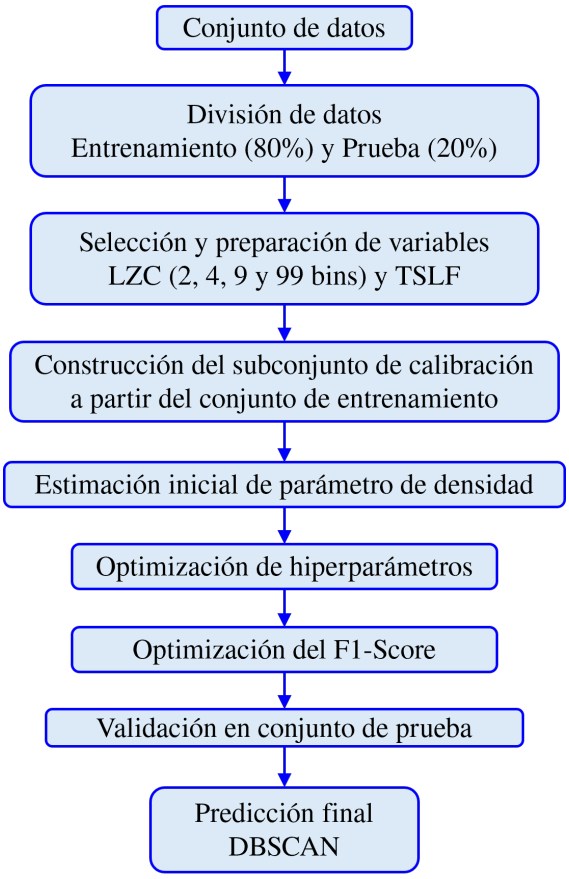

Figura 5.3: Flujograma de la modelación no supervisada sin enfoque de negocio (optimización de F1-Score). Fuente: Elaboración propia.

5.9.2. Modelación DBSCAN sin enfoque de negocio con optimización del AUC-PR (fraude)

Esta versión de la modelación siguió la misma estructura metodológica descrita en la sección anterior, manteniendo el proceso de calibración semi-supervisado y el procedimiento de búsqueda de hiperparámetros. La única diferencia radica en la métrica de optimización empleada: en lugar del *F1-Score*, se utilizó el *Area Under the Precision-Recall Curve* (*AUC-PR*) como criterio principal de selección. En la Figura 5.4 se presenta el flujograma general de esta metodología.

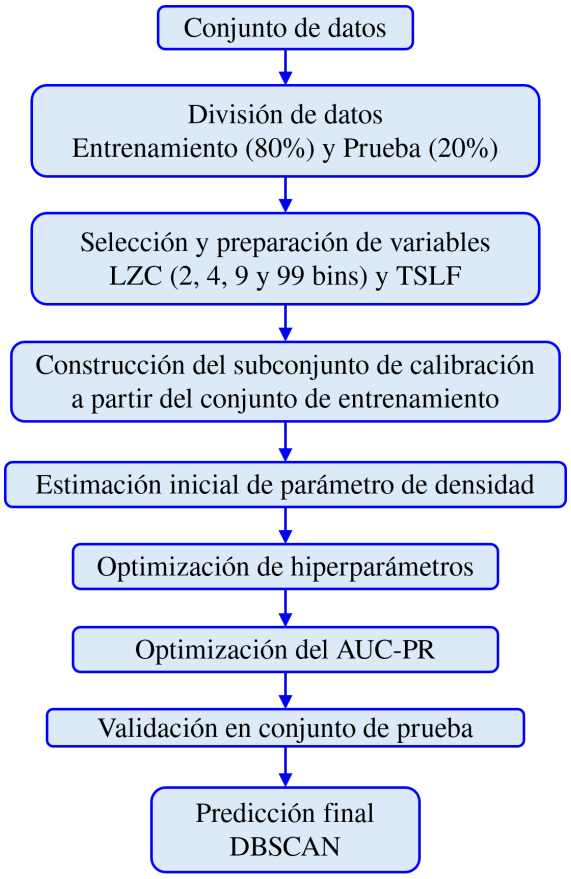

Figura 5.4: Flujograma de la modelación no supervisada sin enfoque de negocio (optimización del AUC-PR). Fuente: Elaboración propia.

5.10. REGRESIÓN LOGÍSTICA FRAUDE

El proceso de modelación logística se estructuró en tres fases principales: la división de los datos, el ajuste y diagnóstico del modelo logístico, y la extensión del análisis mediante técnicas aditivas y espaciales. Cada una de estas etapas se diseñó para garantizar la validez estadística del modelo y su capacidad para capturar los patrones asociados con el comportamiento fraudulento en el consumo de agua.

En primer lugar, los datos se dividieron en conjuntos de entrenamiento y prueba con el fin de evaluar la capacidad de generalización del modelo. La muestra de entrenamiento se empleó para el ajuste, mientras que la de prueba se destinó a la validación final del desempeño. Dado el marcado desbalance entre las clases (fraude y no fraude), se aplicaron estrategias de balanceo para mejorar la representatividad de la clase minoritaria y reducir el sesgo. Asimismo, se prepararon los factores categóricos, garantizando una codificación adecuada de las variables simbólicas para su correcta interpretación en el modelo.

Posteriormente, se ajustaron diversos Modelos Lineales Generalizados (GLM) con distribución binomial y enlace logit, denominados Modelos 1 a 6 (M1 – M6), con el propósito de analizar la contribución de los factores técnicos, geográficos y de consumo al fenómeno de fraude. Es importante destacar que la evaluación de un modelo logit involucra dos dimensiones distintas: primero, la validez del modelo,

la cual depende del cumplimiento de supuestos como la linealidad en el logit, la independencia de los errores y la ausencia de influencia espacial en los residuos, y segundo, el desempeño predictivo del modelo, que describe qué tan bien clasifica los casos de fraude.

En este trabajo de grado, la validez fue evaluada mediante pruebas sobre los supuestos estructurales que incluyen linealidad, diagnóstico de residuos y autocorrelación espacial, mientras que el desempeño del modelo se analizó mediante métricas como el área bajo la curva ROC, la prueba de Hosmer–Lemeshow (HL) y el pseudo-R^2. Esto permitió identificar qué modelos presentaban simultáneamente una buena capacidad de discriminación y una calibración adecuada, sin perder de vista la validez estadística necesaria para una interpretación confiable de los coeficientes e inferencias. Además, se verificó la suposición de linealidad en el logit mediante un Modelo Aditivo Generalizado (GAM), utilizado exclusivamente como prueba diagnóstica para evaluar la validez de los modelos de regresión logística estimados (M1–M6). Esta verificación es importante, ya que el modelo logit exige que la relación entre cada predictor y el logit de la probabilidad sea lineal. Los resultados del GAM permitieron identificar que varios predictores presentan relaciones no lineales con la probabilidad de fraude, lo que indica un incumplimiento parcial de este supuesto. En la Figura 5.5 se presenta el flujograma general de esta metodología.

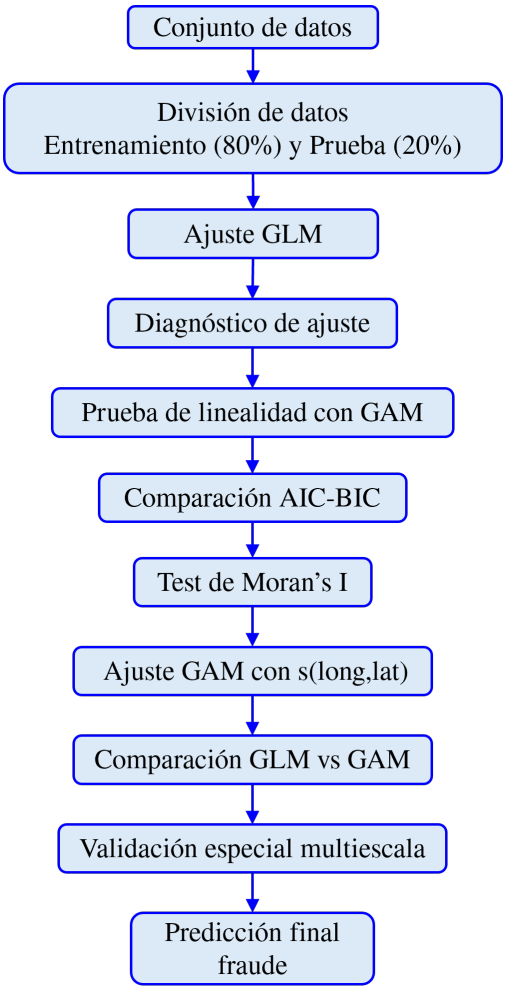

Figura 5.5: Flujograma de la regresión logística para la detección de fraude.
Fuente: Elaboración propia.

METODOLOGÍA: MODELOS ANOMALÍAS

5.11. MODELACIÓN SUPERVISADA ANOMALÍAS

La metodología se estructuró en dos fases principales: la *selección competitiva de modelos* y la *evaluación robusta del modelo ganador*. En la primera fase, se compararon los algoritmos *Random Forest (RF)* y *XGBoost (XGB)*, considerados por su capacidad para capturar relaciones no lineales y manejar interacciones complejas entre variables. Ambos modelos fueron sometidos a un proceso de *calibración de probabilidades* en cada *fold*, aplicando dos métodos complementarios: la *regresión isotónica*, de naturaleza no paramétrica y monotónica, y el *Platt scaling* o regresión sigmoidal, de carácter paramétrico. El calibrador seleccionado correspondió a aquel que obtuvo el menor *Brier Score*, garantizando una mejor fiabilidad en las probabilidades estimadas.

Los datos se dividieron en un 80 % para entrenamiento y validación y un 20 % para prueba (*hold-out*), asegurando la estratificación de las clases para preservar la proporción de observaciones minoritarias. La optimización de hiperparámetros se realizó mediante *búsqueda exhaustiva (GridSearchCV)* con cinco *folds* estratificados, empleando como métrica principal el *F1-Score*. Además, se exploró un rango de umbrales de decisión entre 0,1 y 0,9, seleccionando aquel que, junto con los hiperparámetros óptimos, maximizó el desempeño promedio. La selección final del modelo ganador se basó en dos criterios jerárquicos: el *F1-Score* promedio como métrica principal y el *Brier Score* promedio como medida de desempate y fiabilidad.

En la segunda fase, denominada *evaluación robusta*, se empleó el modelo ganador (*Random Forest*) dentro de un protocolo de *Validación Cruzada Estratificada y Repetida (Repeated Stratified K-Fold)* configurado con 10 repeticiones y 5 *folds*, totalizando 50 evaluaciones independientes. En cada iteración se entrenó el modelo *Random Forest* con sus mejores hiperparámetros, se aplicó la calibración isotónica (seleccionada en la fase previa por su menor *Brier Score*) y se utilizó el umbral óptimo definido en la fase de selección. En cada ejecución se registraron las métricas *F1-Score*, *Precision* y *Recall*, junto con las matrices de confusión correspondientes, con el fin de obtener estimaciones estables del desempeño promedio y de su variabilidad.

Finalmente, se realizó una *evaluación base* sobre el conjunto de entrenamiento (80 %) y una *evaluación final* sobre el conjunto de prueba (20 %), reportando la matriz de confusión y las métricas *F1-Score*, *Precision* y *Recall* con el umbral óptimo previamente determinado. Este procedimiento garantizó la consistencia entre el proceso de optimización, calibración y evaluación, maximizando la fiabilidad estadística de las predicciones del modelo. En la Figura 5.6 se presenta el flujograma general de esta metodología.

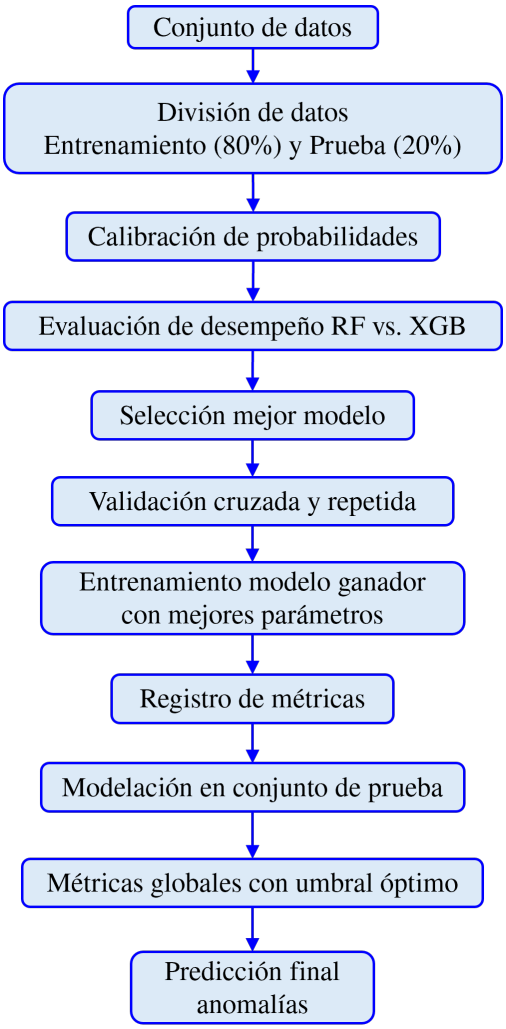

Figura 5.6: Flujograma de la modelación supervisada para la detección de anomalías.
Fuente: Elaboración propia.

5.12. REGRESIÓN LOGÍSTICA ANOMALÍAS

Esta sección sigue la misma metodología descrita previamente en el apartado de regresión logística para fraude (Sección 5.10), manteniendo las mismas etapas de preparación, modelado y validación. La única diferencia corresponde a la variable de respuesta, que en este caso se asocia a la detección de anomalías en el consumo, en lugar de la identificación de casos de fraude. En la Figura 5.7 se presenta el flujograma general de esta metodología.

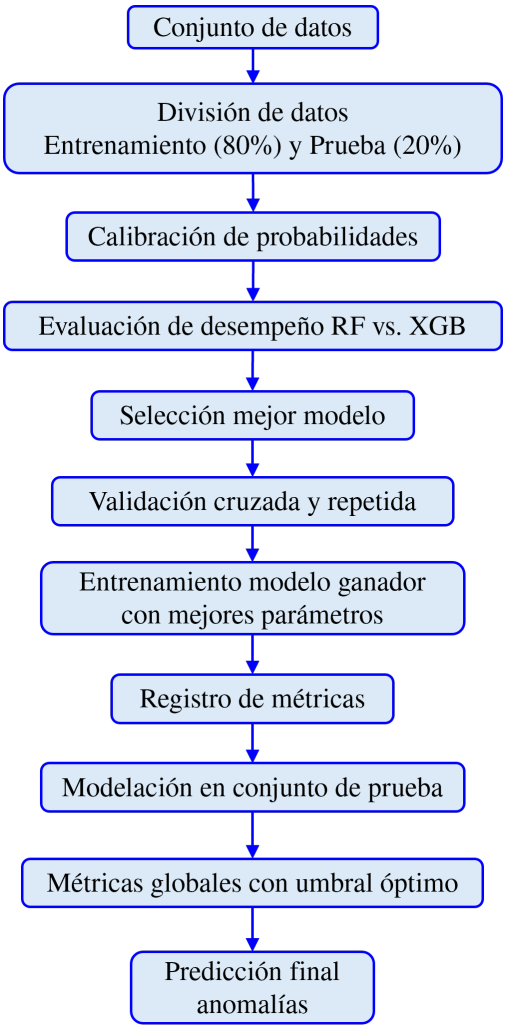

Figura 5.7: Flujograma de la modelación supervisada para la detección de anomalías.
Fuente: Elaboración propia.

6 DATOS: PREPARACIÓN Y ANÁLISIS

6.1. BASES SUMINISTRADAS

El conjunto de datos incluía registros mensuales de consumo de agua por cliente, correspondientes al periodo comprendido entre febrero de 2008 y septiembre de 2013, en diversas localidades de Chile. Adicionalmente, contenía información detallada sobre las características técnicas de los medidores mecánicos de agua, tales como la marca, diámetro, ruedas, transmisión y año de instalación, así como variables administrativas u operativas relacionadas con el servicio prestado.

Se emplearon como insumo dos bases de datos. La primera, denominada *Base A*, sirvió como fuente principal para el desarrollo de los modelos relacionados con la detección de fraude, cuyas características se describen a continuación:

Este conjunto incluía 32.114 registros y un total de 93 variables, de las cuales 68 correspondían a registros mensuales de consumo de agua por parte de los clientes. A partir del análisis de estas variables, se identificaron las siguientes categorías según su naturaleza y propósito.

- **Identificadores y ubicación:** ID, código de región, código de ciudad, dirección, localidad, nombre de localidad, ubicación, latitud, longitud y sociedad.
- **Consumos:** Desde febrero de 2008 hasta septiembre de 2013 (68 meses).
- **Características del medidor:** Marca, diámetro, ruedas, transmisión y año de construcción.
- **Datos administrativos u operativos:** Unidad de lectura, ruta, tarifa, contrato, tipo de cliente, tipo de despacho, clase, punto de suministro, objeto de conexión, equipo y tipo de servicio.
- **Variable objetivo:** Número de fraudes e indicador de fraude.

Sin embargo, se identificó que no todas las variables eran necesarias para los objetivos del proyecto. Por esta razón, se realizó una depuración en la que se descartaron aquellas variables que no aportaban información relevante para la predicción de fraudes. La selección final priorizó las variables relacionadas con el consumo de agua, las características físicas y técnicas del medidor y los indicadores de fraude. De esta manera, las variables seleccionadas se muestran en la Tabla 6.1:

Tabla 6.1: Descripción de las variables seleccionadas de la base de datos
Fuente: Elaboración propia.

Variable	Tipo	Descripción
ID	Numérica	Identificador único del cliente.
Localidad	Categórica	Localidad del cliente.
Dirección	Cadena de texto	Dirección del cliente en el sistema, que corresponde al domicilio registrado oficialmente.
Consumos mensuales	Numérica	Consumo de agua en metros cúbicos (m^3) registrado mensualmente desde febrero de 2008 hasta septiembre de 2013.
Año	Categórica	Corresponde al año de instalación del medidor.
Marca	Categórica	Marca del medidor de agua.
Clase metrológica	Categórica	Categoría que indica el grado de exactitud y precisión del medidor, conforme a normativas técnicas y metrológicas.
Diámetro	Numérica	Diámetro del medidor, expresado en milímetros (mm).
Ruedas	Categórica	Cantidad de ruedas utilizadas por el medidor para registrar el volumen de agua consumido.
Transmisión	Categórica	Tipo de transmisión del medidor.
Fraude	Binaria	Indicador de fraude (0 = no fraude, 1 = fraude).

Por otra parte, el segundo conjunto de datos, denominado *Base B*, correspondía a un subconjunto de la *Base A* compuesto por clientes sin registro de fraude (fraude = 0). Este conjunto incluía la misma información que la *Base A*, pero únicamente para 116 clientes, y adicionalmente incorporaba la clasificación del tipo de medición: subconsumo, sobreconsumo o consumo normal. Para su utilización en los modelos de predicción de anomalías, las tres categorías originales se recodificaron en dos clases: 1 para anomalía, que agrupa subconsumo y sobreconsumo, y 0 para consumo normal. De esta manera, el resumen es el siguiente:

Base A (principal):

- **Clientes:** 32.114
- **Base a utilizar en:** Modelación supervisada, modelación no supervisada y regresión logística para la detección de fraude.

Base B (subconjunto de la Base A):

- **Clientes:** 116
- **Base a utilizar en:** Modelación supervisada y regresión logística para la detección de anomalías.

Dado que la *Base B* es un subconjunto de la *Base A*, el tratamiento de los datos, incluyendo la eliminación de duplicados, la imputación de valores faltantes y otros procesos, realizado sobre la base principal también aplica para la *Base B*. En consecuencia, no es necesario ejecutar dicho procedimiento de forma independiente en ambas bases.

6.2. TRATAMIENTO DE DATOS

6.2.1. Limpieza de datos

Los resultados de esta revisión se detallan a continuación:

- **Datos duplicados:** Se identificaron 8 registros duplicados, los cuales fueron eliminados para evitar redundancias. En cada caso se conservó únicamente un registro por ID, eliminando las repeticiones. Los IDs afectados fueron: 1883625, 2410419, 2459713, 2752593, 2933585, 4153643, 4462868 y 4719831.
- **Medidores ficticios:** Según las recomendaciones técnicas de Essbio (empresa generadora de los datos), el medidor ZM-MED FICTICIO no debe ser considerado en el análisis. Se identificaron 13 registros de este tipo, los cuales también fueron excluidos del conjunto de datos.

6.2.2. Datos faltantes

Los resultados de esta revisión se detallan a continuación:

- **Consumos:** Fueron detectados 353 registros de consumo con datos faltantes, equivalentes al 0,2 % del total de consumos. Estos registros corresponden a 41 clientes, cuyo detalle se muestra a continuación (Tabla 6.3):

Tabla 6.3: Clientes con valores faltantes de consumo.
Fuente: Elaboración propia.

Faltante	Registros totales	Registros faltantes	% faltantes
Faltante 1	68	36	52,94 %
Faltante 2	68	33	48,53 %
Faltante 3	68	28	41,18 %
Faltante 4	68	26	38,24 %
Faltante 5	68	25	36,76 %
Faltante 6	68	25	36,76 %
Faltante 7	68	24	35,29 %
Faltante 8	68	23	33,82 %
Faltante 9	68	21	30,88 %
Faltante 10	68	7	10,29 %
Faltante 11	68	7	10,29 %
Faltante 12	68	7	10,29 %
Faltante 13	68	7	10,29 %
Faltante 14	68	6	8,82 %
Faltante 15	68	6	8,82 %
Faltante 16	68	6	8,82 %
Faltante 17	68	5	7,35 %
Faltante 18	68	4	5,88 %
Faltante 19	68	4	5,88 %
Faltante 20	68	4	5,88 %
Faltante 21	68	4	5,88 %
Faltante 22	68	3	4,41 %
Faltante 23	68	3	4,41 %
Faltante 24	68	3	4,41 %
Faltante 25	68	3	4,41 %

Continuación de la Tabla 6.3

Faltante	Registros	Faltantes	% faltantes
Faltante 26	68	3	4,41 %
Faltante 27	68	3	4,41 %
Faltante 28	68	3	4,41 %
Faltante 29	68	3	4,41 %
Faltante 30	68	3	4,41 %
Faltante 31	68	2	2,94 %
Faltante 32	68	2	2,94 %
Faltante 33	68	2	2,94 %
Faltante 34	68	2	2,94 %
Faltante 35	68	2	2,94 %
Faltante 36	68	2	2,94 %
Faltante 37	68	2	2,94 %
Faltante 38	68	1	1,47 %
Faltante 39	68	1	1,47 %
Faltante 40	68	1	1,47 %
Faltante 41	68	1	1,47 %

Para reducir el ruido, se eliminaron los registros con un porcentaje superior al 30 % de valores faltantes (resaltados en rojo en la Tabla 6.3). Esto correspondió a 9 usuarios, cuya proporción dentro del conjunto de datos es baja (0,2 %), obteniéndose un total de 32.084 usuarios. Posteriormente, se identificaron 112 valores faltantes. Estos fueron imputados mediante *spline* siempre que existieron más de 3 valores presentes, y mediante *interpolación lineal* en caso contrario.

- **Marca:** Se evidenciaron 7 registros con datos faltantes en esta variable categórica, esto equivale al 0,02 % del total de registros. Dado que se trata de una variable categórica, su imputación se realizó utilizando la moda de la marca.

6.2.3. Datos erróneos

Los resultados de esta revisión se detallan a continuación:

- **Diámetro del medidor con mala clasificación según la Norma Chilena 3274/1:** Dado que los diámetros de 75 y 80 mm son considerados equivalentes según la normatividad, y que el valor estándar aceptado es 75 mm, se encontraron 12 medidores con un diámetro de 80 mm. Estos fueron reemplazados por 75 mm para garantizar la consistencia en los datos.
- **Consumos de agua negativos:** Una vez ejecutada la imputación de datos faltantes mediante *spline* e *interpolación lineal*, se encontraron 52 registros con consumos de agua negativos, lo cual no es físicamente posible y representa un error en la captura o procesamiento de los datos. Dichos registros correspondieron a 20 clientes, cuyo detalle se muestra a continuación (Tabla 6.4):

Tabla 6.4: Clientes con registros negativos de consumo.
Fuente: Elaboración propia.

Negativo	Registros negativos
Negativo 1	3
Negativo 2	1
Negativo 3	2

Continuación de la Tabla 6.4

Negativo	Registros negativos
Negativo 4	3
Negativo 5	3
Negativo 6	6
Negativo 7	1
Negativo 8	2
Negativo 9	2
Negativo 10	2
Negativo 11	2
Negativo 12	2
Negativo 13	5
Negativo 14	2
Negativo 15	4
Negativo 16	5
Negativo 17	1
Negativo 18	1
Negativo 19	4
Negativo 20	1

Estos valores fueron imputados mediante *interpolación lineal*, técnica que permitió estimar valores intermedios de forma coherente con la tendencia temporal de la serie, sin introducir distorsiones abruptas y preservando la continuidad de los datos.

6.2.4. Georreferenciación de registros

La distribución espacial de los registros georreferenciados de la Figura 6.1 mostró una concentración predominante de clientes a lo largo del territorio continental chileno, siguiendo la estructura longitudinal del país y reflejando el patrón poblacional característico de las zonas urbanas y costeras. Las mayores densidades se observaron en el centro-sur, especialmente en torno a la Región Metropolitana y las regiones contiguas, donde se concentra la mayor parte de los usuarios del sistema.

Además, se identificaron dos agrupamientos fuera del continente que corresponden a territorios insulares bajo jurisdicción chilena: el punto A, ubicado en la Isla de Pascua, en pleno océano Pacífico, y el punto B, correspondiente al Archipiélago Juan Fernández, situado más próximo a la costa.

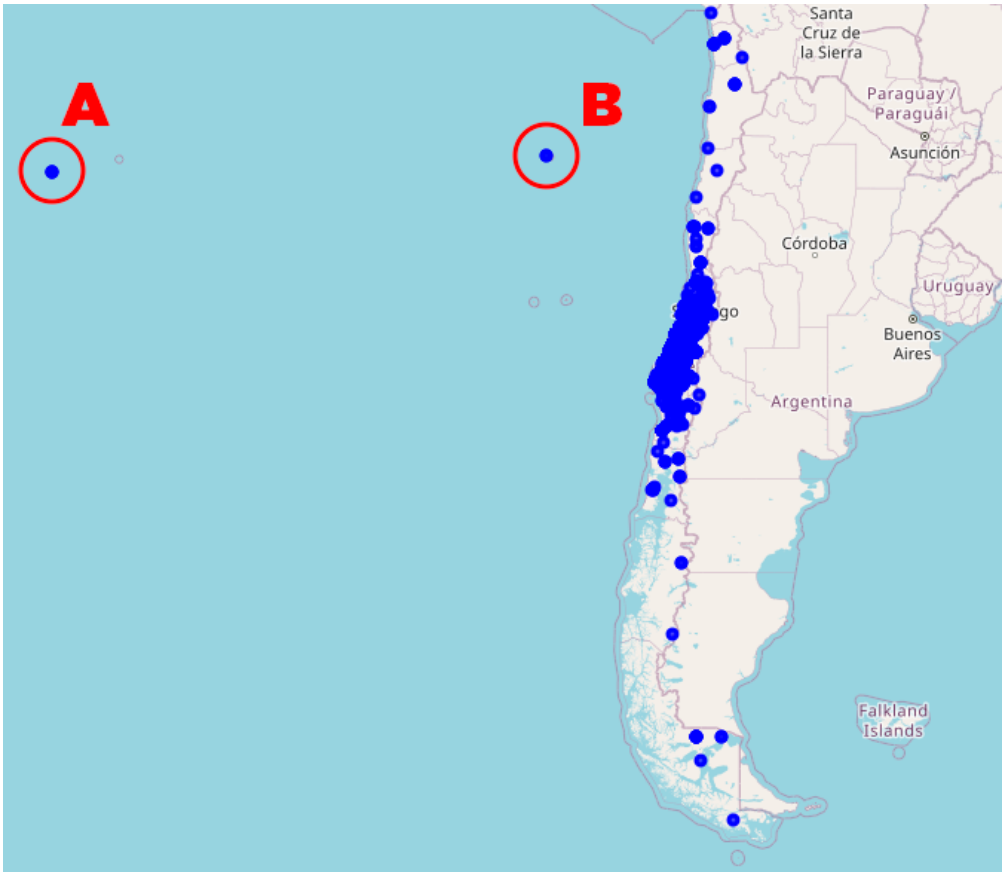

Figura 6.1: Mapa de la distribución de las coordenadas geográficas obtenidas a partir del proceso de geocodificación de direcciones. Fuente: Elaboración propia.

6.3. ANÁLISIS EXPLORATORIO

6.3.1. Estadísticas descriptivas antes y después del tratamiento de datos

En esta sección se presentan las estadísticas descriptivas de las variables numéricas y categóricas incluidas en la base de datos, lo que permite tener una visión general de su distribución y características principales antes y después del tratamiento de los datos. Se exceptúa la variable de consumo, cuyo análisis se desarrolla con mayor detalle en la sección dedicada a series de tiempo, dado su carácter temporal. Asimismo, las coordenadas geográficas de latitud y longitud no son incluidas en este apartado, pues por su naturaleza espacial no resulta pertinente calcular sobre ellas medidas descriptivas tradicionales.

- **Diámetro (variable numérica):**

Tabla 6.5: *Diámetro:* Resumen estadístico antes y después del tratamiento de datos. Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Registros	32.114	32.084
Promedio	14,45	14,46
Desviación	3,63	3,62
Mínimo	2	13
25 %	13	13
50 %	13	13
75 %	13	13
Máximo	100	100

Para el caso de la variable *Diámetro*, cuyos resultados se muestran en la Tabla 6.5, se evidenció que el proceso de imputación no generó cambios relevantes en la estructura estadística de los datos. El número de registros se redujo levemente (de 32.114 a 32.084) como resultado de la depuración, pero tanto el promedio como la desviación estándar permanecieron prácticamente iguales (14,45 a 14,46 en el promedio y 3,63 a 3,62 en la desviación), lo que indicó que la tendencia central y la variabilidad de la variable se conservaron estables. Asimismo, los percentiles 25 %, 50 % y 75 % se mantuvieron constantes en 13, indicando que la concentración de valores alrededor de la mediana no fue afectada por el tratamiento.

El cambio más significativo se presentó en el valor mínimo, que pasó de 2 a 13. Este ajuste responde a la eliminación de los registros de los medidores ficticios que estaban presentes en la base de datos antes del proceso de limpieza y que no correspondían a condiciones reales de los medidores. Con este ajuste, el rango de la variable se alineó con los datos reales, manteniendo el máximo en 100. En conjunto, los resultados evidencian que la imputación fortaleció la calidad de la información corrigiendo errores sin alterar la distribución general de la variable.

- **Año (variable categórica):**

Tabla 6.6: *Año:* Resumen estadístico antes y después del tratamiento de datos. Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	2009	2009
Frecuencia moda	4.135	4.135
Frecuencia relativa	12,88 %	12,89 %
Categorías únicas	42	42
Entropía estandarizada	0,75	0,75

Para el caso de la variable *Año*, cuyos resultados se encuentran en la Tabla 6.6, no se evidenciaron cambios tras el tratamiento de datos. La moda se mantuvo en 2009, con una frecuencia de 4.135 registros, equivalente aproximadamente al 12,9 % del total, prácticamente igual al valor inicial. Asimismo, el número de categorías únicas permaneció en 42 y la entropía estandarizada no presentó

variaciones (0,75). Esto indica que el proceso de tratamiento de datos no afectó la distribución de esta variable y que su estructura general se mantuvo completamente estable.

- **Clase metrológica (variable categórica):**

Tabla 6.7: *Clase metrológica:* Resumen estadístico antes y después del tratamiento de datos. Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	100	100
Frecuencia moda	29.669	29.659
Frecuencia relativa	92,39 %	92,44 %
Categorías únicas	4	3
Entropía estandarizada	0,198	0,246

Para el caso de la variable *Clase metrológica*, cuyos resultados se encuentran en la Tabla 6.7, mostaron que la estructura central se mantuvo tras el tratamiento de datos, dado que la moda continuó siendo la misma (valor 100) y su frecuencia relativa poco varió (92,39 % antes frente a 92,44 % después). La reducción de categorías únicas de 4 a 3 está asociada a la eliminación de registros correspondientes a medidores ficticios, lo cual contribuyó a depurar la base de datos. Como consecuencia, la entropía estandarizada presentó un leve incremento, reflejando una mejor consistencia interna en la distribución de la variable.

- **Localidad (variable categórica):**

Tabla 6.8: *Localidad:* Resumen estadístico antes y después del tratamiento de datos. Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	Concepción	Concepción
Frecuencia moda	2.902	2.898
Frecuencia relativa	9,04 %	9,03 %
Categorías únicas	127	120
Entropía estandarizada	0,8	0,79

Para el caso de la variable *Localidad*, cuyos resultados se encuentran en la Tabla 6.8, su estructura central se mantuvo tras el tratamiento de datos, ya que la localidad más frecuente permaneció inalterada y su peso relativo se mantuvo casi constante (9,04 a 9,03). No obstante, se redujo el número total de categorías y, en consecuencia, la entropía estandarizada, efecto asociado a la eliminación de registros con más del 30 % de datos faltantes.

- **Marca (variable categórica):**

Tabla 6.9: *Marca:* Resumen estadístico antes y después del tratamiento de datos.
Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	CCM-Maipo-Actaris (TMI-TMII)	CCM-Maipo-Actaris (TMI-TMII)
Frecuencia moda	19.641	19.636
Frecuencia relativa	61,17 %	61,20 %
Categorías únicas	31	30
Entropía estandarizada	0,37	0,37

Para el caso de la variable *Marca*, cuyos resultados se muestran en la Tabla 6.9, se observó una disminución leve tanto en la frecuencia de la moda (de 19.641 a 19.636) como en el número de categorías únicas (de 31 a 30). Estos cambios están relacionados con la eliminación de registros que presentaban más del 30 % de datos faltantes, así como con la depuración de medidores ficticios. A pesar de estas modificaciones, la distribución general de la variable se mantuvo estable, lo que se refleja en el resultado de la entropía estandarizada.

- **Ruedas (variable categórica):**

Tabla 6.10: *Ruedas:* Resumen estadístico antes y después del tratamiento de datos.
Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	4	4
Frecuencia moda	22.697	22.685
Frecuencia relativa	70,68 %	70,71 %
Categorías únicas	4	4
Entropía estandarizada	0,45	0,52

Para el caso de la variable *Ruedas*, cuyos resultados se muestran en la Tabla 6.10, la moda se mantuvo en 4 tanto antes como después del tratamiento, reflejando estabilidad en la categoría más frecuente. No obstante, la frecuencia absoluta de la moda presentó una ligera disminución (de 22.697 a 22.685), lo cual está asociado a la eliminación de registros que contenían más del 30 % de datos faltantes. Pese a este ajuste, el número de categorías únicas permaneció en 4 y la entropía estandarizada aumentó de 0,45 a 0,52, indicando una mayor heterogeneidad relativa tras el tratamiento de datos.

- **Transmisión (variable categórica):**

Tabla 6.11: *Transmisión:* Moda y heterogeneidad antes y después del tratamiento. Fuente: Elaboración propia.

Estadístico	Antes del tratamiento	Después del tratamiento
Moda	2	2
Frecuencia moda	30.473	30.458
Frecuencia relativa	94,93 %	94,93 %
Categorías únicas	5	5
Entropía estandarizada	0,144	0,144

Para el caso de la variable *Transmisión*, cuyos resultados se muestran en la Tabla 6.11, se observó que la moda se mantuvo en la categoría 2 antes y después del tratamiento, es decir, hubo estabilidad en la categoría predominante. La frecuencia absoluta de la moda disminuyó levemente (de 30.473 a 30.458), cambio explicado por la eliminación de registros con más del 30 % de datos faltantes. A pesar de esta reducción mínima, el número de categorías únicas permaneció constante en 5 y la entropía estandarizada se mantuvo en 0,144, lo que indicó que la distribución y la heterogeneidad de la variable no se vieron afectadas de manera significativa tras el proceso de depuración.

6.3.2. Análisis univariado

El análisis univariado permitió examinar cada variable de forma independiente para identificar su distribución y características principales. Este paso facilitó la detección de concentraciones, valores dominantes y posibles diferencias entre la base de fraude y la de anomalías, proporcionando una primera aproximación a los patrones presentes en los datos.

- **Diámetro (variable numérica):**

Tabla 6.12: Distribución de clientes según la variable *Diámetro* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Diámetro	#	%
13 mm	25.625	79,87 %
19 mm	5.605	17,47 %
38 mm	90	0,28 %
25 mm	708	2,21 %
50 mm	39	0,12 %
75 mm	16	0,05 %
100 mm	1	0,003 %

(b)

Diámetro	#	%
13 mm	101	87,07 %
19 mm	15	12,93 %
38 mm	0	0,00 %
25 mm	0	0,00 %
50 mm	0	0,00 %
75 mm	0	0,00 %
100 mm	0	0,00 %

Para el caso de la variable *Diámetro*, cuyos resultados se muestran en la Tabla 6.12, evidenciaron que la gran mayoría de los clientes, tanto en la base de fraude como en la de anomalías, correspondieron a medidores de diámetro 13 mm, lo que refleja un claro predominio del uso residencial o doméstico. En la base general, este grupo concentró cerca del 80 % de los registros, mientras que en la base de anomalías representó más del 87 %, lo que indicó que las irregularidades detectadas se concentraron

principalmente en el mismo segmento de usuarios con medidores pequeños. Los diámetros superiores a 19 mm mostraron una participación marginal en ambos casos, y en la base de anomalías prácticamente desaparecieron.

- **Año (variable categórica):**

Tabla 6.13: Distribución de clientes según la variable *Año* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Año	#	%
Otros	508	1,58 %
1991	58	0,18 %
1992	67	0,21 %
1993	106	0,33 %
1994	135	0,42 %
1995	205	0,64 %
1996	290	0,90 %
1997	550	1,71 %
1998	702	2,19 %
1999	792	2,47 %
2000	1.236	3,85 %
2001	2.851	8,89 %
2002	3.153	9,83 %
2003	1.528	4,76 %
2004	1.073	3,34 %
2005	1.422	4,43 %
2006	1.397	4,35 %
2007	2.200	6,86 %
2008	2.817	8,78 %
2009	4.135	12,89 %
2010	3.055	9,52 %
2011	2.445	7,62 %
2012	1.135	3,54 %
2013	224	0,70 %

(b)

Año	#	%
Otros	0	0,00 %
1991	0	0,00 %
1992	0	0,00 %
1993	1	0,88 %
1994	1	0,88 %
1995	1	0,88 %
1996	3	2,63 %
1997	2	1,75 %
1998	7	6,14 %
1999	0	0,00 %
2000	2	1,75 %
2001	8	7,02 %
2002	12	10,53 %
2003	0	0,00 %
2004	6	5,26 %
2005	4	3,51 %
2006	3	2,63 %
2007	4	3,51 %
2008	6	5,26 %
2009	12	10,53 %
2010	11	9,65 %
2011	28	24,56 %
2012	3	2,63 %
2013	0	0,00 %

Para el caso de la variable *Año*, cuyos resultados se muestran en la Tabla 6.13, evidenciaron que los casos de fraude se concentraron entre 2008–2010, con un pico en 2009 (12,89 %), mientras que para las anomalías se observó una marcada concentración en 2011 (24,56 %), indicando un patrón temporal más acotado hacia el final del periodo.

- **Clase metrológica (variable categórica):**

Tabla 6.14: Distribución de clientes según la variable *Clase metrológica* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Clase metrológica	#	%
100	29.659	92,44 %
102	2.412	7,52 %
103	13	0,04 %

(b)

Clase metrológica	#	%
100	85	73,28 %
102	31	26,72 %
103	0	0,00 %

Para el caso de la varibale *Clase metrológica*, cuyos resultados se muestran en la Tabla 6.14, evidenciaron que la mayoría de los medidores, en ambas bases, pertenecían a la clase 100, evidenciando una alta homogeneidad en el tipo de equipos instalados. La base de anomalías presentó una mayor proporción relativa de medidores clase 102, lo que sugiere una ligera concentración de casos en este grupo. La clase 103 tuvo una participación mínima y sin relevancia en los resultados.

- **Localidad (variable categórica):**

Tabla 6.15: Distribución *top 20* de clientes según la variable *Localidad* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Localidad	#	%
Concepción	2.898	9,03 %
Talcahuano	2.493	7,77 %
Chillán	2.315	7,22 %
Talca	2.054	6,40 %
Rancagua	1.915	5,97 %
Los Ángeles	1.671	5,21 %
Coronel	1.212	3,78 %
Curicó	1.173	3,66 %
San Pedro	1.158	3,61 %
Chiguayante	857	2,67 %
San Fernando	814	2,54 %
Linares	782	2,44 %
Lota	520	1,62 %
Machalí	515	1,61 %
Rengo	489	1,52 %
Tome	469	1,46 %
Cauquenes	373	1,16 %
Penco	372	1,16 %
Curanilahue	369	1,15 %
Santa Cruz	310	0,97 %

(b)

Localidad	#	%
Arauco	8	0,025 %
Concepción	7	0,022 %
Yungay	5	0,016 %
Dichato	5	0,016 %
El Carmen	5	0,016 %
Ramadillas	5	0,016 %
Talcahuano	4	0,012 %
Chiguayante	4	0,012 %
Penco	4	0,012 %
Chillán	3	0,009 %
San Pedro	3	0,009 %
Cañete	3	0,009 %
Lebu	3	0,009 %
Coihueco	3	0,009 %
Los Álamos	3	0,009 %
Cerro Alto	3	0,009 %
San Rosendo	3	0,009 %
Rafael	3	0,009 %
Los Ángeles	2*	0,006 %
Curanilahue	2*	0,006 %

*La base de datos de anomalía incluye 13 localidades adicionales con 2 clientes: Cabrero, Carampangue, Cobquecura, Contulmo, Laja, Lirquén, Lomas Coloradas, Mulchén, Quirihue, San Carlos, San Ignacio, Santa Bárbara y Santa Juana.

Para el caso de la variable *Localidad*, cuyos resultados se muestran Tabla 6.15, evidenciaron en la una clara diferencia en la distribución geográfica entre ambas bases. En la base de fraude, las localidades con mayor número de clientes se concentraron en ciudades de gran tamaño como Concepción, Talcahuano y Chillán, lo que refleja una relación directa con la densidad poblacional y el volumen de usuarios atendidos. En contraste, la base de anomalías presentó una dispersión marcada, con casos aislados en múltiples localidades de menor tamaño, donde ninguna concentró una proporción significativa de registros. Esto indica que las anomalías no se concentraron en zonas específicas, sino que se distribuyeron de forma más aleatoria en distintas áreas.

- **Marca (variable categórica):**

Tabla 6.16: Distribución *top 5* de clientes según la variable *Marca* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Marca	#	%
CCM-Maipo-Actaris (TMI-TMII)	19.636	61,2 %
Lautaro-Invensys-Sensus	6.628	20,7 %
MM-NO VIGENTE	1.780	5,5 %
CCM-Maipo-Actaris	795	2,48 %
Tavira-Iberconta-abb	788	2,46 %

(b)

Marca	#	%
CCM-Maipo-Actaris (TMI-TMII)	70	60,34 %
Lautaro-Invensys-Sensus	32	27,59 %
MR-NO VIGENTE	7	6,03 %
CCM-Maipo-Actaris	4	3,45 %
Tavira-Iberconta-abb	1*	0,86 %

*La base de datos de anomalías incluye 2 marcas adicionales con un solo cliente: Elster (Ex Tavira) y CCM-Actaris (Flostar M).

Para el caso de la variable *Marca*, cuyos resultados se muestran en la Tabla 6.16, evidenciaron que la marca CCM-Maipo-Actaris (TMI-TMII) concentró la mayoría de los medidores tanto en la base de fraude (61,2 %) como en la de anomalías (60,34 %), lo que evidencia su amplia presencia en el conjunto analizado. La segunda marca, Lautaro-Invensys-Sensus, también mantuvo una proporción similar en ambas bases, alrededor del 20 % y 27 %, respectivamente. Las demás marcas presentaron participaciones marginales, lo que indica que las irregularidades detectadas se concentraron principalmente en los mismos tipos de medidores más utilizados, sin un sesgo claro hacia marcas minoritarias.

- **Ruedas (variable categórica):**

Tabla 6.17: Distribución de clientes según la variable *Ruedas* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Ruedas	#	%
4	22.685	70,71 %
6	59	0,18 %
5	8.234	25,66 %
7	1.106	3,45 %

(b)

Ruedas	#	%
4	77	66,38 %
6	0	0 %
5	29	25 %
7	10	8,60 %

Para el caso de la variable *Ruedas*, cuyos resultados se muestran en la Tabla 6.17, mostraron que la mayoría de los medidores, tanto en la base de fraude como en la de anomalías, correspondieron a

equipos con 4 ruedas, con participaciones del 70,71 % y 66,38 %, respectivamente. Las configuraciones con 5 y 7 ruedas presentaron una participación menor, aunque mantuvieron proporciones similares en ambas bases, mientras que los medidores de 6 ruedas no estuvieron presentes en los casos anómalos.

- **Transmisión (variable categórica):**

Tabla 6.18: Distribución de clientes según la variable *Transmisión* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Transmisión	#	%
1	1.167	3,64 %
2	30.458	94,93 %
507	1	0,003 %
511	457	1,42 %
512	1	0,003 %

(b)

Transmisión	#	%
1	7	6,03 %
2	97	83,62 %
507	0	0 %
511	12	10,34 %
512	0	0 %

Para el caso de la variable *Transmisión*, cuyos resultados se muestran en la Tabla 6.18, evidenciaron que la categoría 2 concentró la mayoría de los registros tanto en la base de fraude (94,93 %) como en la de anomalías (83,62 %), indicando una distribución similar entre ambas. Sin embargo, se observó un ligero aumento en la proporción de la categoría 1 y 511 dentro de la base de anomalías, lo que podría indicar que las irregularidades se presentaron con mayor frecuencia en medidores con estos tipos de transmisión, aunque su participación general siguió siendo minoritaria.

- **Fraude y Anomalías (variables de respuesta):**

Tabla 6.19: Distribución de registros en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Indicador	#	%
Fraude	531	1,65 %
No fraude	31.553	98,35 %

(b)

Indicador	#	%
Anomalía	84	27,59 %
No anomalía	32	72,41 %

Para el caso de la variables de respuesta, cuyos resultados se muestran en la Tabla 6.19, evidenciaron una marcada desproporción en la cantidad de registros entre las clases. En la base de datos de fraude, únicamente el 1,65 % de los casos correspondió a eventos fraudulentos, mientras que el 98,35 % representó registros normales. Esta diferencia indicó que los fraudes fueron eventos poco frecuentes dentro del conjunto analizado, lo cual coincidió con el comportamiento esperado en contextos reales de detección de fraude, donde las ocurrencias irregulares suelen ser excepcionales.

En contraste, en la base de datos de anomalías el 72,41 % de los registros se clasificó como no anómalo y el 27,59 % como anómalo. Al igual que el fraude, en las anomalías las ocurrencias irregulares también suelen ser excepcionales.

6.3.3. Análisis multivariado

El análisis multivariado permitió examinar la relación conjunta entre múltiples variables con el propósito de identificar patrones de asociación, dependencias y comportamientos compartidos entre los datos. Este paso posibilitó detectar combinaciones de atributos que distinguieron los registros de fraude y los de anomalías, así como posibles correlaciones que no fueron evidentes en el análisis univariado.

- **Diámetro (variable numérica) vs. variables de respuesta:**

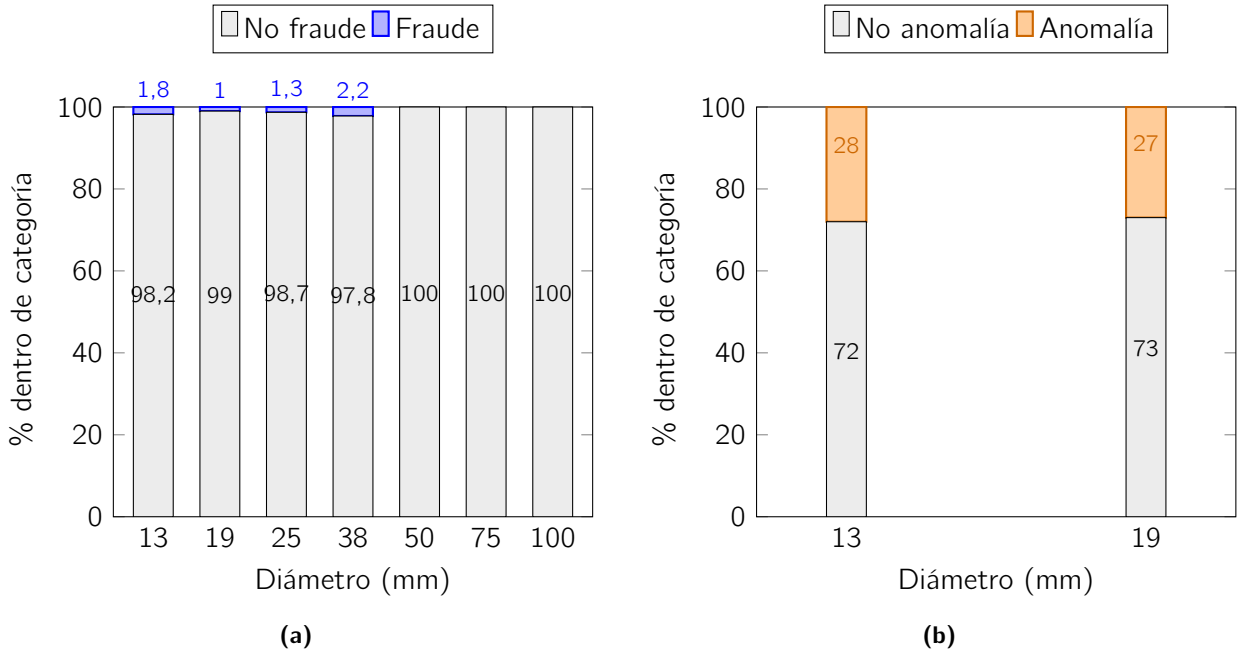

Figura 6.2: Distribución porcentual de **(a)** fraude y **(b)** anomalía según el *Diámetro*.
Fuente: Elaboración propia.

Para el caso de la variable *Diámetro*, cuyos resultados se muestran en la Figura 6.2, evidenciaron que la mayoría de los medidores en la base de fraude correspondió a tamaños pequeños, especialmente de 13 mm y 19 mm, los cuales concentraron la mayor cantidad de registros. En ambos casos, el porcentaje de fraude fue muy bajo (entre 1 % y 2 %), lo que sugiere que la incidencia de fraudes no estuvo asociada directamente con el tamaño del medidor.

En contraste, la base de anomalías presentó un comportamiento distinto: los medidores de 13 mm y 19 mm registraron valores de 28 % y 27 % de anomalías, respectivamente. Esto indica que, aunque los fraudes fueron escasos en estos diámetros, las anomalías se concentraron en los medidores más pequeños, posiblemente debido a su mayor presencia en la red o a condiciones operativas más variables.

- **Año (variable categórica) vs. variables de respuesta:**

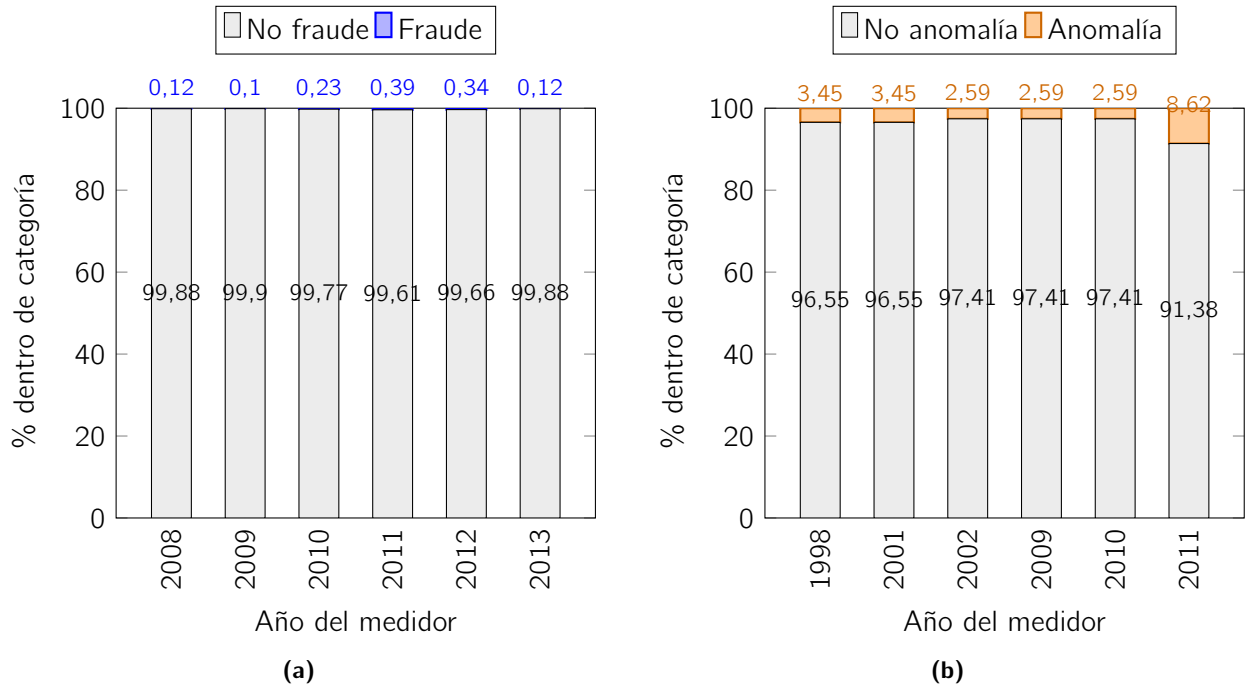

Figura 6.3: Distribución porcentual *top 5* de **(a)** fraude y **(b)** anomalía según el *Año del medidor*.
Fuente: Elaboración propia.

Para el caso de la variable *Año*, cuyos resultados se muestran en la Figura 6.3, evidenciaron que la incidencia de fraude fue superior en los periodos más recientes (2008–2013). Sin embargo, sigue siendo un fenómeno de baja incidencia (menor al 0,4 %). En contraste, las anomalías muestran una variabilidad a lo largo de los periodos y su máximo se da en 2011 (8,62 %), es decir, también dentro de los periodos recientes.

- **Clase metrológica (variable categórica) vs. variables de respuesta:**

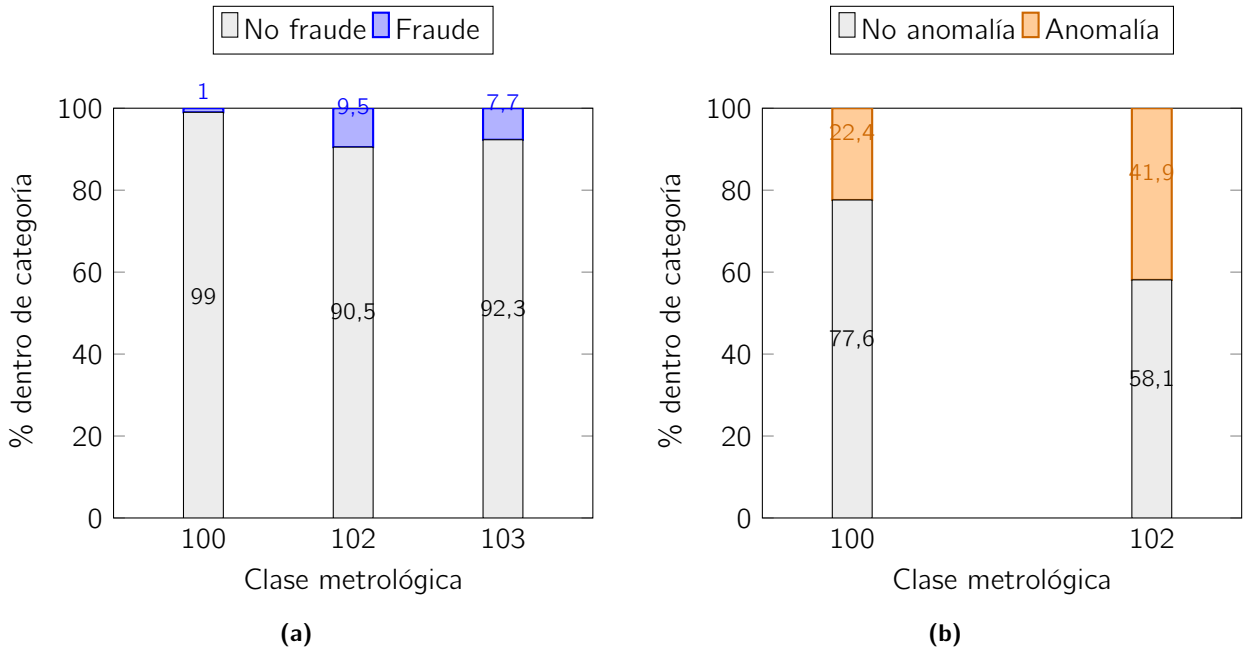

Figura 6.4: Distribución porcentual de **(a)** fraude y **(b)** anomalía según la *Clase metrológica*.
Fuente: Elaboración propia.

Para el caso de la variable *Clase metrológica*, cuyos resultados se muestran en la Figura 6.4, mostraron que la gran mayoría de los medidores en ambas bases pertenecieron a la clase metrológica 100, lo que reflejó la predominancia de este tipo de dispositivos en el conjunto. En el caso del fraude, aunque los porcentajes fueron bajos en general, se observó una ligera variación entre clases: los medidores de clase 103 presentaron un 7,7 % de fraude, superior al 1 % registrado en la clase 100, lo que pudo estar relacionado con diferencias en su uso, ubicación o mantenimiento.

En cuanto a las anomalías, la clase 102 tuvo una mayor proporción (41,9 %) frente a la clase 100 (22,4 %), lo que indicó una posible mayor sensibilidad o propensión de estos equipos a registrar comportamientos fuera de lo esperado. En conjunto, los datos evidenciaron que la clase metrológica pudo influir en la frecuencia tanto de fraudes como de anomalías, posiblemente debido a las características técnicas o al contexto de operación de cada tipo de medidor.

- **Localidad (variable categórica) vs. variables de respuesta:**

Fraude:

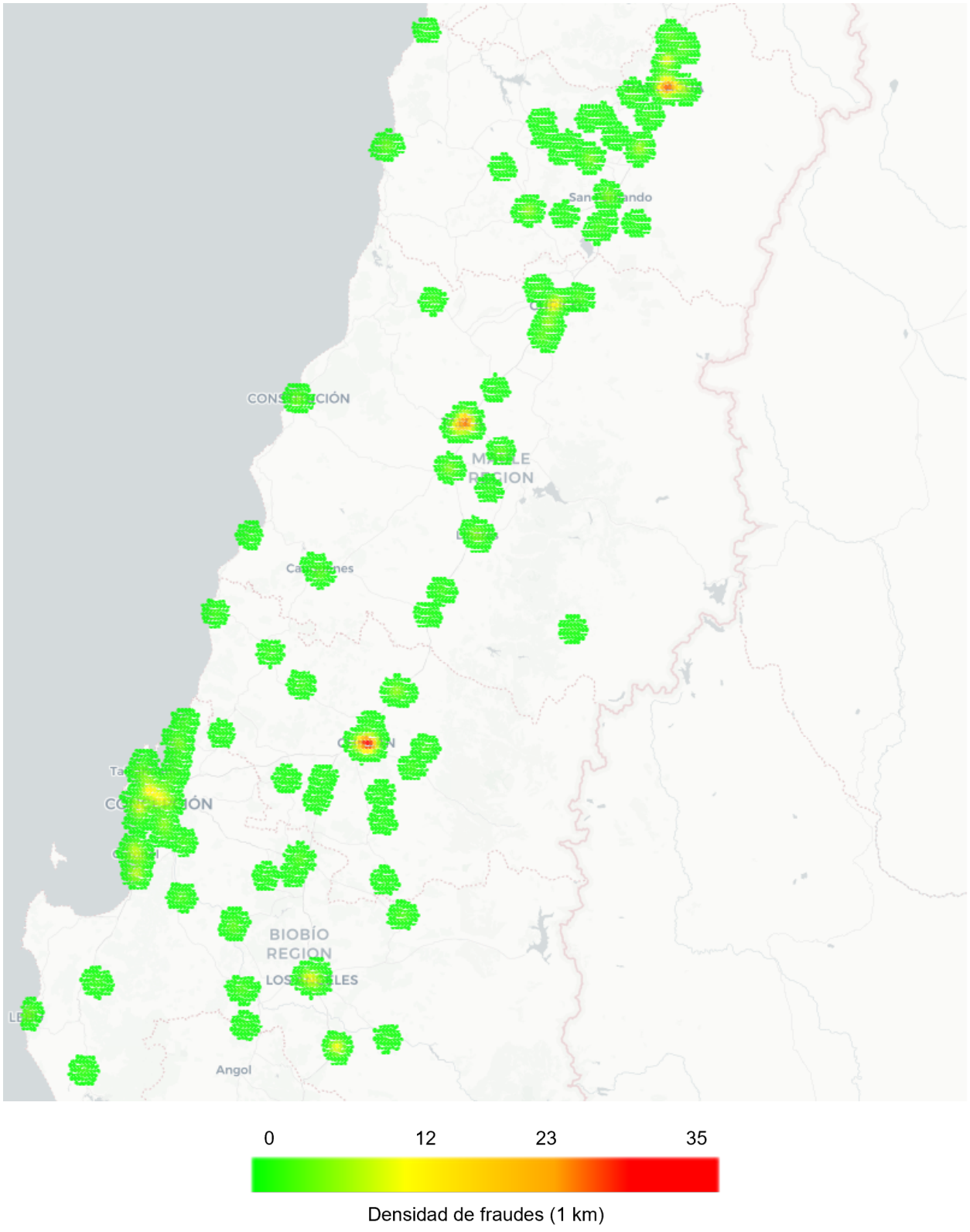

Figura 6.5: Kernel de densidad de fraudes (radio 1 km) a partir del número de casos por área. Fuente: Elaboración propia.

El mapa del kernel de densidad de fraude a partir del número de casos por área (radio de 1 km), mostró que los casos de fraude se distribuyen a lo largo del territorio chileno, y no en una zona específica. En términos generales, la mayor parte de los hexágonos presentó valores bajos, representados en color verde. No obstante, se identificaron áreas específicas con concentraciones elevadas, evidenciadas por tonalidades amarillas y rojas.

La zona con mayor intensidad correspondió a la costa centro-sur, particularmente en el área metropolitana de Concepción y sus alrededores, donde se observaron agrupaciones continuas de hexágonos en colores de alta densidad. Este patrón sugirió una concentración significativa de fraudes en entornos urbanos y periurbanos de alta densidad poblacional. Adicionalmente, se identificaron focos secundarios en las regiones del Maule y Ñuble, principalmente en sectores interiores y urbanos de tamaño intermedio, donde la densidad se mantuvo en rangos medios. Finalmente, hacia la zona sur (Biobío interior, Los Ángeles y Angol), los casos se encontraron más dispersos, con prevalencia de densidades bajas y ausencia de conglomerados marcados.

En complemento al análisis espacial, la estimación de la prevalencia directa por localidad permitió matizar los hallazgos. Si bien el mapa de densidad reflejó una alta concentración de fraudes en áreas metropolitanas como Concepción, la prevalencia directa por localidad evidenció que en comunas pequeñas, aunque con menor volumen absoluto de casos, la proporción de clientes afectados resultó mayor. La Tabla 6.20 mostró que localidades con menor número de clientes, tales como Ninhue (9,68 %) o Ñipas (8,00 %), alcanzaron proporciones relativas de fraude incluso más altas. En contraste, comunas de gran tamaño como Chillán (3,46 %) y Talca (2,87 %) registraron prevalencias más bajas, aunque con volúmenes absolutos de fraude relevantes. Esto indicó que el fenómeno combinó dos dinámicas: por un lado, altas concentraciones absolutas en áreas metropolitanas, y por otro, altas prevalencias relativas en localidades pequeñas.

Tabla 6.20: Top 20 de prevalencia directa de fraude por *Localidad*.
Fuente: Elaboración propia.

Localidad	Clientes con fraude	Clientes totales	Prevalencia fraude
Ninhue	3	31	9,68 %
Ñipas	2	25	8,00 %
Lontue	6	88	6,82 %
Requinoa	6	98	6,12 %
Retiro	2	35	5,71 %
Graneros	16	307	5,21 %
Puente Negro	1	20	5,00 %
Codegua	4	84	4,76 %
Romeral	3	63	4,76 %
La Punta	3	67	4,48 %
San Francisco	5	115	4,35 %
Mulchen	13	308	4,22 %
Peralillo	2	54	3,70 %
Chillán	80	2.315	3,46 %
San Rafael	1	32	3,13 %
Coltauco	3	99	3,03 %

Continuación Tabla 6.20

Localidad	Clientes con fraude	Clientes totales	Prevalencia
Santa Bárbara	3	102	2,94 %
San Javier	7	240	2,92 %
Talca	59	2.056	2,87 %
Lo Miranda	3	107	2,80 %

Anomalías:

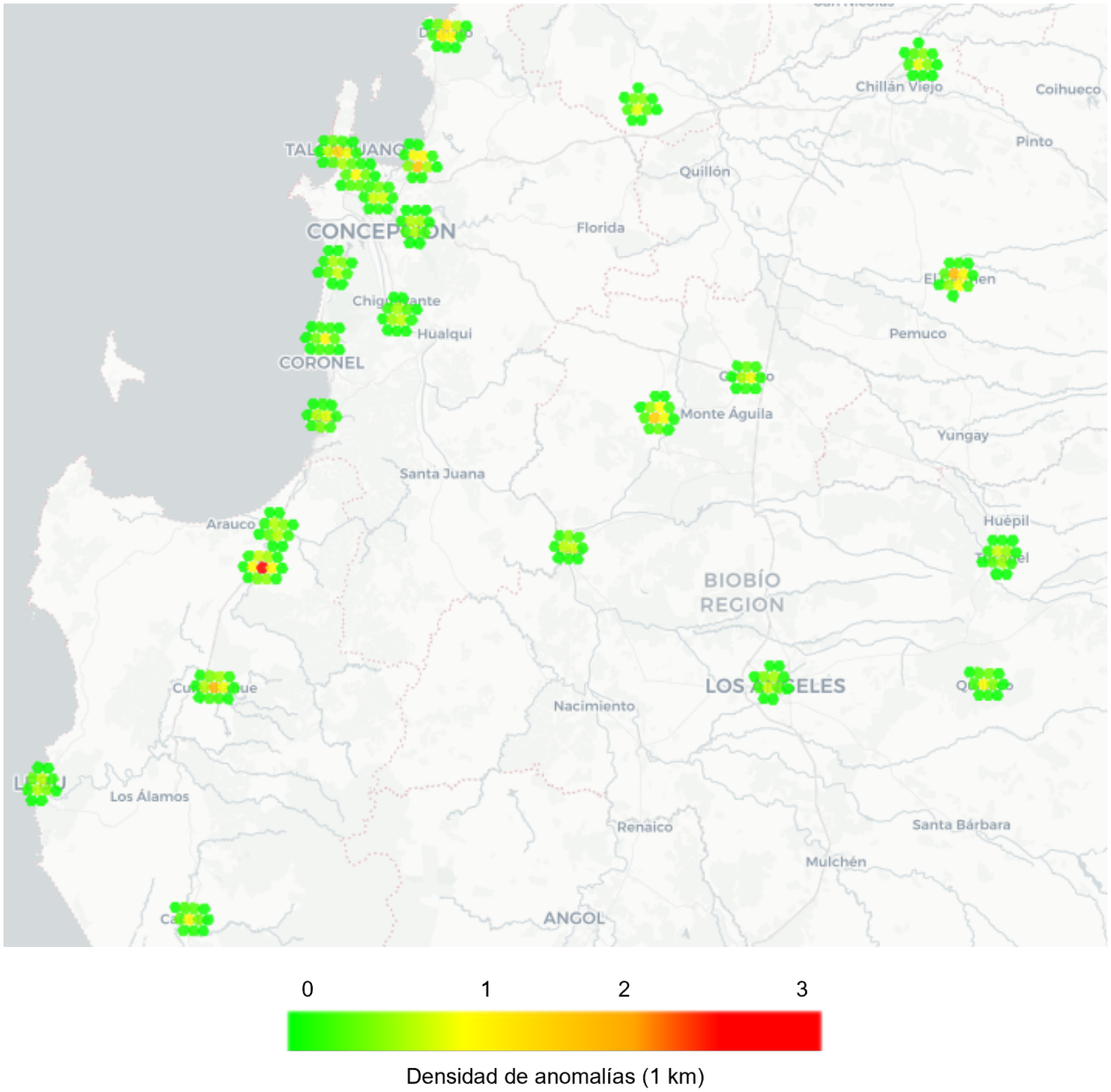

Figura 6.6: Kernel de la densidad de anomalías (radio 1 km) a partir del número de casos por área. Fuente: Elaboración propia.

El mapa del kernel de densidad de anomalías a partir del número de casos por área (radio de 1 km), al igual que en el caso de fraude, también mostró que los casos de anomalías se distribuyen a lo

largo del territorio chileno, y no en una zona específica. En general, la mayor parte de los hexágonos presentó valores bajos, representados en color verde. Sin embargo, se identificaron áreas puntuales con densidades más elevadas, evidenciadas por tonalidades amarillas y rojas.

La concentración principal se localizó en el área metropolitana de Concepción y sus alrededores, donde se observaron varias celdas con densidad superior al promedio regional. Este patrón indicó que las anomalías tendieron a acumularse en entornos urbanos de mayor tamaño poblacional y con infraestructura más compleja. Adicionalmente, se observaron focos adicionales en localidades intermedias como Arauco, Monte Águila y Los Ángeles, donde si bien los niveles absolutos fueron menores, la densidad local alcanzó valores relevantes en comparación con el resto del territorio.

En contraste, en las zonas periféricas y rurales los casos se distribuyeron de manera más dispersa, predominando hexágonos con valores cercanos a cero y ausencia de conglomerados de alta densidad. En complemento al análisis espacial, la estimación de la prevalencia directa de anomalías por localidad permitió profundizar en la magnitud relativa del fenómeno. Mientras el mapa de densidad reflejó la existencia de focos concentrados en áreas metropolitanas como Concepción y ciudades intermedias de la región del Biobío, la Tabla 6.21 mostró que localidades con muy pocos clientes alcanzaron porcentajes de prevalencia excepcionalmente altos. Casos como Lota, Yumbel, Tucapel o Quilleco registraron un 100 % de prevalencia al presentar un único cliente con anomalía, lo que evidenció la sensibilidad de este indicador en comunas de baja cobertura. Asimismo, localidades como Lebu (67 %), Ramadillas (60 %) y Concepción (57 %) también destacaron con valores superiores al 50 %, lo que reforzó la idea de que el análisis proporcional reveló dinámicas distintas a las observadas en el conteo absoluto de anomalías.

Tabla 6.21: Top 20 de prevalencia directa de anomalía por *Localidad*.
Fuente: Elaboración propia.

Localidad	Clientes con anomalía	Clientes totales	Prevalencia anomalía
Lota	1	1	100 %
Yumbel	1	1	100 %
Tucapel	1	1	100 %
Quilleco	1	1	100 %
Ñipas	1	1	100 %
Lebu	2	3	67 %
Ramadillas	3	5	60 %
Concepción	4	7	57 %
Talcahuano	2	4	50 %
Curanilahue	1	2	50 %
Penco	2	4	50 %
Los Ángeles	1	2	50 %
Carampangue	1	2	50 %
Cabrero	1	2	50 %
Laja	1	2	50 %
San Ignacio	1	2	50 %
Lomas Coloradas	1	2	50 %
Dichato	2	5	40 %
El Carmen	2	5	40 %

Continuación Tabla 6.21

Localidad	Clientes con anomalía	Clientes totales	Prevalencia anomalía
Cañete	1	3	33 %

- **Marca (variable categórica) vs. variables de respuesta**

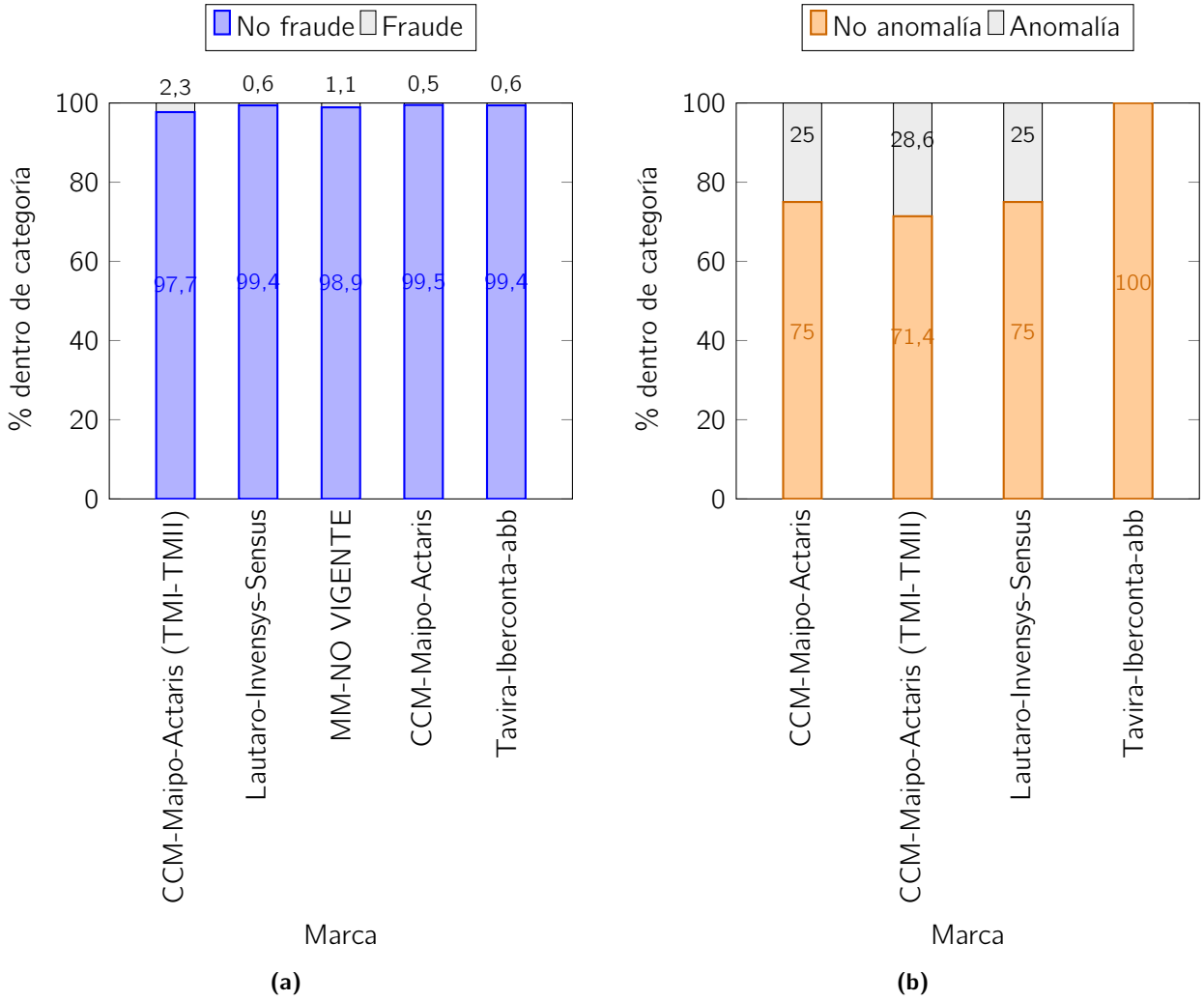

Figura 6.7: Distribución porcentual de **(a)** fraude y **(b)** anomalía según la *Marca* del medidor (top 5). Fuente: Elaboración propia.

Los resultados de la variable *Marca* mostraron que el fraude fue bajo en todas las categorías, con porcentajes inferiores al 3 %, lo que reflejó una distribución homogénea y sin concentraciones significativas por fabricante. Sin embargo, se observó que la marca CCM–Maipo–Actaris (TMI–TMII) presentó el valor más alto (2,3 %), lo que podría deberse a su amplia representación dentro del conjunto.

En contraste, las anomalías presentaron una mayor variabilidad entre marcas: mientras la marca Tavira–Iberconta–abb alcanzó el 100 % de registros anómalos, las marcas CCM–Maipo–Actaris y Lautaro–Invensys–Sensus registraron entre el 25 % y el 28,6 %. Esto indicó que ciertas marcas pudieron haber estado más expuestas a condiciones anómalos, sin que ello implicara necesariamente una relación directa con el fraude.

- **Ruedas (variable categórica) vs. variables de respuesta**

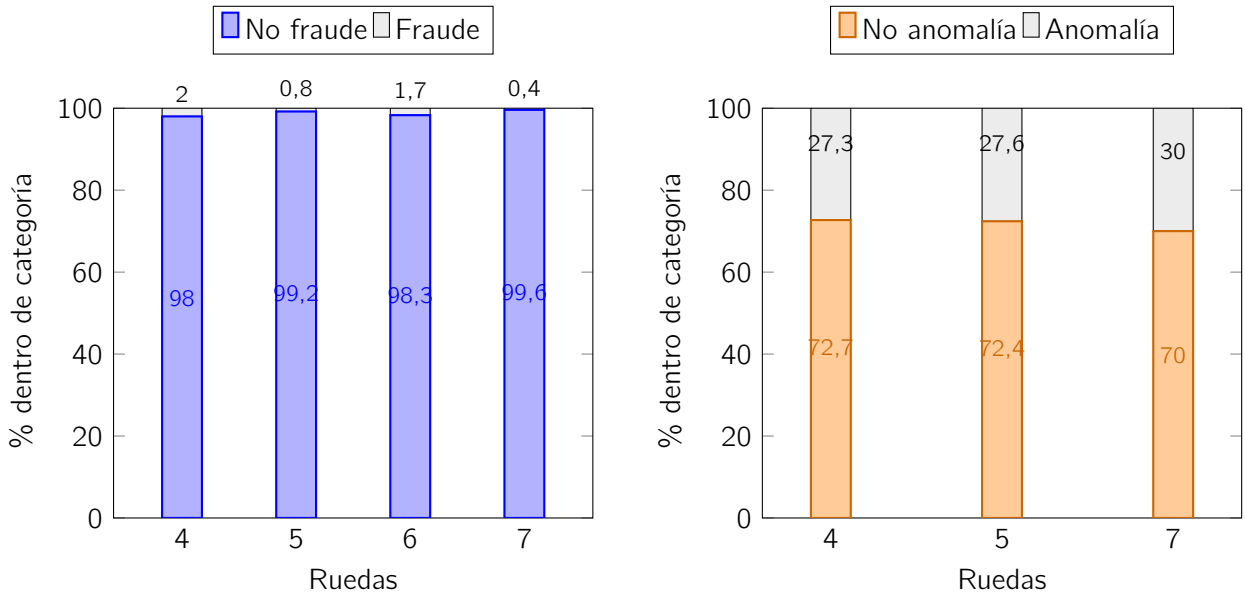

Figura 6.8: Distribución porcentual de **(a)** fraude y **(b)** anomalía según las *Ruedas* del medidor. Fuente: Elaboración propia.

El análisis del número de ruedas mostró que la ocurrencia de fraude fue prácticamente nula en todas las categorías, con valores por debajo del 2 %, lo que evidenció que esta característica estructural del medidor no tuvo una relación significativa con los casos de manipulación detectados. En contraste, las anomalías presentaron una variación ligeramente mayor: mientras que en medidores de 4 y 5 ruedas los valores se mantuvieron alrededor del 27 %, en los de 7 ruedas alcanzaron el 30 %. Este comportamiento indicó que, si bien el número de ruedas no se vinculó con prácticas fraudulentas, podría estar relacionado con una mayor susceptibilidad a fallos técnicos o desajustes operativos en ciertos tipos de dispositivos.

6.3.4. Análisis de consumos

El análisis de consumos permitió examinar el comportamiento histórico del uso total de agua a lo largo del tiempo, identificando tendencias, patrones estacionales y valores atípicos que afectaron la representación del fenómeno. En esta sección se presentó la evaluación del consumo total anual y mensual, con el propósito de reconocer periodos de estabilidad e incrementos sostenidos del comportamiento general del consumo.

- **Análisis de consumo anual total**

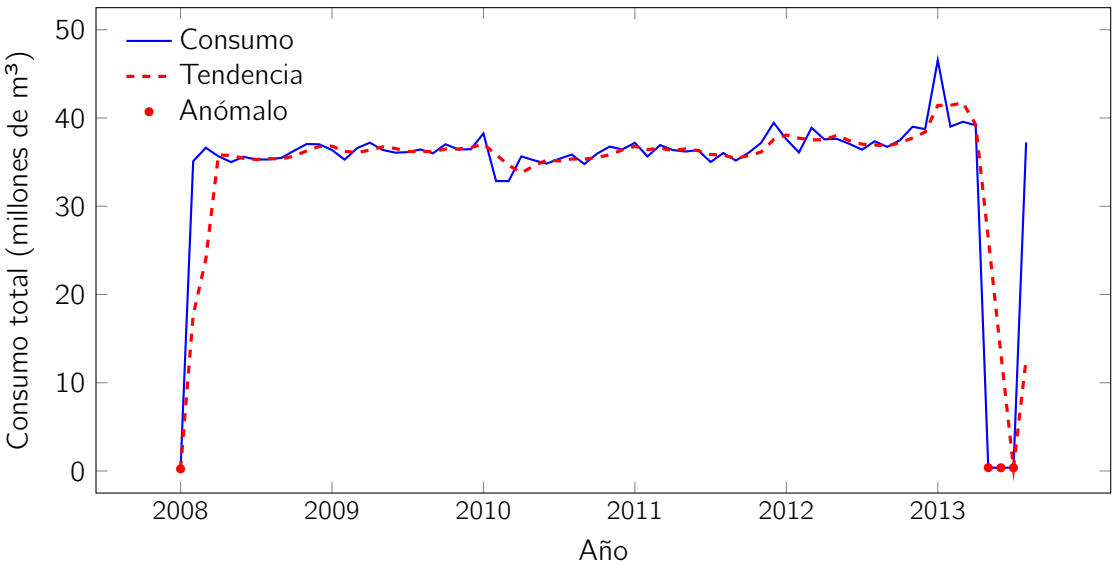

Figura 6.9: Serie de tiempo del consumo total mensual, su tendencia y la identificación de valores anómalos. Fuente: Elaboración propia.

Al analizar los 68 meses de consumo total, se identificó que febrero de 2008 y los meses comprendidos entre junio y agosto de 2013 presentaron valores anómalos (menos de 1 millón de m^3) en comparación con la tendencia general de la serie. En particular, septiembre de 2013 no mostró un comportamiento atípico; sin embargo, también se excluyó con el fin de mantener la continuidad del análisis y evitar la generación de un quiebre artificial en la serie temporal. De esta forma, el periodo de estudio para los modelos pasó de 68 a 63 meses de consumo mensual, garantizando así una representación más consistente y robusta del comportamiento observado.

Al eliminar estos registros, el nuevo comportamiento de la serie mostró una tendencia al alza más estable (Figura 6.10), con valores de consumo mensual comprendidos principalmente entre 32 y 39 millones de m^3, y picos puntuales cercanos a 46 millones de m^3 en 2013. Este ajuste permitió identificar con mayor claridad la tendencia general del consumo, sin las distorsiones introducidas por los valores atípicos.

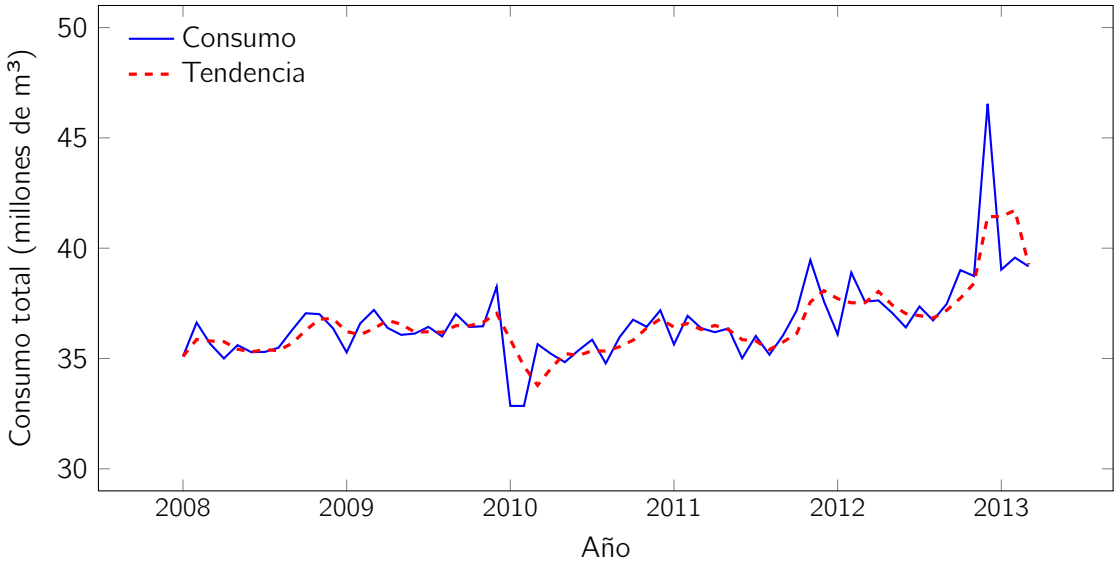

Figura 6.10: Serie de tiempo del consumo total mensual y su tendencia entre marzo de 2008 y mayo de 2013. Fuente: Elaboración propia.

Posteriormente, el análisis del componente estacional presentado en la Figura 6.11 evidenció un patrón recurrente a lo largo de los años: los valores más altos se concentraron en los meses de diciembre, enero y, especialmente, febrero, alcanzando este último el máximo de la estacionalidad con incrementos cercanos a 2,38 millones de m^3. En contraste, los meses comprendidos entre julio y octubre presentaron los niveles más bajos, con variaciones negativas que oscilaron entre –0,28 y –0,93 millones de m^3. Este comportamiento resultó consistente con la estacionalidad climática de Chile [93], dado que los mayores consumos coincidieron con el periodo de verano (diciembre a marzo), caracterizado por temperaturas más elevadas y una mayor demanda de agua, mientras que los descensos se alinearon con los meses de invierno y principios de primavera.

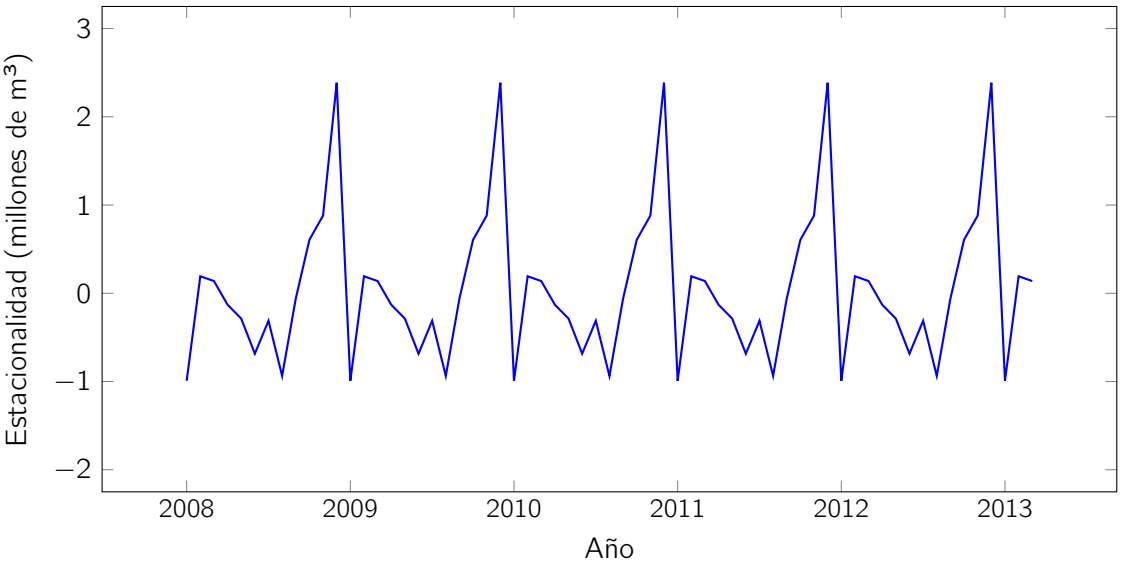

Figura 6.11: Componente estacional estimada de la serie de consumo mensual entre marzo de 2008 y mayo de 2013. Fuente: Elaboración propia.

Por último, la serie de residuos del consumo mensual mostró, en general, valores cercanos a cero y sin un patrón sistemático, lo que indicó que el modelo aditivo logró capturar adecuadamente la tendencia y la estacionalidad de la serie (Figura 6.12). Sin embargo, se detectaron algunos picos atípicos que

sobresalieron respecto al resto de observaciones.

En particular, se registraron valores negativos pronunciados en febrero de 2009 (-2,21 millones de m^3), marzo (-1,76 millones de m^3) y abril de 2010 (-2,91 millones de m^3), así como en febrero de 2011 (-1,35 millones de m^3) y febrero de 2012 (-1,77 millones de m^3). Estos episodios reflejaron meses en los que el consumo real fue significativamente inferior al explicado por la tendencia y la estacionalidad.

Por otro lado, en febrero de 2013 se observó un residuo positivo de aproximadamente 5,16 millones de m^3, constituyendo la desviación más extrema en toda la serie, lo que señaló un consumo muy superior al esperado para ese periodo. La alternancia de residuos positivos y negativos, junto con la ausencia de un patrón de autocorrelación visible, respaldó la validez del ajuste del modelo.

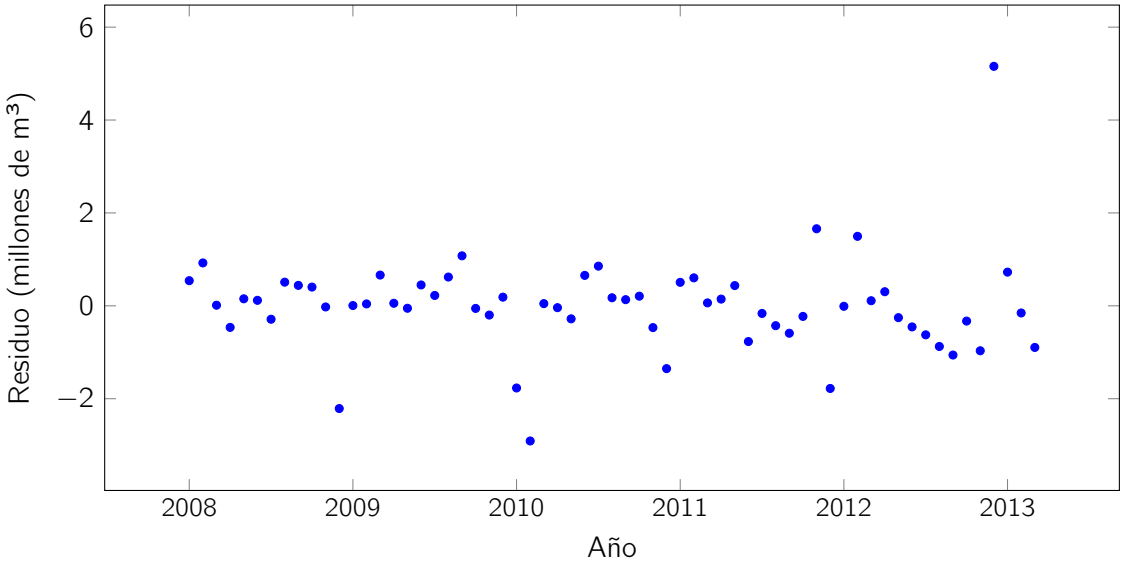

Figura 6.12: Serie de residuos del consumo mensual (modelo aditivo) entre marzo de 2008 y mayo de 2013.
Fuente: Elaboración propia.

- **Análisis de consumo anual mensual**

Tabla 6.22: Medidas mensuales de consumo por cliente.
Fuente: Elaboración propia.

Fecha	N° clientes	Media	Mediana	Q1	Q3	Mínimo	Máximo	Coef. Var.	Atípicos	Anómalos
Mar. 2008	32.084	1.093,90	36	14	1.614	0	5.186	146,17 %	12,07 %	0 %
Abr. 2008	32.084	1.141,63	34	13	1.631	0	5.444	147,68 %	12,64 %	0 %
May. 2008	32.084	1.111,45	29	12	1.715	0	5.079	146,74 %	11,34 %	0 %
Jun. 2008	32.084	1.090,93	27	11	1.671	0	4.833	146,11 %	9,21 %	0 %
Jul. 2008	32.084	1.109,75	26	11	1.674	0	4.987	146,42 %	11,91 %	0 %
Ago. 2008	32.084	1.099,93	26	11	1.684	0	4.876	145,47 %	9,15 %	0 %
Sept. 2008	32.084	1.100,20	27	11	1.686,25	0	4.847	145,38 %	9,12 %	0 %
Oct. 2008	32.084	1.106,06	30	12	1.711	0	4.992	145,69 %	11,14 %	0 %
Nov. 2008	32.084	1.131,15	34	13	1.718	0	5.230	146,85 %	11,09 %	0 %
Dic. 2008	32.084	1.154,77	42	14	1.742	0	5.453	146,82 %	10,67 %	0 %
Ene. 2009	32.084	1.153,57	46	15	1.753	0	5.572	146,65 %	9,88 %	0 %

Continuación de Tabla 6.22

Fecha	N° clientes	Media	Mediana	Q1	Q3	Mínimo	Máximo	Coef. Var.	Atípicos	Anómalos
Feb. 2009	32.084	1.133,55	47	15	1.747	0	5.443	146,13 %	9,90 %	0 %
Mar. 2009	32.084	1.099,58	44	14	1.614	0	5.117	145,09 %	10,86 %	0 %
Abr. 2009	32.084	1.140,53	45	13	1.742	0	5.352	146,15 %	10,52 %	0 %
May. 2009	32.084	1.159,52	78	12	1.767	0	5.132	142,29 %	8,94 %	0 %
Jun. 2009	32.084	1.134,14	52	11	1.725	0	4.947	143,02 %	9,23 %	0 %
Jul. 2009	32.084	1.124,28	50	11	1.711	0	4.870	143,09 %	9,15 %	0 %
Ago. 2009	32.084	1.126,11	49	11	1.749	0	4.874	142,33 %	9,17 %	0 %
Sept. 2009	32.084	1.135,56	46	11	1.839	0	4.866	141,87 %	3,20 %	0 %
Oct. 2009	32.084	1.122,13	45	12	1.719	0	5.010	142,88 %	10,98 %	0 %
Nov. 2009	32.084	1.153,97	46	13	1.743	0	5.231	144,47 %	11,08 %	0 %
Dic. 2009	32.084	1.135,41	66	13	1.715	0	5.285	145,95 %	10,82 %	0 %
Ene. 2010	32.084	1.136,41	102,5	15	1.755	0	5.362	144,26 %	9,74 %	0 %
Feb. 2010	32.084	1.192,14	112	16	1.790	0	5.683	143,61 %	9,66 %	0 %
Mar. 2010	32.084	1.023,95	58	14	1.737	0	4.267	137,74 %	0 %	0 %
Abr. 2010	32.084	1.023,81	34	14	1.560	0	4.620	144,28 %	10,80 %	0 %
May. 2010	32.084	1.111,10	40	13	1.601	0	5.229	147,11 %	12,98 %	0 %
Jun. 2010	32.084	1.097,26	36	12	1.707	0	4.960	145,52 %	8,86 %	0 %
Jul. 2010	32.084	1.085,71	30	11	1.693	0	4.777	145,00 %	9,28 %	0 %
Ago. 2010	32.084	1.102,41	33	11	1.692	0	4.931	145,62 %	11,58 %	0 %
Sept. 2010	32.084	1.117,49	37	12	1.692	0	4.986	145,06 %	11,87 %	0 %
Oct. 2010	32.084	1.084,02	46	13	1.643,5	0	4.898	144,57 %	11,08 %	0 %
Nov. 2010	32.084	1.120,68	67,5	13	1.576	0	5.258	146,49 %	13,18 %	0 %
Dic. 2010	32.084	1.145,67	84	14	1.729	0	5.436	146,55 %	10,47 %	0 %
Ene. 2011	32.084	1.135,68	102	15	1.730	0	5.268	144,75 %	10,51 %	0 %
Feb. 2011	32.084	1.159,05	108	16	1.746	0	5.464	143,21 %	9,83 %	0 %
Mar. 2011	32.084	1.110,75	108	14	1.733	0	4.966	141,52 %	8,24 %	0 %
Abr. 2011	32.084	1.151,11	101	14	1.719	0	5.261	143,30 %	10,72 %	0 %
May. 2011	32.084	1.133,64	91	8	1.696	0	5.170	144,46 %	11,05 %	0 %
Jun. 2011	32.084	1.128,06	76	8	1.738,25	0	5.127	144,82 %	9,00 %	0 %
Jul. 2011	32.084	1.133,32	85	7	1.718,5	0	5.114	144,25 %	11,55 %	0 %
Ago. 2011	32.084	1.091,23	68,5	7	1.770	0	4.806	143,44 %	6,07 %	0 %
Sept. 2011	32.084	1.122,74	92	7	1.782	0	4.917	143,20 %	6,44 %	0 %
Oct. 2011	32.084	1.096,24	88	9	1.684	0	4.964	142,81 %	8,52 %	0 %
Nov. 2011	32.084	1.123,59	86	10	1.690	0	5.109	143,19 %	10,82 %	0 %
Dic. 2011	32.084	1.158,79	168	10	1.705	0	5.411	143,84 %	10,42 %	0 %
Ene. 2012	32.084	1.229,86	177	12	1.874	0	5.866	141,32 %	9,13 %	0 %
Feb. 2012	32.084	1.171,47	188	11,14	1.858	0	5.413	140,82 %	7,60 %	0 %
Mar. 2012	32.084	1.124,97	175	11	1.708	0	4.965	140,69 %	8,54 %	0 %
Abr. 2012	32.084	1.212,31	187	11	1.874	0	5.508	141,52 %	10,49 %	0 %
May. 2012	32.084	1.171,43	187	10	1.824	0	5.046	140,10 %	6,38 %	0 %
Jun. 2012	32.084	1.172,81	191	9	1.817	0	4.994	140,29 %	6,36 %	0 %
Jul. 2012	32.084	1.155,26	126	8	1.818	0	4.927	140,64 %	6,27 %	0 %
Ago. 2012	32.084	1.134,78	135	8	1.778	0	4.847	140,80 %	6,19 %	0 %
Sept. 2012	32.084	1.164,36	181	9	1.846	0	4.974	140,36 %	6,32 %	0 %
Oct. 2012	32.084	1.144,69	188	10	1.795	0	5.087	140,76 %	8,34 %	0 %
Nov. 2012	32.084	1.168,02	193	10	1.817	0	5.195	141,14 %	8,70 %	0 %
Dic. 2012	32.084	1.215,71	195	11	1.860	0	5.362	140,13 %	8,71 %	0 %
Ene. 2013	32.084	1.207,43	193	12	1.863	0	5.414	139,26 %	8,19 %	0 %
Feb. 2013	32.084	1.450,72	918	25	2.305	0	5.470	111,45 %	0 %	0 %
Mar. 2013	32.084	1.216,35	198	12	1.895	0	5.295	137,08 %	8,26 %	0 %
Abr. 2013	32.084	1.233,39	188	12	1.866	0	5.383	138,27 %	8,60 %	0 %

Continuación de Tabla 6.22

Fecha	N° clientes	Media	Mediana	Q1	Q3	Mínimo	Máximo	Coef. Var.	Atípicos	Anómalos
May. 2013	32.084	1.221,06	198	10	1.976	0	5.176	137,45 %	3,17 %	0 %

El análisis de las medidas mensuales de consumo por cliente (Tabla 6.22) entre 2008 y 2013 mostró un crecimiento sostenido en los promedios de consumo, que pasaron de valores cercanos a 1.100 unidades a cifras superiores a 1.200 e incluso 1.400 en los últimos periodos, junto con una estacionalidad marcada por aumentos hacia finales de cada año e inicios del siguiente. El coeficiente de variación (CV), que se mantuvo de manera consistente por encima del 140 %, indica que la variabilidad del consumo mensual es muy alta en relación con su media. En términos prácticos, un CV superior al 100 % implica que la dispersión de los valores es mayor que el promedio mismo, lo que indica patrones de consumo diferentes entre los clientes. Adicionalmente, aunque los porcentajes de atípicos se mantuvieron entre el 9 % y el 12 %, y los casos anómalos fueron prácticamente nulos, estos valores no debían eliminarse del análisis. Los registros atípicos no representaban necesariamente errores, sino posibles señales de comportamientos relevantes, como fraudes, fugas o alteraciones en los patrones de uso, que podían ofrecer información crítica para la toma de decisiones. Por ello, en lugar de excluirlos, se calcularon variables o indicadores que permitieron capturar su presencia y frecuencia.

6.4. TRANSFORMACIÓN DE VARIABLES

6.4.1. Colapso de variables

Durante el análisis, la transformación de variables permitió observar patrones más claros y comparables entre los grupos, al reducir la dispersión y consolidar categorías con baja frecuencia. Esta etapa resultó fundamental para interpretar de manera más robusta el comportamiento de las variables categóricas del conjunto de datos. Gracias a esta simplificación, se logró una mejor comprensión de las diferencias entre grupos, evitando que las categorías poco representativas afectaran los resultados.

- **Año a Año simplificado (colapso):**

Tabla 6.23: Distribución de clientes según la variable *Año simplificado* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Año simplificado	#	%
Antes 1990	444	1,38 %
Años 90	2.969	9,25 %
Años 2000	21.812	67,98 %
Años 2010	6.859	21,38 %

(b)

Año simplificado	#	%
Antes 1990	2	1,72 %
Años 90	15	12,93 %
Años 2000	57	49,14 %
Años 2010	42	36,21 %

La distribución de la variable *Año simplificado* (Tabla 6.23) mostró que en la base de fraude predominaban los medidores correspondientes a los años 2000, con un 67,98 % del total, seguidos por los fabricados en los años 2010, con un 21,38 %. En la base de anomalías, esta tendencia se mantuvo, aunque con una mayor representación relativa de equipos recientes: los medidores de los años 2000 y 2010 concentraron el 49,14 % y el 36,21 % respectivamente. En ambos casos, las categorías anteriores a 1990 y de los años 90 presentaron una participación reducida.

- **Clase metrológica a Clase metrológica simplificada (colapso):**

Tabla 6.24: Distribución de clientes según la variable *Clase metrológica simplificada* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Clase metrológica simpl.	#	%
Clase 100	29.659	92,44 %
Otras clases	2.425	7,56 %

(b)

Clase metrológica simpl.	#	%
Clase 100	85	73,28 %
Otras clases	31	26,72 %

La distribución de la variable *Clase metrológica simplificada* (Tabla 6.24) mostró que, en la base de fraude, la gran mayoría de los medidores correspondían a la clase 100, con un 92,44 % del total. En la base de anomalías, aunque esta clase continuó siendo la predominante, su proporción disminuyó a 73,28 %, mientras que las demás clases aumentaron su participación al 26,72 %. Estos resultados reflejaron una mayor diversidad en la clase metrológica dentro de los casos asociados a anomalías.

- **Diámetro a Diámetro simplificado (colapso):**

Tabla 6.25: Distribución de clientes según la variable *Diámetro simplificado* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Diámetro simplificado	#	%
Diámetro bajo	31.230	97,34 %
Diámetro medio o alto	854	2,66 %

(b)

Diámetro simplificado	#	%
Diámetro bajo	116	100 %
Diámetro medio o alto	0	0 %

La distribución de la variable *Diámetro simplificado* (Tabla 6.25) mostró un predominio de los medidores con diámetro bajo en la base de fraude, con un 97,34 % del total. En la base de anomalías, esta categoría concentró la totalidad de los registros (100 %), sin presencia de medidores de diámetro medio o alto. Estos resultados reflejaron que las anomalías se presentaron exclusivamente en medidores de menor diámetro.

- **Localidad a Localidad simplificada (colapso):**

Tabla 6.26: Distribución de clientes según la variable *Localidad simplificada* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Localidad	#	%
Zona Bio Bío Interior	2.659	8,29 %
Zona Colchagua	814	2,54 %
Zona Concepción Arauco	11.706	36,49 %
Zona Litoral Maule	711	2,22 %
Zona Maule	5.205	16,22 %
Zona Ñuble	3.370	10,50 %
Zona O'Higgins	4.726	14,73 %
Otras zonas	2.893	9,02 %

(b)

Localidad	#	%
Zona Bio Bío Interior	13	11,21 %
Zona Colchagua	0	0 %
Zona Concepción Arauco	59	50,86 %
Zona Litoral Maule	0	0 %
Zona Maule	0	0 %
Zona Ñuble	22	18,97 %
Zona O'Higgins	0	0 %
Otras zonas	22	18,97 %

La distribución de la variable *Localidad simplificada* (Tabla 6.26) mostró que, en la base de fraude, la mayor proporción de registros correspondió a la zona Concepción Arauco (36,49 %), seguida por las zonas Maule (16,22 %) y O'Higgins (14,73 %). En la base de anomalías, la zona Concepción-Arauco también concentró la mayor proporción de casos (50,86 %), mientras que Ñuble y Otras zonas representaron el 19 % cada una. En contraste, Colchagua, Litoral Maule, Maule y O'Higgins no registraron casos de anomalías.

- **Marca a Marca simplificada (colapso):**

Tabla 6.27: Distribución de clientes según la variable *Marca simplificada* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Marca simplificada	#	%
CCM-Maipo-Actaris	20.431	63,68 %
Elster-Ex-Tavira	642	2 %
Lautaro-Sensus	6.628	20,66 %
Tavira-Iberconta	788	2,46 %
Marca no vigente	3.517	10,96 %
Otras marcas	78	2,46 %

(b)

Marca simplificada	#	%
CCM-Maipo-Actaris	74	63,79 %
Elster-Ex-Tavira	1	0,86 %
Lautaro-Sensus	32	27,59 %
Tavira-Iberconta	1	0,86 %
Marca no vigente	7	6,03 %
Otras marcas	1	0,86 %

La distribución de clientes según la variable Marca simplificada (Tabla 6.27) mostró una fuerte concentración en la marca CCM–Maipo–Actaris, que representó aproximadamente el 64 % tanto en la base de fraude como en la de anomalías. Este predominio indicó que la alta proporción de casos asociados a dicha marca respondía principalmente a su mayor presencia en el conjunto de medidores, más que a una mayor propensión al fraude o a lecturas atípicas.

- **Ruedas a Ruedas simplificada (colapso):**

Tabla 6.28: Distribución de clientes según la variable *Ruedas simplificada* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Ruedas simplificada	#	%
4-5 ruedas	30.919	96,37 %
6-7 ruedas	1.165	3,63 %

(b)

Ruedas simplificada	#	%
4-5 ruedas	106	91,38 %
6-7 ruedas	10	8,62 %

La distribución de la variable *Ruedas simplificada* (Tabla 6.28) mostró que, en la base de fraude, la mayoría de los medidores correspondían a aquellos con 4 o 5 ruedas, representando el 96,37 % del total. En la base de anomalías, esta categoría continuó siendo la predominante (91,38 %), aunque con una ligera mayor proporción de medidores con 6 o 7 ruedas (8,62 %). Estos resultados reflejaron una composición similar entre ambas bases, con predominio de medidores de menor número de ruedas.

- **Tranmisión a Transmisión simplificada (colapso):**

Tabla 6.29: Distribución de clientes según la variable *Transmisión simplificada* en las bases de datos de **(a)** fraude y **(b)** anomalías. Fuente: Elaboración propia.

(a)

Transmisión simplificada	#	%
T1	1.167	3,64 %
T2	30.458	94,93 %
T511	457	1,42 %
Otro tipo	2	0,01 %

(b)

Transmisión simplificada	#	%
T1	7	6,03 %
T2	97	83,62 %
T511	12	10,34 %
Otro tipo	0	0 %

La distribución de la variable *Transmisión simplificada* (Tabla 6.29) evidenció que en la base de fraude predominaban los medidores con transmisión T2, que representaron el 94,93 % del total, seguidos por los medidores con T1 con un 3,64 %. En la base de anomalías se mantuvo esta tendencia, con un 83,62 % de medidores T2, aunque se observó una participación mayor del tipo T511 (10,34 %), lo que indicó que las anomalías estuvieron relativamente más asociadas a este tipo de transmisión específica.

6.4.2. Validación de independencia y asociación

Los resultados del análisis de independencia y asociación mostraron diferencias notables entre las variables categóricas simplificadas. La prueba Chi-cuadrado permitió identificar cuáles de ellas presentaron dependencia significativa con la variable de referencia, mientras que el coeficiente V de Cramer permitió estimar la magnitud de dichas relaciones.

- **Fraude:**

Tabla 6.30: Resultados de la prueba Chi-cuadrado por variable categórica simplificada de la base de datos de fraude. Fuente: Elaboración propia.

Variable	p-value	Celdas < 5	Celdas $= 0$
Año simplificado	≈ 0	0	0
Clase metrológica simplificada	≈ 0	0	0
Diámetro simplificado	0,47	0	0
Localidad simplificada	≈ 0	0	0
Marca simplificada	≈ 0	0	0
Ruedas simplificada	0,0012	0	0
Transmisión simplificada	≈ 0	2	0

Los resultados de la prueba Chi-cuadrado (Tabla 6.30) mostraron que, con excepción de la variable *Diámetro simplificado*, todas las variables categóricas presentaron asociaciones estadísticamente significativas con la variable de referencia, al obtener valores de *p*-value cercanos a cero (p <0,05). Esto indicó dependencia estadística entre las categorías analizadas, lo cual sugiere que dichas variables podrían influir en la ocurrencia de fraude. En contraste, el *Diámetro simplificado* presentó un valor de p = 0,47, lo que evidenció independencia respecto a la variable dependiente.

Tabla 6.31: Resultados del coeficiente V de Cramer por variable categórica simplificada de la base de datos de fraude. Fuente: Elaboración propia.

Variable	V de Cramer
Año simplificado	0,13
Clase metrológica simplificada	0,17
Diámetro simplificado	0,0047
Marca simplificada	0,06
Localidad simplificada	0,05
Ruedas simplificada	0,02
Transmisión simplificada	0,08

El coeficiente V de Cramer (Tabla 6.31) confirmó que la magnitud de las asociaciones fue en general baja, siendo las más altas las correspondientes a las variables *Clase metrológica simplificada* (V = 0,17) y *Año simplificado* (V = 0,13). Estas asociaciones, aunque débiles, fueron más consistentes con los resultados obtenidos en la prueba de independencia, mientras que el *Diámetro simplificado* (V = 0,0047) mostró una relación prácticamente nula. No obstante, esta variable se mantuvo dentro de los modelos con el propósito de verificar empíricamente su escasa capacidad explicativa y confirmar, a partir de los resultados del modelado, que su aporte predictivo era efectivamente limitado. En conjunto, los resultados indicaron que, aunque varias variables categóricas estuvieron asociadas con la presencia de fraude, la fuerza de dichas relaciones fue reducida.

- **Anomalías:**

Tabla 6.32: Resultados de la prueba Chi-cuadrado por variable categórica simplificada de la base de datos de anomalías. Fuente: Elaboración propia.

Variable	p-value	Celdas < 5	Celdas $= 0$
Año simplificado	0,5	3	0
Clase metrológica simplificada	0,06	0	0
Diámetro simplificado	No aplica	No aplica	No aplica
Marca simplificada	0,83	7	0
Localidad simplificada	0,72	1	0
Ruedas simplificada	1	1	0
Transmisión simplificada	0,66	2	0

Los resultados de la prueba Chi-cuadrado para las variables categóricas simplificadas (Tabla 6.32) mostraron que no se encontraron asociaciones estadísticamente significativas entre estas y la variable objetivo, dado que los valores de p fueron superiores a 0,05. Esto indicó que las categorías analizadas no presentaron diferencias relevantes en su distribución con respecto a la presencia o ausencia de anomalías.

Por otra parte, la variable *Diámetro simplificado* no presentó variabilidad, es decir, todos los registros correspondieron a una única categoría. Debido a ello, no fue posible aplicar la prueba Chi-cuadrado, ya que la falta de variabilidad impide estimar la asociación estadística. En consecuencia, esta variable fue descartada del análisis.

Tabla 6.33: Resultados del coeficiente V de Cramer por variable categórica simplificada de la base de datos de anomalías. Fuente: Elaboración propia.

Variable	V de Cramer
Año simplificado	0,143
Clase metrológica simplificada	0,194
Diámetro simplificado	No aplica
Marca simplificada	0,134
Localidad simplificada	0,107
Ruedas simplificada	0,017
Transmisión simplificada	0,084

Los resultados del coeficiente V de Cramer (Tabla 6.33) evidenciaron que las asociaciones entre las variables categóricas simplificadas y la variable objetivo fueron débiles en la mayoría de los casos (V <0,20). En particular, las variables *Año simplificado* y *Clase metrológica simplificada* mostraron los valores más altos (0,143 y 0,194, respectivamente), indicando una ligera dependencia con la ocurrencia de anomalías. No obstante, estos niveles de asociación no fueron lo suficientemente elevados como para considerarse relaciones fuertes. En contraste, las variables *Marca simplificada*, *Localidad simplificada*, *Ruedas simplificada* y *Transmisión simplificada* presentaron valores de V inferiores a 0,15, lo que indicó que hay independencia casi total respecto a la variable de anomalías.

6.4.3. Análisis de correlación

El coeficiente de correlación de Spearman permitió identificar la fuerza y dirección de las asociaciones entre las variables numéricas del conjunto de datos. Este análisis evidenció relaciones monotónicas altas entre varios indicadores de consumo, reflejando que el aumento o disminución de una variable tiende a estar acompañado por variaciones consistentes en otras.

- **Fraude:**

Tabla 6.34: Resultados de la prueba de correlación de Spearman ($|\rho|$ >0,7) entre variables numéricas de la base de datos de fraude. Fuente: Elaboración propia.

Variable 1	Variable 2	Spearman	$\|\rho\|$
Rango consumo	*Máx. consumo*	0,99	0,99
Z mín.	*CV consumo*	0,96	0,96
Rango consumo	*SD consumo*	0,95	0,95
Mediana consumo	*Media consumo*	0,94	0,94
Máx. consumo	*SD consumo*	0,94	0,94
Z máx.	*Asimetría consumo*	0,92	0,92
LZC 4 bins	*LZC 9 bins*	0,91	0,91
Máx. consumo	*Media consumo*	0,90	0,90
Z máx.	*Curtosis consumo*	0,89	0,89
TSFL	*Mediana consumo*	0,87	0,87
LZC 2 bins	*Curtosis consumo*	-0,87	0,87
Rango consumo	*Media consumo*	0,87	0,87
TSFL	*Media consumo*	0,87	0,87
SD consumo	*Media consumo*	0,87	0,87
Entropía SAX	*Mediana consumo*	0,84	0,84
Z mín.	*Asimetría consumo*	0,84	0,84
LZC 4 bins	*LZC 2 bins*	0,84	0,84
TSFL	*Entropía SAX*	0,84	0,84
Entropía SAX	*Media consumo*	0,81	0,81
LZC 2 bins	*Asimetría consumo*	-0,81	0,81
LZC 2 bins	*Z máx.*	-0,81	0,81
Curtosis consumo	*Asimetría consumo*	0,81	0,81
Máx. consumo	*Mediana consumo*	0,80	0,80
LZC 9 bins	*Z mín.*	-0,80	0,80
LZC 4 bins	*Asimetría consumo*	-0,79	0,79
TSFL	*Máx. consumo*	0,79	0,79
LZC 9 bins	*CV consumo*	-0,79	0,79
LZC 9 bins	*LZC 2 bins*	0,77	0,77
Rango consumo	*Mediana consumo*	0,77	0,77
TSFL	*Rango consumo*	0,77	0,77
LZC 9 bins	*Asimetría consumo*	-0,76	0,76
LZC 4 bins	*Curtosis consumo*	-0,76	0,76
TSFL	*SD consumo*	0,76	0,76

Continuación de la Tabla 6.34

Variable 1	**Variable 2**	**Spearman**	$\|\rho\|$
Mediana consumo	*SD consumo*	0,76	0,76
LZC 4 bins	*Z mín.*	-0,76	0,76
Asimetría consumo	*CV consumo*	0,76	0,76
Secuencias cero	*Consumo ceros*	0,75	0,75
Entropía SAX	*Máx. consumo*	0,74	0,74
Mín. consumo	*Mediana consumo*	0,74	0,74
LZC 4 bins	*CV consumo*	-0,73	0,73
Entropía SAX	*SD consumo*	0,73	0,73
N° atípicos extremos	*N° atípicos moderados*	0,73	0,73
Entropía SAX	*Rango consumo*	0,72	0,72
LZC 99 bins	*LZC 9 bins*	0,72	0,72
LZC 4 bins	*Z máx.*	-0,72	0,72
LZC 99 bins	*Entropía SAX*	0,72	0,72
TSFL	*LZC 99 bins*	0,71	0,71
LZC 9 bins	*Curtosis consumo*	-0,71	0,71
LZC 99 bins	*Mediana consumo*	0,71	0,71
N° atípicos extremos	*Curtosis consumo*	0,70	0,70
TSFL	*Mín. consumo*	0,70	0,70

Los resultados de la prueba de correlación de Spearman evidenciaron una fuerte relación entre varias de las variables numéricas del conjunto de datos. En particular, las correlaciones mostradas en la Tabla 6.34 revelaron que gran parte de las variables de consumo (como el rango, la media, la desviación estándar, el máximo y la mediana) se comportaron de manera similar, reflejando que midieron dimensiones estrechamente relacionadas del comportamiento del consumo de agua. Asimismo, las variables derivadas de medidas de complejidad y entropía, como las asociadas a *LZC* y *Entropía SAX*, también presentaron correlaciones elevadas entre sí, lo que indicó redundancia en la información capturada por esos indicadores.

Durante la revisión de correlaciones entre variables numéricas se identificaron relaciones de alta dependencia que podían generar redundancia en el conjunto de características. El coeficiente de variación del consumo (*CV consumo*) presentó correlaciones altas con *Rango consumo*, *Media consumo*, *SD consumo*, *Máx. consumo* y *Mín. consumo*. De manera similar, *Rango consumo* mostró correlaciones elevadas con *Máx. consumo*, *Media consumo*, *SD consumo* y *Mín. consumo*, mientras que *Máx. consumo*, *SD consumo*, *Media consumo* y *Mediana consumo* presentaron entre sí asociaciones, evidenciando una clara multicolinealidad entre métricas de tendencia central y dispersión.

Por otra parte, las variables *Z máx.* y *Asimetría consumo* se correlacionaron fuertemente, reflejando información similar sobre la distribución del consumo. La variable *Consumo ceros* también mostró correlaciones relevantes con *Rango consumo*, *Media consumo* y *Máx. consumo*, por lo que se consideró redundante. De igual manera, la variable *Cambios de símbolo* se eliminó por presentar correlaciones significativas con *Entropía SAX*, *Mín. consumo* y *Mediana consumo*, lo que indicó que aportaba información equivalente sobre la variabilidad y la estructura interna de las series de consumo.

En cuanto a las medidas de complejidad basadas en la complejidad de Lempel-Ziv (*LZC*), las versiones

calculadas con distintos niveles de discretización (2, 4, 9 y 99 bins) mostraron correlaciones. Dado que todas representaron transformaciones de la misma métrica, se decidió conservar únicamente una de ellas para reducir la redundancia y simplificar el modelo. En los anexos (Sección 11.1) se amplía el detalle de las correlaciones.

- **Anomalías:**

Para este análisis de las variables de anomalías, es importante recordar que no se incluyeron las variables *Consumo ceros* ni *Secuencias cero*, dado que los registros analizados no presentan consumos iguales a cero, lo que impide su construcción.

Tabla 6.35: Resultados de la prueba de correlación de Spearman ($|\rho|$ >0,7) entre variables numéricas de la base de datos de anomalías. Fuente: Elaboración propia.

Variable 1	**Variable 2**	**Spearman**	$\|\rho\|$
Z mín.	*CV consumo*	0,964	0,964
Rango consumo	*Máx. consumo*	0,956	0,956
Mediana consumo	*Media consumo*	0,901	0,901
Asimetría consumo	*Media consumo*	-0,892	0,892
Z máx.	*Asimetría consumo*	0,853	0,853
Z mín.	*Mediana consumo*	-0,850	0,850
Asimetría consumo	*Mediana consumo*	-0,846	0,846
LZC 4 bins	*LZC 9 bins*	0,837	0,837
Z máx.	*Media consumo*	-0,816	0,816
Z máx.	*Curtosis consumo*	0,813	0,813
Rango consumo	*SD consumo*	0,809	0,809
Mediana consumo	*CV consumo*	-0,786	0,786
Máx. consumo	*SD consumo*	0,768	0,768
LZC 2 bins	*Curtosis consumo*	-0,746	0,746
Z mín.	*Asimetría consumo*	0,741	0,741
Z mín.	*Media consumo*	-0,741	0,741
Z outliers	*N° atípicos moderados*	0,735	0,735
LZC 4 bins	*LZC 2 bins*	0,709	0,709
N° atípicos extremos	*N° atípicos moderados*	0,704	0,704

La Tabla 6.35 evidenció una alta redundancia entre varias de las variables numéricas analizadas, lo que indicó la necesidad de depurar el conjunto de predictores para evitar problemas de multicolinealidad y sobreajuste en los modelos posteriores. Las correlaciones más elevadas se observaron entre *Z mín.*, *CV consumo* y *Mediana consumo*, así como entre *Rango consumo*, *Máx. consumo* y *SD consumo*, lo que reflejó una dependencia estructural entre las medidas de tendencia central y dispersión. De igual manera, las medidas de forma, como *Asimetría consumo* y *Curtosis consumo*, mostraron correlaciones significativas con variables de magnitud extrema (*Z máx.* y *Z mín.*), mientras que los indicadores de complejidad temporal (*LZC*) presentaron correlaciones altas entre sus configuraciones de bins, lo que evidenció que aportaban información equivalente sobre la variabilidad de las series. En función de estos resultados, se eliminaron las variables altamente correlacionadas manteniendo aquellas que ofrecían una mejor representación del comportamiento general del consumo y mayor estabilidad estadística. En primer lugar, se eliminaron *CV consumo* y *Z mín.*, debido a su fuerte relación mutua ($\rho = 0{,}965$) y con

Mediana consumo; la mediana fue conservada por reflejar mejor la tendencia central sin depender de valores extremos. De forma similar, *Rango consumo* y *Máx. consumo* se descartaron por su alta correlación con *SD consumo*, medida que resultó más estándar y robusta para representar la variabilidad.

Asimismo, *Media consumo* se eliminó por su correlación con *Mediana consumo*, *Asimetría consumo*, dado que la mediana ofreció una representación más confiable en presencia de atípicos. La variable *Z máx.* también se excluyó por su redundancia con las medidas de forma (*Asimetría consumo* y *Curtosis consumo*), mientras que *Curtosis consumo* se descartó por su correlación con *LZC 2 bins*, dado que la asimetría ya describía la forma de la distribución de manera suficiente.

En cuanto a los indicadores de complejidad, *LZC 2 bins* y *LZC 4 bins* se eliminaron por su redundancia con *LZC 9 bins*, que se mantuvo por ofrecer una mayor granularidad en la descripción de la señal. Finalmente, las variables relacionadas con valores atípicos (*Z atípicos* y *N° atípicos extremos*) se eliminaron por su alta correlación con *N° atípicos moderados*, conservándose esta última por ser más interpretativa y representativa del comportamiento anómalo sin sesgar la distribución general de los datos. En los anexos (Sección 11.1) se amplía el detalle de las correlaciones.

6.5. VARIABLES FINALES PARA LOS MODELOS

- **Fraude:**

En un escenario ideal, habría sido deseable que los tres enfoques (supervisado, no supervisado y de regresión logística) utilizaran el mismo conjunto de variables, con el fin de realizar una comparación más directa del desempeño entre los modelos. Sin embargo, esto no fue posible en todos los casos, dado que los modelos estadísticos, particularmente los de tipo paramétrico, requieren la validación de supuestos específicos (como la normalidad, la independencia o la ausencia de multicolinealidad), lo que condicionó la selección final de variables. En este estudio, únicamente fue viable mantener el mismo conjunto de características entre el modelo supervisado y la regresión logística, ya que ambos compartieron criterios de preparación y estandarización de datos. En contraste, el modelo no supervisado no permitió una correspondencia exacta, puesto que se basó en la metodología propuesta por Ghamkhar et al. [47], la cual empleó un conjunto de variables diferente, ajustado a los principios del aprendizaje no supervisado.

Modelación supervisada vs. Regresión logística

Tabla 6.36: Variables utilizadas en la modelación supervisada y de regresión logística para la detección de fraude. Fuente: Elaboración propia.

Variable	Modelación supervisada	Regresión logística
Latitud	Original	Original + Modelo GAM
Longitud	Original	Original + Modelo GAM
Curtosis consumo	Original	Transf. completa* + estandarizado
N° atípicos moderados	Original	Transf. completa* + estandarizado
N° atípicos extremos	Original	Transf. completa* + estandarizado
Z mín.	Original	Transf. completa* + estandarizado
Z atípicos	Original	Transf. completa* + estandarizado

Continuación de la Tabla 6.36

Variable	Modelación supervisada	Regresión logística
Delta brusco	Original	Transf. completa* + estandarizado
Secuencia ceros	Original	Transf. completa* + estandarizado
Entropía SAX	Original	Transf. completa* + estandarizado
LZC 9 bins	Original	Transf. completa* + estandarizado
TSFL	Original	Transf. completa* + estandarizado
Año simplificado	Original	Original
Clase metrológica simplificada	Original	Original
Diámetro simplificado	Original	Original
Localidad simplificada	Original	Original
Marca simplificada	Original	Original
Ruedas simplificada	Original	Original
Transmisión simplificada	Original	Original
Fraude (variable objetivo)	Original	Original

*Transformación completa hace referencia a la aplicación de *winsor* + *shift* + *log1p*.

De acuerdo con la Tabla 6.36, las variables empleadas en la modelación supervisada no requirieron transformaciones adicionales, dado que estos algoritmos utilizados son inherentemente robustos frente a escalas y distribuciones no normales, al basarse en divisiones jerárquicas y medidas de ganancia de información [60], [62]. En consecuencia, el uso de las variables en su forma original no afectó la capacidad predictiva ni la estabilidad del modelo. Por el contrario, la regresión logística sí exigió la estandarización y transformación de las variables numéricas (mediante *winsorization*, *shift* y *log1p*) para cumplir con los supuestos de linealidad, homocedasticidad y ausencia de multicolinealidad. Esta diferenciación metodológica explica por qué, aunque las variables analizadas son conceptualmente equivalentes, su tratamiento estadístico varió según el tipo de modelo implementado.

Modelación no supervisada

Para este modelo se emplearon exclusivamente las versiones normalizadas de la complejidad de Lempel–Ziv (*LZC*) con discretizaciones de 2, 4, 9 y 99 *bins*, junto con la métrica *TSLF*, siguiendo la metodología propuesta por Ghamkhar et al. [47].

- **Anomalías:**

Dado que los resultados de las pruebas de independencia y asociación para las variables categóricas de anomalías no evidenciaron relaciones estadísticamente significativas que permitieran definir conjuntos de variables a utilizar en combinación, la selección de variables se realizó en función del tipo de modelo aplicado. En este caso, se priorizó la coherencia metodológica entre la modelación supervisada y la regresión logística, garantizando la validez estadística y el cumplimiento de los supuestos propios de cada técnica, más que la homogeneidad en el conjunto de predictores.

Modelación supervisada

Réplica de las variables utilizadas en la modelación de fraude para hacer la comparativa con la modelación de anomalías.

Tabla 6.37: Variables utilizadas en la modelación supervisada para la detección de anomalías. Fuente: Elaboración propia.

Variable	Modelación supervisada
Latitud	Original
Longitud	Original
Curtosis consumo	Original
N° atípicos moderados	Original
N° atípicos extremos	Original
Z mín.	Original
Z atípicos	Original
Delta brusco	Original
Secuencia ceros	Original
Entropía SAX	Original
LZC 9 bins	Original
TSFL	Original
Año simplificado	Original
Clase metrológica simplificada	Original
Diámetro simplificado	Original
Localidad simplificada	Original
Marca simplificada	Original
Ruedas simplificada	Original
Transmisión simplificada	Original
Anomalías (variable objetivo)	Original

Modelación no supervisada

No aplica.

Regresión logística

Tabla 6.38: Variables utilizadas en la regresión logística para la detección de anomalías (transformaciones + estandarizado). Fuente: Elaboración propia.

Variable	Regresión logística
Latitud	Original + Modelo GAM
Longitud	Original + Modelo GAM
Asimetría consumo	winsor + estandarizado
Cambios de símbolo	Transformación completa* + estandarizado
Delta brusco	winsor + estandarizado
Entropía SAX	winsor + estandarizado
LZC 9 bins	Transformación completa* + estandarizado
LZC 99 bins	Transformación completa* + estandarizado
Mediana consumo	winsor + estandarizado
Mín. consumo	winsor + estandarizado
N° atípicos moderados	winsor + estandarizado
SD consumo	winsor + estandarizado

Continuación de la Tabla 6.38

Variable	**Regresión logística**
TSFL	Transformación completa* + estandarizado
Año simplificado	Original
Clase metrológica simplificada	Original
Localidad simplificada	Original
Marca simplificada	Original
Ruedas simplificada	Original
Transmisión simplificada	Original
Anomalías (variable objetivo)	Original

*Transformación completa hace referencia a la aplicación de winsor + shift + log1p.

6.6. ELECCIÓN DE LOS MODELOS SUPERVISADOS

De acuerdo con las características de los datos y la naturaleza del fraude y anomalías como *eventos raros*, la estrategia inicial consistió en priorizar modelos que maximizan el área bajo la curva de precisión vs. recall (AUC-PR), tales como *Random Forest* y *XGBoost*. La elección de *Random Forest* y *XGBoost* como modelos prioritarios para este problema se fundamenta en su capacidad de manejar conjuntos de datos complejos y altamente desbalanceados, características típicas en escenarios de detección de fraude o anomalías [62], [94]. Ambos modelos pertenecen a la familia de métodos de ensamble basados en árboles de decisión, lo que les permite capturar relaciones no lineales y explorar interacciones entre variables de manera más efectiva que los modelos lineales o más simples. Además, tienen la capacidad de ajustar la importancia relativa de las clases minoritarias y permiten calibrar umbrales de decisión para adaptarse a métricas como el AUC-PR, que resulta fundamental en contextos donde la clase de interés es poco frecuente [62], [94]. Por estas razones, se consideran opciones sólidas para maximizar la detección de fraudes reales y minimizando el error tipo II.

6.7. MÉTRICAS DE SELECCIÓN PARA LOS MODELOS

- **Fraude:**

Tabla 6.39: Métricas evaluadas por modelo y enfoque para la detección de fraudes. Fuente: Elaboración propia.

Caso	Métricas
Modelación supervisada: Random Forest vs. XGBoost	
Sin enfoque de negocio	*F1-Score*: métrica principal de optimización y selección, adecuada para conjuntos desbalanceados al equilibrar precisión y sensibilidad. *Brier Score*: métrica secundaria para evaluar la calibración probabilística; valores bajos indican estimaciones más confiables. *Precision*: mide la proporción de predicciones positivas que son correctas, útil para minimizar falsos positivos. *Recall*: refleja la capacidad del modelo para identificar correctamente los casos de fraude, reduciendo falsos negativos.
Con enfoque de negocio	*AUC-PR*: métrica principal de selección estadística, más representativa que el AUC-ROC en escenarios con alta desproporción de clases. *Brier Score*: métrica secundaria de calibración que permite validar la fiabilidad de las probabilidades de fraude estimadas. *Precisión y Recall*: empleadas conjuntamente para verificar que el modelo cumpla los umbrales operativos definidos por el negocio.
Modelación no supervisada: DBSCAN	
Sin enfoque de negocio con optimización F1-Score	*F1-Score*: métrica principal para calibración semi-supervisada, útil para ajustar parámetros del algoritmo en presencia de etiquetas parciales. *Precisión y Recall*: utilizadas para medir el desempeño final y la capacidad de detección de anomalías dentro del conjunto de prueba.
Sin enfoque de negocio con optimización AUC-PR	*AUC-PR*: métrica principal de optimización basada en el ranking de *scores*, apropiada para priorizar las detecciones más confiables. *F1-Score, Precisión y Recall*: métricas finales que permiten una evaluación integral del equilibrio entre detección efectiva y errores de clasificación.

- **Anomalías:**

Para la modelación supervisada orientada a la detección de anomalías, se empleó el *F1-Score promedio* como criterio principal de selección, dado que equilibra la precisión y la sensibilidad del modelo en escenarios con clases desbalanceadas. Adicionalmente, se utilizó el *Brier Score promedio* como métrica complementaria para el desempate y la evaluación de la fiabilidad probabilística, permitiendo validar la coherencia entre las probabilidades estimadas y los resultados observados.

7 RESULTADOS: MODELOS FRAUDE

7.1. MODELACIÓN SUPERVISADA FRAUDE

7.1.1. Modelación Random Forest vs. XGBoost sin enfoque de negocio (fraude)

- **Random Forest:**

Random Forest: Búsqueda de hiperparámetros sin enfoque de negocio (fraude)
De acuerdo con la Tabla 7.1, los resultados de los mejores hiperparámetros mostraron que los modelos con un mayor número de estimadores (`n estimators`) y profundidad del árbol (`Max Depth`) tendieron a alcanzar los valores más altos del *F1 promedio*. En particular, el modelo con 250 árboles y profundidad máxima de 20 obtuvo el mejor desempeño (*F1 promedio* = 0,053), evidenciando que un incremento en la complejidad del modelo contribuyó a una mejor capacidad de clasificación en un contexto de fuerte desbalance de clases. Sin embargo, también se observa una variabilidad moderada en los valores de *F1 desviación*, lo que indica cierta inestabilidad asociada a la sensibilidad del modelo frente a los datos de entrenamiento.

En contraste, los modelos con profundidades menores (10 o Ninguno) y menos estimadores (150) mostraron desempeños más bajos, con valores de *F1 promedio* cercanos a 0,022 – 0,031, lo que indica una pérdida de capacidad predictiva al reducir la complejidad del bosque. Estos resultados confirman que, en problemas de detección de fraude con alta desproporción de clases, la combinación de un número elevado de árboles y una profundidad controlada favorece el equilibrio entre sesgo y varianza, maximizando la detección de la clase minoritaria sin comprometer la estabilidad general del modelo.

Tabla 7.1: Resultados del Grid Search *Random Forest* (fraude) sin enfoque de negocio.
Fuente: Elaboración propia.

`n estimators`	`Max Depth`	`Min Split`	`Min Leaf`	`Max Features`	**F1 Prom.**	**F1 Desv. Est.**	**Ranking**
250	20	5	2	sqrt	0,053	0,032	1
250	Ninguno	5	2	sqrt	0,049	0,008	2
150	Ninguno	5	2	sqrt	0,049	0,016	3
150	20	5	2	sqrt	0,049	0,025	4
150	10	5	2	sqrt	0,040	0,026	5

Continuación de la Tabla 7.1

n est.	Max Depth	Min Split	Min Leaf	Max Features	**F1 Prom.**	**F1 Desv. Est.**	**Ranking**
150	Ninguno	10	2	sqrt	0,040	0,026	6
250	10	10	2	sqrt	0,040	0,025	7
150	20	10	2	sqrt	0,040	0,029	8
150	10	10	2	sqrt	0,040	0,037	9
250	10	5	2	sqrt	0,036	0,022	10
250	20	5	2	sqrt	0,036	0,022	11
150	Ninguno	5	4	sqrt	0,031	0,022	12
250	Ninguno	5	4	sqrt	0,031	0,022	13
150	10	5	4	sqrt	0,027	0,021	14
250	20	5	4	sqrt	0,027	0,016	15
150	10	10	4	sqrt	0,027	0,021	16
250	10	5	4	sqrt	0,027	0,021	17
250	10	5	2	sqrt	0,022	0,014	18
150	Ninguno	10	4	sqrt	0,022	0,014	19
250	Ninguno	10	4	sqrt	0,022	0,014	20
250	10	5	4	sqrt	0,022	0,014	21
150	20	10	4	sqrt	0,022	0,020	22
250	10	10	4	sqrt	0,022	0,020	23
150	10	10	4	sqrt	0,022	0,024	24

Random Forest: Métricas en cada fold sin enfoque de negocio (fraude)

En la Tabla 7.2 se presentan los resultados obtenidos para el modelo *Random Forest* en los cinco folds de validación. Los valores del *Brier-Score* se mantuvieron bajos y estables (entre 0,013 y 0,014), lo que refleja un desempeño consistente en la estimación de probabilidades. En cuanto al *F1-Score*, se observa una tendencia creciente a lo largo de los folds, pasando de 0,275 en el primero a 0,376 en el quinto, lo que indica una mejora progresiva en la capacidad del modelo para equilibrar la precisión y la exhaustividad en la detección de fraudes. De forma paralela, la precisión aumentó de 0,293 a 0,420, mientras que el *Recall* se incrementó de 0,258 a 0,341, evidenciando un fortalecimiento paulatino del modelo en la identificación correcta de casos positivos sin sacrificar estabilidad.

Tabla 7.2: *Random Forest* (fraude) sin enfoque de negocio - Métricas detalladas por fold
Fuente: Elaboración propia.

Fold	Calibrador	Brier-Score	Mejor umbral	F1-Score	Precision	Recall
1	Isotónico	0,014	0,166	0,275	0,293	0,258
2	Isotónico	0,014	0,187	0,284	0,285	0,282
3	Isotónico	0,014	0,200	0,319	0,321	0,317
4	Isotónico	0,013	0,166	0,323	0,342	0,305
5	Isotónico	0,013	0,250	0,376	0,420	0,341

Random Forest: Métricas promedio y desviación sin enfoque de negocio (fraude)

La Tabla 8.6 mostró que el modelo *Random Forest* mantuvo un desempeño estable en las métricas de validación. El *F1-Score* promedio fue de 0,315, con una desviación de 0,04, lo que reflejó un equilibrio razonable entre precisión y exhaustividad. El mejor umbral se ubicó en 0,194, confirmando

una calibración adecuada del modelo, mientras que el *Brier-Score* promedio de 0,014 indicó una alta coherencia entre las probabilidades estimadas y los valores observados.

Tabla 7.3: *Random Forest* (fraude) sin enfoque de negocio - Promedio y desviación estándar de las métricas. Fuente: Elaboración propia.

Métrica	Promedio	Desviación
Brier-Score	0,014	0,000
Mejor umbral	0,194	0,034
F1-Score	0,315	0,04
Precision	0,332	0,054
Recall	0,301	0,031

- **XGBoost:**

XGBoost: Búsqueda de hiperparámetros sin enfoque de negocio (fraude)

Los resultados del *Grid Search* para el modelo *XGBoost* de la Tabla 7.4 evidenciaron que los mejores desempeños se obtuvieron con un número elevado de estimadores y profundidades intermedias del árbol. El modelo con 250 árboles, profundidad máxima de 5 y una tasa de aprendizaje de 0,10 alcanzó el mayor valor promedio de *F1-Score* (0,154), seguido de configuraciones similares con profundidades de 7 y tasas de aprendizaje de 0,10 o 0,05. En términos generales, los valores de *F1-Score* se mantuvieron por debajo de 0,15, lo que reflejó un rendimiento inferior al observado en el *Random Forest*. Este comportamiento se asoció con la mayor sensibilidad de *XGBoost* a los desequilibrios de clase y a la regularización aplicada, que redujo el sobreajuste pero también limitó la capacidad del modelo para detectar correctamente la clase minoritaria. Aun así, la baja desviación estándar en los resultados indicó una estabilidad razonable entre las configuraciones evaluadas.

Tabla 7.4: Resultados del Grid Search *XGBoost* (fraude) sin enfoque de negocio. Fuente: Elaboración propia.

n est.	Max Depth	Learn. Rate	Subsample	Colsample ByTree	F1 Prom.	F1 Desv.	Ranking
250	5	0,10	0,80	0,80	0,154	0,046	1
150	7	0,10	0,80	0,80	0,146	0,025	2
250	7	0,05	0,80	0,80	0,126	0,038	3
250	7	0,10	0,80	0,80	0,126	0,032	4
150	5	0,10	0,80	0,80	0,126	0,042	5
250	5	0,05	0,80	0,80	0,118	0,025	6
250	3	0,10	0,80	0,80	0,118	0,024	7
150	7	0,05	0,80	0,80	0,118	0,032	8
150	5	0,05	0,80	0,80	0,099	0,027	9
150	3	0,10	0,80	0,80	0,095	0,028	10
250	3	0,05	0,80	0,80	0,082	0,024	11
150	3	0,05	0,80	0,80	0,040	0,016	12

XGBoost: Métricas en cada fold sin enfoque de negocio (fraude)

Teniendo en cuenta la Tabla 8.5, los resultados obtenidos para el modelo *XGBoost* calibrado con regresión isotónica mostraron un desempeño moderado y consistente entre los cinco folds de validación. El *F1-Score* varió entre 0,274 y 0,353, con una ligera superioridad en el primer fold, mientras que el

Brier-Score se mantuvo estable en torno a 0,014, lo que indicó una buena calibración probabilística del modelo. En cuanto al umbral, los valores oscilaron entre 0,142 y 0,285, reflejando ajustes adaptativos según la distribución de las clases en cada partición. Aunque las métricas de *Precision* y *Recall* presentaron leves fluctuaciones, el comportamiento general del modelo evidenció estabilidad y una capacidad razonable para discriminar entre casos de fraude y no fraude, aun cuando su desempeño global fue inferior al alcanzado por el *Random Forest*.

Tabla 7.5: *XGBoost* (fraude) sin enfoque de negocio - Métricas detalladas por fold. Fuente: Elaboración propia.

Fold	Calibrador	Brier-Score	Mejor umbral	F1-Score	Precision	Recall
1	Isotónico	0,014	0,285	0,353	0,419	0,305
2	Isotónico	0,014	0,222	0,281	0,328	0,247
3	Isotónico	0,014	0,200	0,274	0,327	0,235
4	Isotónico	0,014	0,175	0,327	0,337	0,317
5	Isotónico	0,014	0,142	0,274	0,266	0,282

XGBoost: Métricas promedio y desviación sin enfoque de negocio (fraude)

La Tabla 7.6 presentó los valores promedio y la desviación estándar de las métricas obtenidas con el modelo *XGBoost*. El *F1-Score* alcanzó un valor promedio de 0,302 con una desviación de 0,036, lo que evidenció un rendimiento moderado y estable entre los folds. El mejor umbral se ubicó en 0,205, mientras que el *Brier-Score* promedio de 0,014 indicó una buena calibración de las probabilidades estimadas. Las métricas de *Precision* (0,335) y *Recall* (0,277) mostraron un equilibrio razonable.

Tabla 7.6: Promedio y desviación estándar de las métricas del modelo *XGBoost* (fraude) sin enfoque de negocio. Fuente: Elaboración propia.

Métrica	Promedio	Desviación
Brier-Score	0,014	0,000
Mejor umbral	0,205	0,053
F1-Score	0,302	0,036
Precision	0,335	0,054
Recall	0,277	0,035

- **Comparación y discusión Random Forest vs. XGBoost sin enfoque de negocio (fraude):**

Tabla 7.7: Comparación de métricas entre *Random Forest* (RF) y *XGBoost* (XG) sin enfoque de negocio (fraude). Fuente: Elaboración propia.

Métrica	RF Prom.	RF Desv. Est.	XGB Prom.	XGB Desv. Est.
Brier-Score	0,014	0,000	0,014	0,000
Mejor umbral	0,194	0,034	0,205	0,053
F1-Score	0,315	0,040	0,302	0,036
Precision	0,332	0,054	0,335	0,054
Recall	0,301	0,031	0,277	0,035

La Tabla 7.7 presentó los resultados comparativos entre los modelos *Random Forest* (RF) y *XGBoost* (XGB), evidenciando un desempeño superior del primero en la mayoría de las métricas analizadas. Aunque ambos modelos mostraron un *Brier-Score* promedio de 0,014, indicador de una buena calibración probabilística, las diferencias en el resto de las métricas reflejaron un mejor equilibrio general a favor del *Random Forest*.

En cuanto al *F1-Score*, considerado el criterio principal de evaluación, su relevancia radicó en que, frente a datasets altamente desbalanceados, esta métrica permitió valorar de manera más equilibrada la capacidad del modelo para detectar fraudes sin incurrir en un exceso de falsos positivos. Bajo este criterio, el *Random Forest* superó a *XGBoost* en aproximadamente un 4,3 % relativo, lo que evidenció un mejor balance entre *Precision* y *recall* y una mayor capacidad de generalización en la clasificación de casos minoritarios.

Respecto al *Brier-Score*, utilizado como criterio secundario, su importancia se centró en la calidad de las probabilidades predichas, un aspecto crucial en contextos donde las decisiones se sustentan en la estimación del riesgo asociado a cada cliente. Si bien la diferencia entre ambos modelos fue mínima, el *Random Forest* presentó valores ligeramente más consistentes y con menor desviación, lo que indicó una calibración más estable y, por ende, probabilidades más fiables.

Finalmente, tanto la *Precision* (0,332 vs. 0,335) como el *Recall* (0,301 vs. 0,277) reforzaron la ventaja del *Random Forest*, que logró mantener una sensibilidad superior sin sacrificar exactitud. En conjunto, los resultados confirmaron que, bajo la calibración isotónica y el umbral óptimo, el *Random Forest* fue el modelo más robusto y confiable para la detección de fraudes en el conjunto de datos analizado.

El modelo final, correspondiente al *Random Forest* calibrado con regresión isotónica y ajustado con el umbral óptimo, al reentrenarlo con el 80 % de los datos se obtuvo:

Tabla 7.8: Matriz de confusión del modelo *Random Forest* (fraude) sin enfoque de negocio en entrenamiento. Fuente: Elaboración propia.

		Predicho	
		No fraude (0)	Fraude (1)
Real	No fraude (0)	25.161	81
	Fraude (1)	38	387

Tabla 7.9: Reporte de clasificación del modelo *Random Forest* (fraude) sin enfoque de negocio en entrenamiento. Fuente: Elaboración propia.

Clase	**Precision**	**Recall**	**F1-Score**	**Soporte**
0	1	1	1	25.242
1	0,83	0,91	0,87	425
Accuracy			1	25.667
Macro avg	0,91	0,95	0,93	25.667

Continuación de la Tabla 7.9

Clase	Precision	Recall	F1-Score	Soporte
Weighted avg	1	1	1	25.667

De acuerdo con la Tabla 7.8 y la Tabla 7.9, los resultados en el conjunto de entrenamiento evidenciaron un desempeño sobresaliente del modelo *Random Forest* calibrado. La matriz de confusión mostró una clasificación casi perfecta de la clase *No fraude* y un reconocimiento altamente efectivo de la clase *Fraude*, con tan solo 38 falsos negativos y 81 falsos positivos sobre un total de 25.667 observaciones. En términos de métricas, el modelo alcanzó una *Precision* del 83 % y un *Recall* del 91 % para la clase minoritaria, lo que se tradujo en un *F1-Score* de 0,87. Estas cifras reflejaron un equilibrio adecuado entre la identificación de casos fraudulentos y la reducción de falsas alarmas. El *Weighted avg* y el *Macro avg* confirmaron una estabilidad general en el rendimiento del modelo, evidenciando que el proceso de calibración isotónica contribuyó a mejorar la calidad de las predicciones sin comprometer la capacidad de generalización dentro del conjunto de entrenamiento.

Finalmente, con el conjunto de prueba los resultados fueron los siguientes:

Tabla 7.10: Matriz de confusión del modelo *Random Forest* (fraude) sin enfoque de negocio en prueba. Fuente: Elaboración propia.

		Predicha	
		No fraude (0)	Fraude (1)
Real	No fraude (0)	6.258	53
	Fraude (1)	84	22

Tabla 7.11: Reporte de clasificación del modelo *Random Forest* (fraude) sin enfoque de negocio en prueba. Fuente: Elaboración propia.

Clase	Precision	Recall	F1-Score	Soporte
0	0,99	0,99	0,99	6.311
1	0,29	0,21	0,24	106
Accuracy			0,98	6.417
Macro avg	0,64	0,60	0,62	6.417
Weighted avg	0,98	0,98	0,98	6.417

De acuerdo con la Tabla 7.10 y la Tabla 7.11, los resultados del conjunto de prueba mostraron una disminución esperada en el desempeño del modelo *Random Forest* respecto a la fase de entrenamiento. La matriz de confusión evidenció que el modelo mantuvo una alta capacidad para identificar correctamente los casos de *No fraude* (6.258 aciertos frente a 53 falsos positivos), mientras que en la clase minoritaria (*Fraude*) detectó 22 de los 106 casos reales, con 84 observaciones no identificadas. Este comportamiento se reflejó en una precisión del 29 % y un *recall* del 21 % para la clase de fraude, lo que se tradujo en un *F1-Score* de 24 %.

La estrategia de optimización de F1-Score los resultados mostraron que el *Recall* y *Precision* fue insuficiente para una implementación práctica. Esto evidenció la necesidad de metodologías complementarias (ej. enfoque ROI o ranking) que prioricen la detección masiva de fraudes aún a costa de más falsos positivos.

Importancia de características del modelo con mejor desempeño

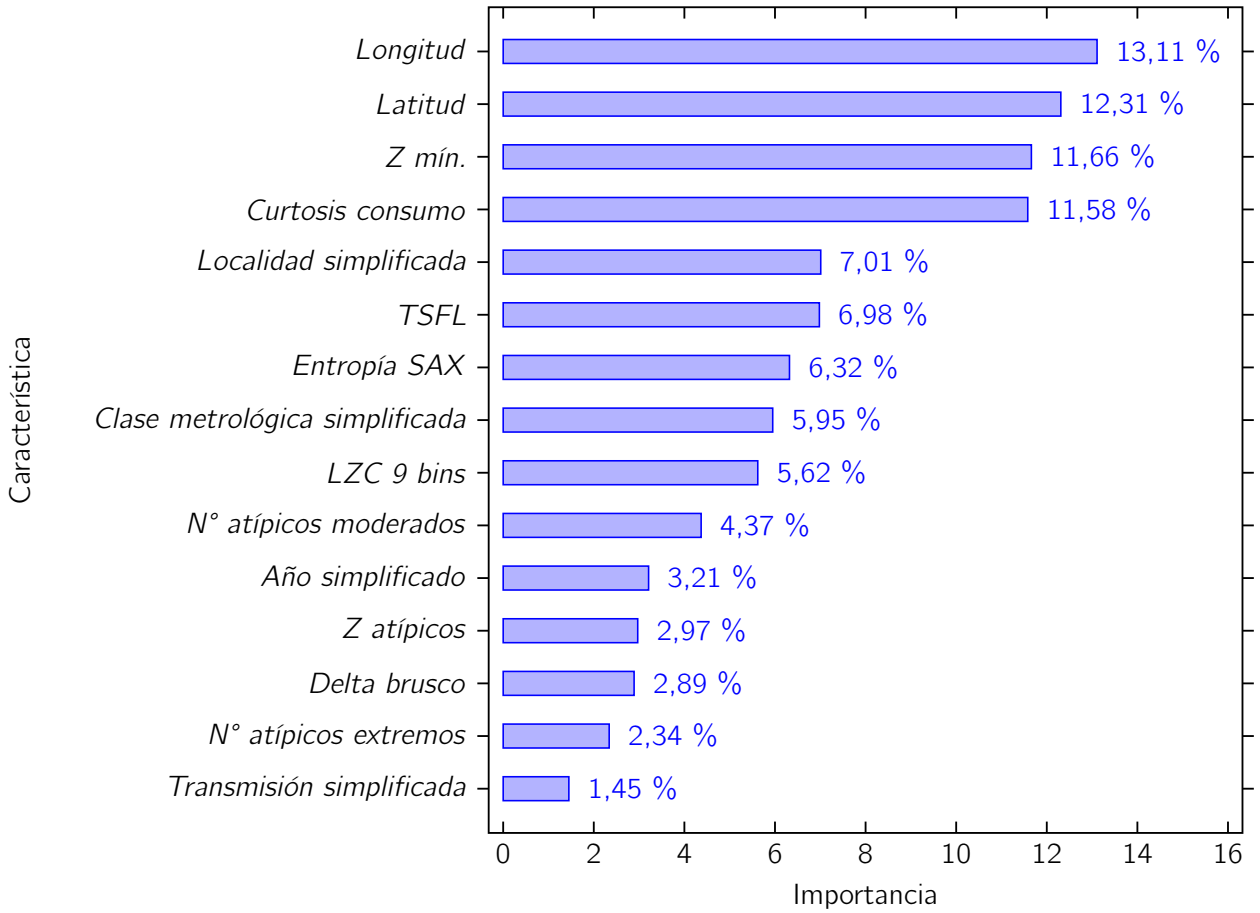

Figura 7.1: Top 15 de características más importantes del modelo *Random Forest* (fraude) sin enfoque de negocio. Fuente: Elaboración propia.

El análisis de importancia de características (Figura 7.1) mostró que las variables de localización geográfica (*Longitud* y *Latitud*) concentraron más del 25 % del poder predictivo del modelo, indicando que los patrones espaciales tuvieron un papel relevante en la detección de fraudes. Las variables asociadas al comportamiento temporal del consumo (*Z mín.*, *Curtosis consumo*, *TSFL* y *Entropía SAX*) también presentaron un peso importante, evidenciando que el modelo identificó irregularidades estadísticas y dinámicas anómalas en las series de consumo. En contraste, las variables técnicas del medidor (*Clase metrológica simplificada*, *Año simplificado* y *Transmisión simplificada*) tuvieron menor influencia, lo que mostró que su aporte fue limitado o mínimo frente a los factores espaciales y de comportamiento del consumo. En conjunto, las 15 variables principales explicaron alrededor del 98 % de la importancia total, confirmando que el *Random Forest* concentró su capacidad predictiva en un conjunto reducido de predictores. Finalmente, la variable *Diámetro simplificado* no se encontró dentro de este grupo, lo que implicó un aporte prácticamente nulo y coincidió con las pruebas de Chi-cuadrado y el coeficiente V de Cramer.

7.1.2. Modelación Random Forest vs. XGBoost con enfoque de negocio (fraude)

- **Random Forest:**

Random Forest: Búsqueda de hiperparámetros con enfoque de negocio (fraude)
La Tabla 7.12 evidenció que los mejores desempeños se alcanzaron con profundidades moderadas (`Max Depth` = 10) y un número de estimadores entre 150 y 250, es decir, el modelo se benefició de una complejidad media sin incurrir en sobreajuste. En particular, el mayor valor promedio de AUC-PR fue de 0,221 con una desviación estándar cercana a 0,027, correspondiente a configuraciones con `Min Split` = 10 y `Min Leaf` = 2 o 4.

Estos resultados indicaron una estabilidad adecuada entre las configuraciones mejor clasificadas, ya que las diferencias entre las primeras posiciones del ranking fueron mínimas (del orden de 0,002). En consecuencia, el modelo demostró un comportamiento robusto ante pequeñas variaciones de los parámetros, manteniendo un nivel de desempeño consistente en la detección de fraudes dentro de un conjunto de datos altamente desbalanceado.

Tabla 7.12: Resultados del Grid Search *Random Forest* (fraude) con enfoque de negocio.
Fuente: Elaboración propia.

n est.	Max Depth	Min Split	Min Leaf	AUC-PR Prom.	AUC-PR Desv.	Ranking
150	10	10	2	0,221	0,027	1
250	10	10	2	0,221	0,027	2
150	10	10	4	0,221	0,027	3
150	10	5	2	0,220	0,029	4
150	10	5	4	0,220	0,023	5
250	20	5	4	0,219	0,025	6
150	20	5	4	0,219	0,023	7
250	10	10	4	0,218	0,027	8
150	Ninguno	5	4	0,218	0,024	9
250	10	5	2	0,218	0,029	10
250	20	10	2	0,217	0,025	11
250	Ninguno	10	2	0,217	0,028	12
150	Ninguno	5	2	0,217	0,024	13
250	Ninguno	5	2	0,217	0,022	14
250	Ninguno	5	4	0,216	0,024	15
250	20	5	2	0,216	0,023	16
250	20	10	4	0,216	0,027	17
150	20	10	2	0,215	0,025	18
250	10	5	4	0,215	0,025	19
250	Ninguno	10	4	0,215	0,023	20
150	Ninguno	10	4	0,215	0,023	21
150	Ninguno	10	2	0,214	0,026	22
150	20	5	2	0,214	0,027	23
150	20	10	4	0,213	0,023	24

Random Forest: Métricas en cada fold con enfoque de negocio (fraude)
Teniendo en cuenta la Tabla 7.13, se observó que el modelo *Random Forest* con calibración isotónica

mostró una mejora progresiva en la métrica *AUC-PR*, pasando de 0,165 en el primer *fold* a 0,242 en el quinto, lo que reflejó una mayor capacidad del modelo para distinguir entre las clases positivas (fraude) y negativas a medida que se validaba el conjunto de datos. El *Brier-Score* se mantuvo estable alrededor de 0,014, indicando una buena calibración de las probabilidades predichas. Asimismo, las métricas de desempeño económico evidenciaron una tendencia de incremento en la *ganancia* y el *mROI*, alcanzando hasta un 79,27 % en los últimos *folds*, lo que indicó que el modelo no solo mejoró su rendimiento predictivo, sino también su rentabilidad potencial. En términos operativos, el número de verdaderos positivos (TP) aumentó ligeramente, mientras que los falsos positivos (FP) disminuyeron, consolidando un equilibrio favorable entre detección efectiva de fraudes y control de errores de clasificación.

Tabla 7.13: *Random Forest* (fraude) con enfoque de negocio - Métricas detalladas por fold. Fuente: Elaboración propia.

Fold	Calibrador	Brier-Score	AUC-PR	Ganancia	mROI	TP	FP
1	Isotónico	0,0146	0,165	$1.230.000	34,45 %	24	78
2	Isotónico	0,0141	0,212	$2.230.000	62,46 %	29	73
3	Isotónico	0,0139	0,221	$2.430.000	68,07 %	30	72
4	Isotónico	0,0138	0,238	$2.830.000	79,27 %	32	70
5	Isotónico	0,0135	0,242	$2.830.000	79,27 %	32	70

Random Forest: Métricas promedio y desviación con enfoque de negocio (fraude)
Los resultados evidenciaron un desempeño estable y coherente del modelo *Random Forest* calibrado isotónicamente bajo el enfoque de negocio (Tabla 7.14). El *AUC-PR* promedio fue de 0,22 con una desviación estándar de 0,03, lo que indicó una buena capacidad de discriminación entre clientes fraudulentos y no fraudulentos en un escenario altamente desbalanceado. El *Brier-Score* se mantuvo bajo (0,01), reflejando una alta fiabilidad en las probabilidades estimadas y demostrando que el modelo no solo clasificó correctamente, sino que también asignó niveles de riesgo consistentes.

En términos económicos, la ganancia promedio fue de $2.310.000, con una variación aproximada de $657.000, lo que reflejó una rentabilidad constante entre las diferentes particiones. Además, el *mROI* promedio fue de 0,65, lo que representó un retorno del 65 % respecto a la inversión operativa. Finalmente, los valores medios de 29 verdaderos positivos (TP) y 73 falsos positivos (FP) mostraron un equilibrio adecuado entre la detección y el costo operativo, consolidando el potencial del modelo como herramienta de apoyo en la toma de decisiones estratégicas orientadas a la reducción de pérdidas por fraude.

Tabla 7.14: Promedio y desviación estándar de las métricas del modelo *Random Forest* con enfoque de negocio. Fuente: Elaboración propia.

Métrica	Promedio	Desviación estándar
Brier-Score	0,01	0,00
AUC-PR	0,22	0,03
Ganancia	$2.310.000	$657.267
mROI	63,30 %	17,62 %

- **XGBoost:**

XGBoost: Búsqueda de hiperparámetros con enfoque de negocio (fraude)

De acuerdo con la Tabla 7.15, los resultados del *Grid Search* para el modelo *XGBoost* con enfoque de negocio mostraron valores de AUC-PR comprendidos entre 0,184 y 0,215, con desviaciones estándar que oscilaron entre 0,025 y 0,034. El mejor desempeño se obtuvo con una combinación de hiperparámetros de 150 árboles, profundidad máxima de 5 y tasa de aprendizaje de 0,05, alcanzando un valor promedio de AUC-PR de 0,215. Los modelos con tasas de aprendizaje menores o profundidades más altas presentaron un rendimiento similar, sin diferencias significativas entre las configuraciones principales. En general, las métricas evidenciaron estabilidad entre los diferentes conjuntos de validación, con variaciones mínimas en la desviación estándar.

Tabla 7.15: Resultados del Grid Search *XGBoost* (fraude) con enfoque de negocio.
Fuente: Elaboración propia.

n est.	Max Depth	Learning Rate	Subsample	Colsample ByTree	**AUC-PR Prom.**	**AUC-PR Desv. Est.**
150	5	0,05	0,80	0,80	0,215	0,032
150	7	0,05	0,80	0,80	0,215	0,034
250	5	0,05	0,80	0,80	0,211	0,034
150	3	0,05	0,80	0,80	0,210	0,033
150	5	0,10	0,80	0,80	0,206	0,025
250	3	0,05	0,80	0,80	0,205	0,030
150	3	0,10	0,80	0,80	0,205	0,03
250	7	0,05	0,80	0,80	0,205	0,028
250	3	0,10	0,80	0,80	0,200	0,032
150	7	0,10	0,80	0,80	0,19	0,03
250	5	0,10	0,80	0,80	0,191	0,03
250	7	0,10	0,80	0,80	0,184	0,025

XGBoost: Métricas en cada fold con enfoque de negocio (fraude)

A partir de los valores mostrados en la Tabla 7.16, el modelo *XGBoost* calibrado de forma isotónica presentó un desempeño variable entre los cinco *folds* evaluados. El valor del *Brier-Score* se mantuvo estable alrededor de 0,014, lo que indicó una adecuada calibración de las probabilidades predichas. En cuanto al *AUC-PR*, los valores oscilaron entre 0,167 y 0,263, alcanzando su mejor desempeño en el quinto *fold*, donde también se observó la mayor ganancia económica estimada (2.430.000) y un retorno medio sobre la inversión (*mROI*) de 68,07

Asimismo, se evidenció una tendencia ascendente tanto en la ganancia como en el *mROI* a medida que aumentaron los verdaderos positivos (TP) y disminuyeron los falsos positivos (FP). En promedio, el modelo detectó entre 27 y 32 casos de fraude por *fold*, con un número de falsos positivos cercano a 73. Estos resultados mostraron que, aunque el modelo mantuvo una calibración adecuada y un desempeño aceptable en términos de *AUC-PR*, su capacidad de detección varió entre las particiones, reflejando sensibilidad al conjunto de validación utilizado en cada iteración.

Tabla 7.16: *XGBoost* (fraude) con enfoque de negocio - Métricas detalladas por fold.
Fuente: Elaboración propia.

Fold	**Calibrador**	**Brier-Score**	**AUC-PR**	**Ganancia**	**mROI**	**TP**	**FP**
1	Isotónico	0,0146	0,167	$1.430.000	40,06 %	25	77

Continuación de la Tabla 7.16

Fold	Calibrador	Brier-Score	AUC-PR	Ganancia	mROI	TP	FP
2	Isotónico	0,0142	0,208	$2.030.000	56,86 %	28	74
3	Isotónico	0,0141	0,211	$1.830.000	51,26 %	27	75
4	Isotónico	0,0141	0,212	$2.830.000	79,27 %	32	70
5	Isotónico	0,0134	0,263	$2.430.000	68,07 %	30	72

XGBoost: Métricas promedio y desviación con enfoque de negocio (fraude)

De acuerdo con la Tabla 7.17, los resultados promediados del modelo *XGBoost* con enfoque de negocio evidenciaron un comportamiento estable y coherente con los obtenidos durante la validación cruzada. El valor medio del *Brier-Score* fue de 0,010, lo que indicó una adecuada calibración de las probabilidades, con predicciones bien ajustadas a las tasas reales de fraude. En cuanto al desempeño discriminativo, el modelo alcanzó un *AUC-PR* promedio de 0,210 con una desviación de 0,030, reflejando una capacidad moderada para distinguir entre clientes fraudulentos y no fraudulentos en un escenario altamente desbalanceado.

La ganancia media proyectada fue de $2.110.000, con una desviación estándar de $540.370, lo que indicó un rendimiento económico positivo y relativamente estable entre los distintos *folds*. El retorno promedio sobre la inversión (*mROI*) fue de 0,59, confirmando que el modelo generó beneficios netos en la mayoría de los escenarios evaluados. Finalmente, el promedio de verdaderos positivos (28,4) y falsos positivos (73,6) mostró que el modelo logró detectar fraudes de manera consistente, manteniendo un equilibrio entre sensibilidad y costo operativo. En conjunto, estos resultados ratificaron la viabilidad del modelo *XGBoost* calibrado con enfoque de negocio para apoyar decisiones en campo bajo criterios de rentabilidad.

Tabla 7.17: Promedio y desviación estándar de las métricas del modelo *XGBoost* (fraude) con enfoque de negocio. Fuente: Elaboración propia.

Métrica	Promedio	Desviación estándar
Brier-Score	0,01	0,00
AUC-PR	0,21	0,03
Ganancia	$2.110.000	$540.370
mROI	59,90 %	14,44 %

- **Comparación y discusión Random Forest vs. XGBoost con enfoque de negocio (fraude):**

Tabla 7.18: Comparación de métricas entre *Random Forest* (RF) y *XGBoost* (XGB) con enfoque de negocio (fraude). Fuente: Elaboración propia.

Métrica	RF Promedio	RF Desv. Est.	XGB Promedio	XGB Desv. Est.
Brier-Score	0,01	0,00	0,01	0,00
AUC-PR	0,22	0,03	0,21	0,03
Ganancia	$2.310.000	$657.267	$2.110.000	$540.370
mROI	64,30 %	17,62 %	59,90 %	14,44 %

Los resultados comparativos entre los modelos *Random Forest* y *XGBoost* con enfoque de negocio de la Tabla 7.18, mostraron un desempeño superior de *Random Forest* en los principales criterios de evaluación. En términos del *AUC-PR*, métrica más representativa en escenarios con alta desproporción entre clases, el modelo *Random Forest* alcanzó un valor promedio de 0,22 frente a 0,21 obtenido por *XGBoost*. Este resultado evidenció una mejor capacidad del *Random Forest* para distinguir entre clientes con y sin fraude a lo largo de distintos umbrales de decisión, optimizando la relación entre precisión y exhaustividad.

En el plano económico, la métrica de rentabilidad (*mROI*) también favoreció al modelo *Random Forest*, con un valor promedio de 0,65, superando el 0,59 obtenido por *XGBoost*. Esta diferencia reflejó que el *Random Forest* generó un mayor retorno sobre la inversión por cada inspección realizada, consolidándose como la alternativa más rentable para el despliegue operativo en campo. De forma coherente, el modelo presentó una ganancia media de $2.310.000 frente a los $2.110.000 de *XGBoost*, manteniendo al mismo tiempo una menor desviación, lo que indicó estabilidad en el rendimiento financiero.

Por último, el *Brier-Score* promedio de 0,01 en ambos modelos confirmó una buena calibración de las probabilidades predichas, aunque el *Random Forest* presentó menor variabilidad, reafirmando su consistencia en la estimación de riesgos. En conjunto, las métricas mostraron que el modelo *Random Forest* no solo logró un mejor equilibrio entre precisión y recall, sino que además ofreció mayor estabilidad y rentabilidad en su aplicación práctica.

El modelo final, correspondiente al *Random Forest* optimizado, al reentrenarlo sobre el 80 % se obtuvo:

Tabla 7.19: *Random Forest* (fraude) con enfoque de negocio - Métricas en datos de entrenamiento. Fuente: Elaboración propia.

Métrica	Valor
Brier-Score	0,011
AUC-PR	0,507
Ganancia Total	$27.045.000
mROI	151 %

Tabla 7.20: Matriz de confusión del modelo *Random Forest* (fraude) con enfoque de negocio en entrenamiento. Fuente: Elaboración propia.

		Predicha	
		No fraude (0)	Fraude (1)
Real	No fraude (0)	24.954	288
	Fraude (1)	200	225

De acuerdo con la Tabla 7.19 y la Tabla 7.20, el modelo *Random Forest* optimizado presentó un desempeño sólido al ser reentrenado sobre el 80 % de los datos de entrenamiento. El valor del *AUC-PR* alcanzó 0,507, lo que indicó una adecuada capacidad para distinguir entre fraudes y no fraudes en un contexto de desbalance severo. El *Brier-Score* de 0,011 evidenció una buena calibración de las probabilidades predichas, reforzando la confiabilidad del modelo al priorizar los casos con mayor probabilidad de fraude. En términos económicos, la ganancia total estimada fue de $27.045.000, con un margen de retorno sobre la inversión (*mROI*) de 1,51, lo que significó que el modelo habría generado un beneficio 1,5 veces superior al costo operativo de las inspecciones. Finalmente, la matriz de confusión mostró un equilibrio razonable entre los verdaderos positivos (225) y los falsos positivos (288), lo cual implicó que el modelo priorizó la detección de casos sospechosos sin sacrificar significativamente la precisión, mientras que los verdaderos negativos (24.954) y los falsos negativos (200) confirmaron un desempeño estable en la clasificación de no fraudes.

Filtrado geográfico (bolsón)

Al analizar el *Top 20* de localidades con mayor prevalencia de fraudes, presentadas en la Tabla 6.20, se observaron los siguientes resultados:

Tabla 7.21: Reporte final del modelo *Random Forest* (fraude) bajo distintas capacidades de inspección. Fuente: Elaboración propia.

Capacidad	Inspecciones	Precision	Recall Global	Recall Bolsón	Ganancia	mROI
0,5 %	32	34,38 %	10,38 %	22,45 %	$1.080.000	96,43 %
1 %	64	31,25 %	18,87 %	40,82 %	$1.760.000	78,57 %
1,5 %	96	25 %	22,64 %	48,98 %	$1.440.000	42,86 %
2 %	128	19,53 %	23,58 %	51,02 %	$520.000	11,61 %

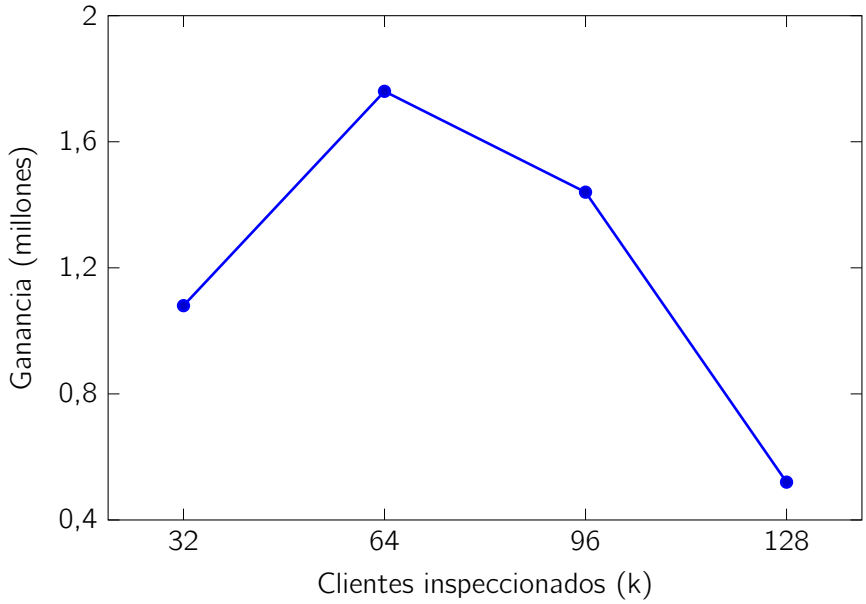

Figura 7.2: Ganancia generada vs. Esfuerzo de inspección en el bolsón.
Fuente: Elaboración propia.

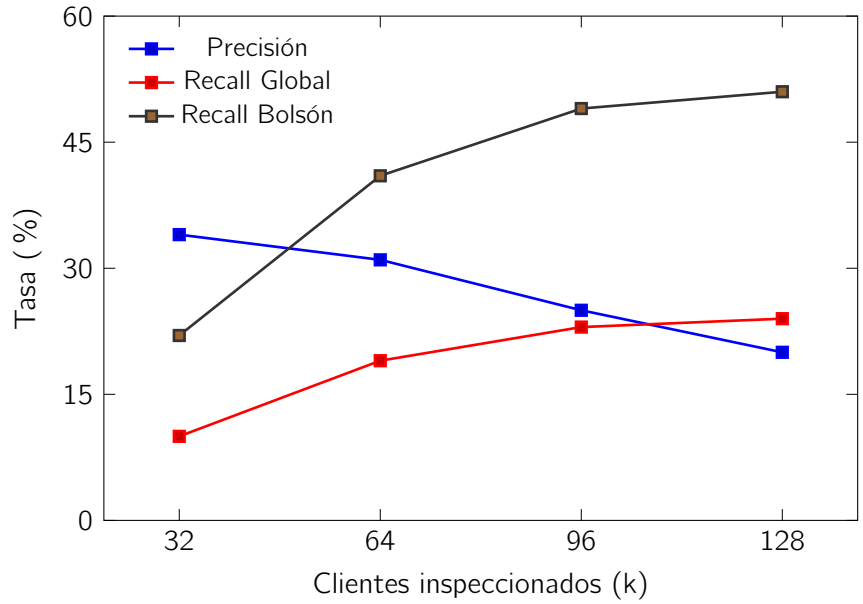

Figura 7.3: Métricas de desempeño del modelo bajo diferentes capacidades de inspección (bolsón).
Fuente: Elaboración propia.

Los resultados de la Tabla 7.21, Figura 7.2 y la Figura 7.3 mostraron que, al incrementar la capacidad de inspección, la ganancia total no aumentó de manera proporcional, mientras que la *Precision* disminuyó progresivamente. Con una capacidad del 0,5 %, el modelo alcanzó el mayor retorno (*mROI* de 96,43 %), aunque con una cobertura reducida. En contraste, al aumentar la capacidad al 2 %, se logró un mayor *Recall Global*, pero la ganancia disminuyó considerablemente debido al incremento de falsos positivos.

El punto de equilibrio se evidenció entre el 1 % y 1,5 % de capacidad de inspección, donde el modelo mantuvo un balance razonable entre la cantidad de fraudes detectados y la utilidad económica. No

obstante, se escogió el 1 % como capacidad óptima, ya que presentó la mayor *Ganancia* ($1.760.000) y un desempeño estable en términos de *Precision* y *Recall Bolsón*. En consecuencia, se consideró que este punto representó la configuración más eficiente para maximizar el beneficio con los recursos disponibles. Al realizar la comparación entre la calibración del conjunto de entrenamiento vs. conjunto de prueba se obtuvieron los siguientes resultados:

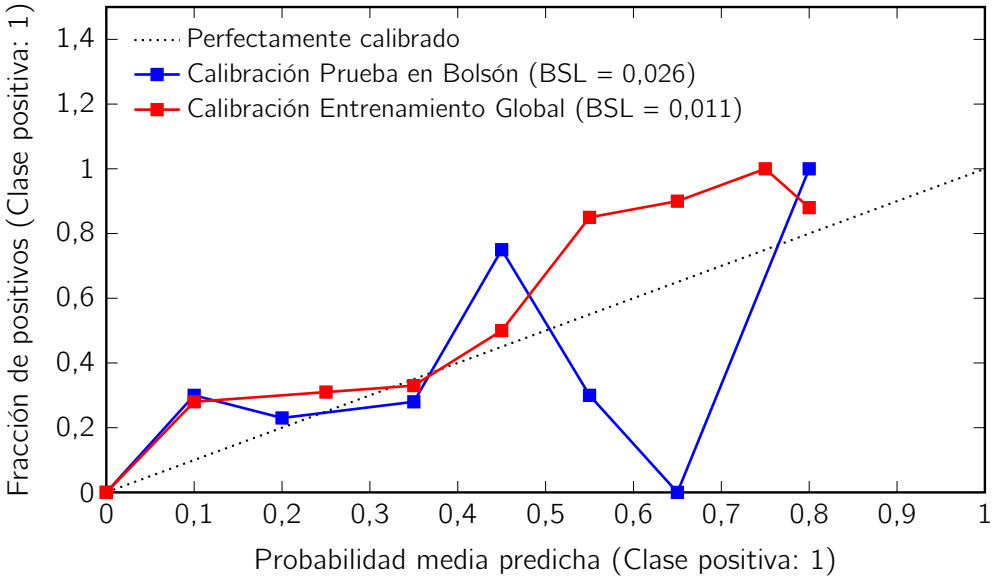

Figura 7.4: Curva de calibración: comparación entre conjuntos de entrenamiento y prueba en bolsón. Fuente: Elaboración propia.

La comparación entre las curvas de ranking y calibración de la Figura 7.4 permitió evidenciar dos aspectos complementarios del desempeño del modelo. Por un lado, la curva de ranking mostró que el modelo ordenó eficazmente a los clientes según su nivel de riesgo, concentrando los casos positivos en las primeras posiciones, lo que resultó útil para priorizar acciones de inspección o seguimiento. Por otro lado, la curva de calibración reveló una sensibilidad al cambio poblacional, ya que en el conjunto de prueba en bolsón el modelo presentó menor fiabilidad probabilística respecto al conjunto de entrenamiento. Esta pérdida parcial de calibración indicó que las probabilidades estimadas no debían interpretarse de forma absoluta, sino relativa, lo que justificó la adopción de un enfoque basado en ranking (*top-k*) en lugar de la aplicación de un umbral fijo de probabilidad para la toma de decisiones.

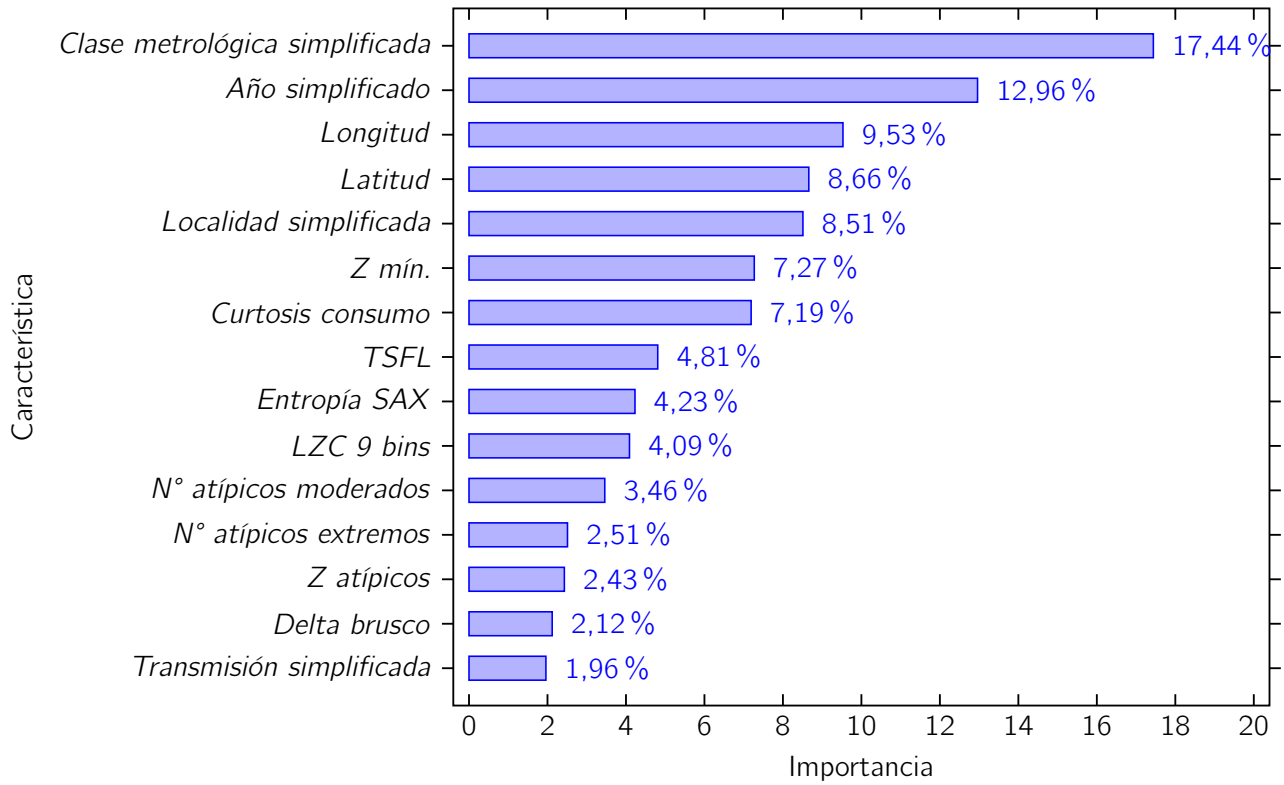

Figura 7.5: Top 15 de características más importantes del modelo *Random Forest* (fraude) con enfoque de negocio. Fuente: Elaboración propia.

De acuerdo con la Figura 7.5, las características con mayor aporte al modelo correspondieron principalmente a variables de tipo físico y metrológico. La *Clase metrológica simplificada* y el *Año simplificado* fueron las más influyentes, con una importancia relativa del 17,44 % y 12,96 %, respectivamente. Estas variables reflejaron la relevancia de la antigüedad y del tipo de instrumento en la detección de patrones asociados al fraude.

En un segundo nivel se ubicaron las variables espaciales (*Longitud*, *Latitud* y *Localidad simplificada*) y las derivadas del comportamiento del consumo (*Z mín.*, *Curtosis consumo*, *TSFL*, *Entropía SAX*, entre otras), que capturaron tanto la localización geográfica de los medidores como la forma de las series de consumo. En conjunto, estos resultados mostraron que el modelo con enfoque de negocio priorizó variables que combinaban información técnica del medidor y dinámica espacial del consumo, reforzando la naturaleza multivariable del fenómeno analizado.

7.2. MODELACIÓN NO SUPERVISADA FRAUDE

7.2.1. Modelación DBSCAN sin enfoque de negocio con optimización del F1-Score (fraude)

DBSCAN: Búsqueda de hiperparámetros sin enfoque de negocio con optimización del F1-Score
De acuerdo con la Figura 7.6, se observó que la curva de k-distancias mostró un crecimiento suave en su primera mitad, lo que indicó que la mayoría de los puntos se ubicaban en regiones densas del espacio de datos. A partir de aproximadamente el punto 2.500 en el eje x (cuando la distancia al décimo vecino

rondaba entre 0,9 y 1) se observó un ascenso abrupto que marcó la transición entre zonas de alta densidad y áreas más dispersas. Este cambio de pendiente evidenció que, desde ese umbral, los puntos requirieron distancias mucho mayores para alcanzar a su décimo vecino, rasgo propio de observaciones aisladas o atípicas. Los parámetros que maximizaron el F1-Score fueron: $\varepsilon = 1{,}43$ y `min samples` = 56, valores que se ajustaron considerando este comportamiento. De esta manera, el punto de inflexión correspondiente a $\varepsilon = 1{,}43$ se seleccionó como umbral óptimo para distinguir entre clústeres y *outliers* en el modelo DBSCAN, maximizando el F1-Score y preservando la estructura de densidad natural de los datos.

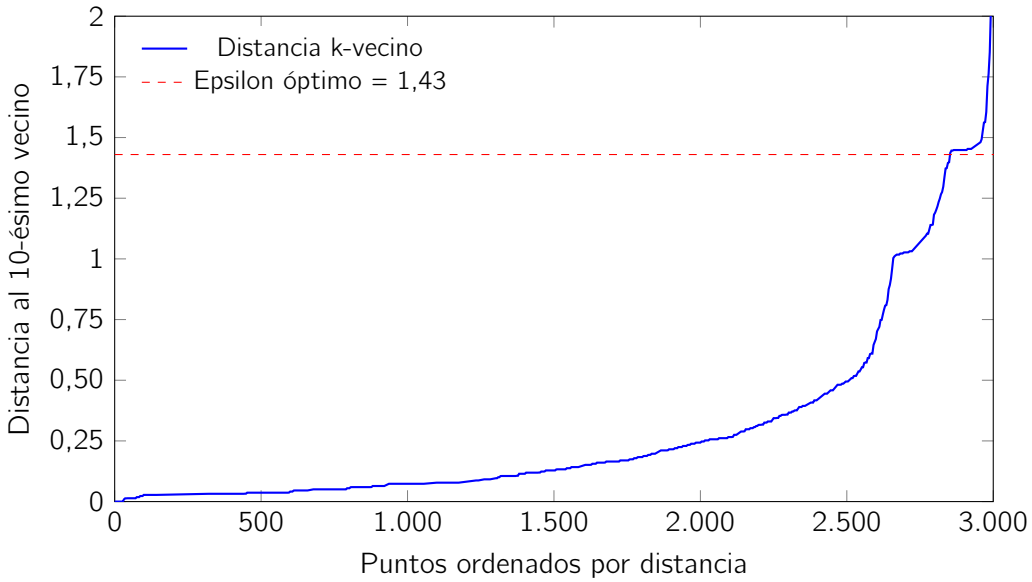

Figura 7.6: Gráfica de k-distancias para la selección del parámetro ε en el modelo DBSCAN sin enfoque de negocio con optimización del F1-Score (fraude). Fuente: Elaboración propia.

Al evaluar el modelo en el conjunto de prueba se obtuvieron los siguientes resultados:

Tabla 7.22: Matriz de confusión del modelo *DBSCAN* (fraude) sin enfoque de negocio con optimización del F1-Score en conjunto de prueba. Fuente: Elaboración propia.

		Predicha	
		No fraude (0)	Fraude (1)
Real	No fraude (0)	6.110	201
	Fraude (1)	105	1

Tabla 7.23: Reporte de clasificación del modelo *DBSCAN* (fraude) sin enfoque de negocio con optimización del F1-Score en conjunto de prueba.

Clase	**Precision**	**Recall**	**F1-Score**	**Soporte**
0	0,98	0,97	0,97	6.311
1	0	0,01	0,01	106
Accuracy			0,95	6.417

Continuación de la Tabla 7.25

Clase	Precision	Recall	F1-Score	Soporte
Macro avg	0,49	0,49	0,49	6.417
Weighted avg	0,97	0,95	0,96	6.417

Los resultados de la Tabla 7.24 y la Tabla 7.25 mostraron que, aunque *DBSCAN* fue un algoritmo potente para la detección de *outliers*, en este caso no logró separar de forma significativa los patrones fraudulentos de los normales. Esta limitación se explicó, en primer lugar, por la escasez de datos fraudulentos, ya que la baja representación de casos de fraude ($\approx 1{,}65\,\%$) restringió la capacidad del modelo para identificar adecuadamente dicha clase, incluso dentro de un marco semi-supervisado y con una propuesta metodológica adaptada al desbalance de clases. En segundo lugar, se observó una superposición de patrones en el espacio de características, lo que impidió que los consumos fraudulentos se diferenciaran claramente de los registros normales. Además, el modelo presentó una alta tasa de falsos positivos, al etiquetar como fraude varios consumos atípicos que correspondían a comportamientos legítimos pero inusuales. Finalmente, mientras que Ghamkhar et al. [47] obtuvieron resultados favorables en la detección de anomalías generales, en este estudio se evidenció que *fraude* no fue equivalente a *anomalía genérica*, dado que el fraude constituyó un evento aún más raro y complejo de distinguir, lo que explicó la disminución en el desempeño observado.

7.2.2. Modelación DBSCAN sin enfoque de negocio con optimización del AUC-PR (fraude)

DBSCAN: Búsqueda de hiperparámetros sin enfoque de negocio con optimización del AUC-PR
En este caso, la gráfica de la k distancias es igual que la Figura 7.6. A diferencia de la optimización del F1-Score, en este escenario (optimización del AUC-PR) cambió el `min samples` de 56 a 66, es decir, los parámetros que maximizaron el AUC-PR fueron: $\varepsilon = 1{,}43$ y `min samples` = 66, valores que se ajustaron considerando este comportamiento.

Al evaluar el modelo en el conjunto de prueba se obtuvieron los siguientes resultados:

Tabla 7.24: Matriz de confusión del modelo *DBSCAN* (fraude) sin enfoque de negocio con optimización del AUC-PR en conjunto de prueba. Fuente: Elaboración propia.

		Predicha	
		No fraude (0)	Fraude (1)
Real	No fraude (0)	6.076	235
	Fraude (1)	105	1

Tabla 7.25: Reporte de clasificación del modelo *DBSCAN* (fraude) sin enfoque de negocio con optimización del AUC-PR en conjunto de prueba.

Clase	Precision	Recall	F1-Score	Soporte
0	0,98	0,96	0,97	6.311

Continuación de la Tabla 7.25

Clase	Precision	Recall	F1-Score	Soporte
1	0	0,01	0,01	106
Accuracy			0,95	6.417
Macro avg	0,49	0,49	0,49	6.417
Weighted avg	0,97	0,95	0,96	6.417

De acuerdo con la Tabla 7.25 y la Tabla 7.24, el desempeño del modelo evidenció que la optimización por ranking no logró transferirse de manera efectiva al contexto práctico. Aunque los hiperparámetros fueron seleccionados para maximizar el *AUC-PR*, el modelo solo detectó un caso de fraude en el conjunto de prueba. Este resultado indicó que los consumos fraudulentos no presentaron una diferenciación suficiente respecto a los consumos atípicos legítimos, incluso después de priorizarlos mediante el proceso de ranking. En consecuencia, la capacidad discriminativa del modelo se vio limitada por la similitud estructural entre las clases y por la naturaleza sutil y poco representativa de los patrones asociados al fraude.

7.3. REGRESIÓN LOGÍSTICA FRAUDE

En esta sección se emplea la nomenclatura de M1 a M6 para identificar los diferentes modelos para la detección de fraude. La Tabla 7.26 detalla las variables incluidas en cada uno de ellos.

Tabla 7.26: Variables incluidas en la regresión logística de fraude (M1–M6).
Fuente: Elaboración propia.

Variable	M1	M2	M3	M4	M5	M6
Curtosis consumo $^{(t)}$	✓					
N° atípicos moderados $^{(t)}$	✓					
N° atípicos extremos $^{(t)}$	✓					
Z mín. $^{(t)}$	✓	✓	✓	✓	✓	
Z atípicos $^{(t)}$	✓	✓				
Delta brusco $^{(t)}$	✓					
Secuencia ceros $^{(t)}$	✓					
Entropía SAX $^{(t)}$	✓	✓	✓	✓	✓	✓
LZC 9 bins $^{(t)}$	✓	✓	✓			
TSFL $^{(t)}$	✓	✓				
Año simplificado	✓	✓	✓	✓		
Clase metrológica simplificada	✓		✓	✓	✓	
Diámetro simplificado	✓					
Localidad simplificada	✓	✓	✓	✓	✓	✓
Marca simplificada	✓	✓	✓	✓	✓	✓
Ruedas simplificada	✓					
Transmisión simplificada	✓	✓				

Nota: Las variables marcadas con (t) hacen referencia a la variable transformada y estandarizada que se aplica a la regresión logística.

7.3.1. Significancia y ajustes del modelo

Tabla 7.27: Medidas de influencia de los modelos de regresión logística (M1–M6) para la detección de fraude. Fuente: Elaboración propia.

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M1	23.390	$4{,}5446 \times 10^{22}$	1	$3{,}3592 \times 10^{14}$
M1	6.724	0,0366	0,1553	0,9533
M1	1.875	0,0362	0,0103	0,3333
M1	5.248	0,0196	0,0035	0,1171
M1	23.084	0,0172	0,0003	0,0463
M1	6.333	0,0168	0,0016	0,0634
M1	10.965	0,0167	0,0053	0,1097
M1	15.692	0,0137	0,0003	0,0236
M1	1.745	0,0136	0,0012	0,0643
M1	17.224	0,0135	0,0004	0,0403
M1	14.646	0,0125	0,0005	0,0478
M1	18.099	0,0118	0,0019	0,0772
M1	5.011	0,0116	0,0051	0,0816
M1	4.903	0,0107	0,0017	0,0098
M1	23.864	0,0102	0,0017	0,0583
M2	6.724	0,0482	0,1567	0,9050
M2	1.875	0,0463	0,0087	0,2884
M2	5.248	0,0266	0,0034	0,1175
M2	23.084	0,0247	0,0003	0,0443
M2	15.692	0,0169	0,0002	0,0248
M2	23.864	0,0137	0,0016	0,0580
M2	22.006	0,0125	0,0020	0,0671
M2	5.011	0,0118	0,0034	0,0845
M2	4.903	0,0104	0,0012	0,0105
M2	6.333	0,0096	0,0011	0,0394
M2	5.274	0,0096	0,0029	0,0159
M2	2.262	0,0085	0,0046	0,0589
M2	9.130	0,0076	0,0008	0,0186
M2	10.235	0,0076	0,0222	0,1433
M2	9.926	0,0076	0,0060	0,0757
M3	6.724	0,0559	0,1440	0,8770
M3	1.875	0,0535	0,0085	0,2871
M3	5.248	0,0298	0,0037	0,1224
M3	23.084	0,0283	0,0003	0,0455
M3	15.692	0,0188	0,0003	0,0267
M3	23.864	0,0148	0,0017	0,0623
M3	22.006	0,0144	0,0020	0,0671
M3	5.011	0,0136	0,0031	0,0812
M3	4.903	0,0113	0,0012	0,0118

Continuación de la Tabla 7.27

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M3	6.333	0,0107	0,0013	0,0392
M3	5.274	0,0102	0,0016	0,0129
M3	9.926	0,0087	0,0050	0,0698
M3	8.785	0,0083	0,0013	0,0436
M3	18.133	0,0082	0,0005	0,0156
M3	695	0,0081	0,0008	0,0190
M4	6.724	0,0694	0,1150	0,8241
M4	5.248	0,0361	0,0028	0,1085
M4	23.084	0,0331	0,0005	0,0606
M4	23.864	0,0176	0,0019	0,0651
M4	15.692	0,0175	0,0003	0,0318
M4	22.006	0,0169	0,0023	0,0718
M4	5.011	0,0160	0,0035	0,0858
M4	4.903	0,0128	0,0035	0,0171
M4	5.274	0,0125	0,0011	0,0094
M4	6.333	0,0124	0,0009	0,0342
M4	5.056	0,0098	0,0030	0,0160
M4	9.926	0,0097	0,0012	0,0412
M4	8.785	0,0096	0,0017	0,0475
M4	3.295	0,0087	0,0012	0,0291
M4	2.262	0,0086	0,0018	0,0436
M5	6.724	0,0739	0,0235	0,4345
M5	5.248	0,0426	0,0018	0,0908
M5	23.084	0,0369	0,0011	0,0809
M5	23.864	0,0198	0,0014	0,0585
M5	22.006	0,0189	0,0024	0,0730
M5	15.692	0,0187	0,0008	0,0467
M5	5.011	0,0185	0,0022	0,0718
M5	4.903	0,0151	0,0034	0,0127
M5	6.333	0,0134	0,0008	0,0322
M5	5.274	0,0129	0,0024	0,0101
M5	5.056	0,0116	0,0029	0,0116
M5	9.926	0,0101	0,0021	0,0513
M5	10.235	0,0101	0,0021	0,0511
M5	8.785	0,0101	0,0021	0,0511
M5	6.680	0,0101	0,0021	0,0509
M6	6.724	0,0739	0,0235	0,4345
M6	5.248	0,0426	0,0018	0,0908
M6	23.084	0,0369	0,0011	0,0809
M6	23.864	0,0198	0,0014	0,0585
M6	22.006	0,0189	0,0024	0,0730
M6	15.692	0,0187	0,0008	0,0467

Continuación de la Tabla 7.27

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M6	5.011	0,0185	0,0022	0,0718
M6	4.903	0,0151	0,0034	0,0127
M6	6.333	0,0134	0,0008	0,0322
M6	5.274	0,0129	0,0024	0,0101
M6	5.056	0,0116	0,0029	0,0116
M6	9.926	0,0101	0,0021	0,0513
M6	10.235	0,0101	0,0021	0,0511
M6	8.785	0,0101	0,0021	0,0511
M6	6.680	0,0101	0,0021	0,0509

7.3.2. GVIF ajustado

Tabla 7.28: GVIF ajustado para los M1 a M6 de fraude.
Fuente: Elaboración propia.

Variable	M1	M2	M3	M4	M5	M6
Curtosis consumo $^{(t)}$	3,08					
N° atípicos moderados $^{(t)}$	1,63					
N° atípicos extremos $^{(t)}$	1,82					
Z mín. $^{(t)}$	1,71	1,43	1,35	1,32	1,31	
Z atípicos $^{(t)}$	1,42	1,08				
Delta brusco $^{(t)}$	1,35					
Secuencia ceros $^{(t)}$	1,15					
Entropía SAX $^{(t)}$	1,19	1,10	1,08	1,08	1,06	1,04
LZC 9 bins $^{(t)}$	3,60					
$(LZC\ 9\ bins)^2$ $^{(t)}$		2,01	1,88			
TSFL $^{(t)}$	2,73	2,62				
Año simplificado	1,18	1,13	1,12	1,12		
Clase metrológica simplificada	1,55		1,47	1,45	1,23	
Diámetro simplificado	1,03					
Localidad simplificada	1,24	1,23	1,12	1,07	1,07	1,02
Marca simplificada	1,07	1,05	1,05	1,05	1,03	1,02
Ruedas simplificada	1,06					
Transmisión simplificada	1,08	1,08				

Nota: Las variables marcadas con (t) hace referencia a la variable transformada + estandarizada que se aplica a la regresión logística.

Los resultados del análisis del GVIF ajustado de la Tabla 7.28 evidenciaron la ausencia de multicolinealidad significativa entre las variables incluidas en los modelos. En el Modelo 1 se observaron los valores más altos para *Curtosis consumo* (3,08), *LZC 9 bins* (3,60) y *TSFL* (2,73); sin embargo, todos permanecieron por debajo del umbral crítico de 5, indicando independencia aceptable entre predictores. A partir del Modelo 2, los GVI ajustados se redujeron considerablemente, alcanzando valores cercanos a 1, lo que refleja el efecto positivo de las transformaciones aplicadas y la depuración de va-

riables redundantes. En general, los resultados confirmaron que los modelos presentan una estructura estable y que las variables seleccionadas no generan colinealidad que pueda afectar la estimación o interpretación de los coeficientes.

7.3.3. Métricas de ajuste (prueba de Hosmer-Lemeshow)

- **Hipótesis nula** (H_0)**:** El modelo se ajusta bien a los datos. Es decir, no hay diferencia significativa entre los valores observados y los esperados, y el modelo tiene una buena calibración.
- **Hipótesis alternativa** (H_1)**:** El modelo no se ajusta bien a los datos. Es decir, existe una diferencia significativa entre los valores observados y los esperados, y el modelo no tiene una buena calibración.

Tabla 7.29: Indicadores de ajuste y prueba de bondad para los modelos de fraude.
Fuente: Elaboración propia.

Modelo	llh	llhNull	McFadden	r_{ML}	r_{CU}	p-value	Decisión (H_0)
M1	-1.774,24	-2.164,35	0,18	0,02	0,19	≈ 0	Rechazar H_0
M2	-1.774,56	-2.164,35	0,18	0,02	0,19	≈ 0	Rechazar H_0
M3	-1.776,85	-2.164,35	0,17	0,02	0,19	≈ 0	Rechazar H_0
M4	-1.783,06	-2.164,35	0,17	0,02	0,18	≈ 0	Rechazar H_0
M5	-1.835,46	-2.164,35	0,15	0,02	0,16	≈ 0	Rechazar H_0
M6	-2.069,17	-2.164,35	0,04	≈ 0	0,04	≈ 0	Rechazar H_0

De acuerdo con los resultados obtenidos en la Tabla 7.29, los valores del estadístico McFadden oscilaron entre 0,04 y 0,18, lo que indicó que, aunque los modelos logísticos presentaron cierta capacidad explicativa, su poder predictivo global fue limitado, especialmente en las primeras versiones. En todos los casos, el p-value asociado a la prueba de Hosmer–Lemeshow resultó menor a 0,05, por lo cual se rechazó la hipótesis nula (H_0) y se afirmó que ninguno de los modelos logró una calibración perfecta entre los valores observados y los esperados.

7.3.4. Métricas de evaluación

Tabla 7.30: Métricas de desempeño por modelo de fraude en validación cruzada y prueba.
Fuente: Elaboración propia.

Métrica	M1	M2	M3	M4	M5	M6
F2 CV entrenamiento	0,07	0,07	0,07	0,07	0,07	0,07
Spec. CV entrenamiento	0,22	0,20	0,28	0,20	0,20	0,27
Recall CV entrenamiento	0,78	0,80	0,72	0,78	0,80	0,75
Prec. CV entrenamiento	0,01	0,01	0,01	0,01	0,01	0,01
Threshold CV	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	0,01
F2 entrenamiento	0,09	0,09	0,09	0,09	0,09	0,09
F1 entrenamiento	0,04	0,03	0,04	0,03	0,03	0,04
Recall	0,96	0,97	0,92	0,96	0,97	0,92
Precision	0,02	0,02	0,02	0,01	0,01	0,02
Specificity	0,22	0,20	0,29	0,20	0,18	0,25

Continuación de la Tabla 8.16

Métrica	M1	M2	M3	M4	M5	M6
Balanced Accuracy	0,59	0,58	0,60	0,58	0,57	0,59
AUC-ROC	0,79	0,79	0,79	0,78	0,78	0,69
AUC-PR	0,12	0,12	0,12	0,12	0,11	0,03
Brier-Score	0,01	0,01	0,01	0,01	0,01	0,01
LogLoss	0,07	0,07	0,07	0,07	0,07	0,08

De acuerdo con la Tabla 8.16, los seis modelos evaluados presentaron un comportamiento similar en términos de desempeño, con métricas consistentes tanto en la validación cruzada como en la prueba. Los valores de F2 estuvieron sobre 0,07, lo que evidenció un rendimiento modesto en la detección de fraudes, privilegiando el recall sobre la precisión. En efecto, los modelos alcanzaron recalls elevados (entre 0,75 y 0,97), lo que indicó una buena capacidad para identificar casos positivos de fraude, aunque con una precisión muy baja (0,02), reflejando un número considerable de falsos positivos.

El *AUC-ROC* se mantuvo estable entre 0,78 y 0,79, lo que evidencia una discriminación aceptable entre clases, mientras que los valores de *AUC-PR* fueron inferiores a 0,13, mostrando la dificultad de los modelos para distinguir fraudes en contextos de desbalance extremo. Las métricas de Specificity (0,18–0,29) y Balanced Accuracy (0,58–0,61) confirman este patrón: los modelos lograron capturar la mayoría de los fraudes, pero con un costo significativo en la clasificación correcta de los no fraudes.

Los errores de calibración, medidos mediante *Brier-Score* y *LogLoss*, se mantuvieron bajos y estables, lo que evidenció que las probabilidades predichas fueron razonablemente coherentes con las observaciones. En conjunto, los resultados indicaron que los modelos priorizaron la sensibilidad, logrando una detección robusta de fraudes a costa de precisión y especificidad, lo cual es consistente con el objetivo de minimizar falsos negativos en contextos de riesgo operativo.

Tabla 7.31: Indicadores de desempeño y calibración por modelo de fraude en los conjuntos de entrenamiento y prueba. Fuente: Elaboración propia.

Métrica	M1	M2	M3	M4	M5	M6
Entrenamiento						
AUC-ROC	0,80	0,80	0,80	0,80	0,76	0,68
AUC-PR	0,18	0,16	0,16	0,16	0,14	0,03
Brier	0,01	0,01	0,01	0,01	0,01	0,01
LogLoss	0,06	0,06	0,06	0,06	0,07	0,08
Hosmer–Lemeshow	0	0	≈ 0	0	≈ 0	≈ 0
Intercepto	0	0	0	0	0	0
Pendiente	1	1	1	1	1	1
Prueba						
AUC-ROC	0,79	0,79	0,79	0,78	0,78	0,69
AUC-PR	0,12	0,12	0,12	0,12	0,11	0,03
Brier	0,01	0,01	0,01	0,01	0,01	0,01
LogLoss	0,07	0,07	0,07	0,07	0,07	0,08

Continuación de la Tabla 7.31

Métrica	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
Hosmer–Lemeshow	≈ 0	0,02	≈ 0	0,28	0,10	0,18
Intercepto	0,03	0,03	0,04	0,03	0,04	-0,01
Pendiente	0,94	0,93	0,93	0,94	0,96	1,07

En cuanto a las métricas de desempeño y calibración de la Tabla 7.31, se observó que los Modelos 1 a 4 presentaron comportamientos consistentes tanto en entrenamiento como en prueba, con valores de *AUC-ROC* cercanos a 0,80 y pérdidas logarítmicas (LogLoss) bajas, lo cual indicó una capacidad adecuada para discriminar entre casos de fraude y no fraude. Los Modelos 5 y 6, en contraste, mostraron un deterioro en el poder predictivo, especialmente el Modelo 6, cuyo *AUC-ROC* fue de 0,69 en prueba, reflejando una menor capacidad de separación. Los valores del *AUC-PR* siguieron una tendencia similar, confirmando una reducción en la precisión del modelo ante clases desbalanceadas. En términos de calibración, los p-valores de la prueba de Hosmer–Lemeshow resultaron muy bajos en entrenamiento, lo que indicó sobreajuste parcial, aunque en el conjunto de prueba se observaron valores más equilibrados, cercanos a 0,28 y 0,37 para algunos modelos. Finalmente, las pendientes de calibración se mantuvieron cercanas a 1, lo cual evidenció una relación lineal adecuada entre las probabilidades predichas y los valores observados, mientras que los interceptos se aproximaron a cero, reflejando un sesgo mínimo en las predicciones.

7.3.5. Modelo seleccionado

En relación con los resultados anteriores, el modelo seleccionado fue el M4. Aunque el modelo cumplió con los supuestos de los modelos aditivos generalizados (GAM) para las variables *Z mín.* $^{(t)}$ y *Entropía SAX* $^{(t)}$, no superó la prueba de Hosmer–Lemeshow, lo que indicó ciertas discrepancias entre las probabilidades observadas y las predichas. En particular, se observó que el modelo tendió a asignar probabilidades bajas a la clase fraude, presentando dificultades para distinguir correctamente entre las clases.

Cabe señalar que la prueba de Hosmer–Lemeshow puede resultar sensible al tamaño muestral y a la desproporción de clases, por lo que su resultado no implica necesariamente un mal ajuste global del modelo. Por esta razón, se complementó el análisis con métricas continuas y curvas de calibración por deciles, a fin de evaluar de manera más precisa la calibración del modelo.

A pesar de la señal de falta de ajuste evidenciada por la prueba de Hosmer–Lemeshow, no fue necesario realizar una calibración adicional, ya que al analizar las métricas *LogLoss* y *Brier-Score*, así como la calibración por deciles, los resultados se mantuvieron dentro de márgenes aceptables.

Asimismo, aunque algunos valores de Cook's Distance fueron relativamente altos, ninguno excedió los umbrales críticos que habrían indicado una influencia desproporcionada de alguna observación en el modelo. Además, los valores de Hat y MaxAbsDFBeta no evidenciaron la presencia de puntos extremadamente influyentes. Por lo tanto, el Modelo 4 pareció encontrarse razonablemente bien ajustado, sin que existiera la necesidad urgente de eliminar observaciones influyentes. De manera complementaria, estos son los valores de los coeficientes estimados para la regresión logística del M4:

Tabla 7.32: Resultados del modelo 4 (M4) de regresión logística para fraude.
Fuente: Elaboración propia.

Variable	Estimación	Odds Ratio	Error estándar	Valor z	Pr(>|z|)	Signif.
Intercepto	-5,71	0,00	1,02	-5,60	≈ 0	***
Z mín. $^{(t)}$	0,18	1,19	0,07	2,64	≈ 0	**
Entropía SAX $^{(t)}$	0,18	1,19	0,06	2,76	≈ 0	**
LS | Zona Colchagua	0,16	1,17	0,40	0,40	0,68	
LS | Zona Concepción Arauco	-1,22	0,29	0,21	-5,57	≈ 0	***
LS | Zona Litoral Maule	-0,04	0,96	0,43	-0,09	0,92	
LS | Zona Maule	0,47	1,59	0,22	2,07	0,03	*
LS | Zona Ñuble	0,49	1,63	0,22	2,19	0,02	*
LS | Zona O'Higgins	0,67	1,95	0,24	2,77	≈ 0	**
LS | Otras zonas	-0,62	0,53	0,24	-2,50	0,01	*
AS | Años 90	0,53	1,69	1,04	0,51	0,60	
AS | Años 2000	0,90	2,45	1,00	0,89	0,36	
AS | Años 2010	2,41	11,13	1,00	2,39	0,01	*
CMS | Otras clases	1,64	5,15	0,14	11,06	≈ 0	***
MS | Elster-Ex-Tavira	-1,63	0,19	0,72	-2,26	0,02	*
MS | Lautaro-Sensus	-0,60	0,54	0,20	-3,00	≈ 0	**
MS | Tavira-Iberconta	-0,50	0,60	0,51	-0,98	0,32	
MS | Marca no vigente	-0,30	0,74	0,23	-1,29	0,19	
MS | Otras marcas	-0,19	0,82	1,03	-0,18	0,85	

Abreviaciones: Localidad simplificada (LS) | Año simplificado (AS) | Clase metrológica simplificada (CMS) | Marca simplificada (MS).

En relación con la Tabla 7.32, las variables categóricas incluidas en el modelo, se establecieron las siguientes categorías de referencia: *Zona Bio Bío interior* para la variable *Localidad simplificada*, *Antes 1990* para *Año simplificado*, *Clase 100* para *Clase metrológica simplificada* y *CCM-Maipo-Actaris* para *Marca simplificada*. Estas categorías se tomaron como base para la comparación de los efectos de las demás modalidades en la estimación de las probabilidades de fraude.

Los resultados del modelo permiten comprender de manera directa cómo se configura el fraude en el parque de medidores de la empresa y qué factores incrementan o reducen la probabilidad de ocurrencia. En primer lugar, las variables de comportamiento asociadas a la serie temporal del consumo, como Z mín. $^{(t)}$ y Entropía SAX $^{(t)}$, presentan coeficientes positivos y odds ratios superiores a uno, lo que indica que los casos de fraude tienden a mostrar consumos con caídas abruptas y patrones temporales inestables. Este comportamiento es consistente con intervenciones manuales sobre el medidor o alteraciones en el flujo normal del consumo, por lo que estos indicadores se consolidan como señales tempranas de riesgo operativo.

Desde la dimensión territorial, algunas zonas muestran odds ratios mayores que la categoría de referencia, lo que refleja la existencia de concentraciones espaciales donde el fraude ocurre con mayor frecuencia. Ello es coherente con patrones observados previamente en el análisis exploratorio y proporciona información útil para priorizar inspecciones en áreas donde el rendimiento esperado del operativo

es más alto. En contraste, otras zonas presentan odds ratios inferiores a uno, lo que indica que, en comparación con la zona de referencia, registran menor propensión al fraude bajo las mismas condiciones de consumo y características técnicas.

En cuanto a las características técnicas del medidor, los dispositivos instalados en la década de 2010 y aquellos pertenecientes a clases metrológicas distintas a la referencia exhiben los odds ratios más elevados de toda la tabla. Esto indica que la probabilidad de fraude aumenta cuando los medidores presentan ciertas configuraciones técnicas o antigüedades específicas. Estos resultados permiten concluir que el riesgo no depende únicamente del comportamiento de consumo, sino también de la interacción entre factores tecnológicos y operativos del parque de medidores. Por otro lado, algunas marcas presentan odds ratios menores que uno, lo que implica una menor probabilidad relativa de fraude en comparación con la marca de referencia, lo que aporta evidencia empírica para respaldar decisiones de renovación y adquisición de equipos.

En conjunto, el modelo revela que el fraude en el sistema de medición no ocurre de manera aleatoria. La probabilidad de incidencia se explica por una combinación de consumo anómalo, condiciones territoriales particulares y características técnicas del medidor. Esta información constituye una base cuantitativa relevante para orientar la planificación operativa de la empresa, optimizar la asignación de recursos destinados a inspección y fortalecer las estrategias de mantenimiento y recambio de equipos.

La curva de calibración (Figura 7.7) complementó estos hallazgos, mostrando una alta correspondencia entre las probabilidades predichas y las tasas observadas. Esto indicó que el modelo, además de incluir predictores estadísticamente significativos, estimó correctamente la magnitud del riesgo, reflejando un buen equilibrio entre discriminación y calibración.

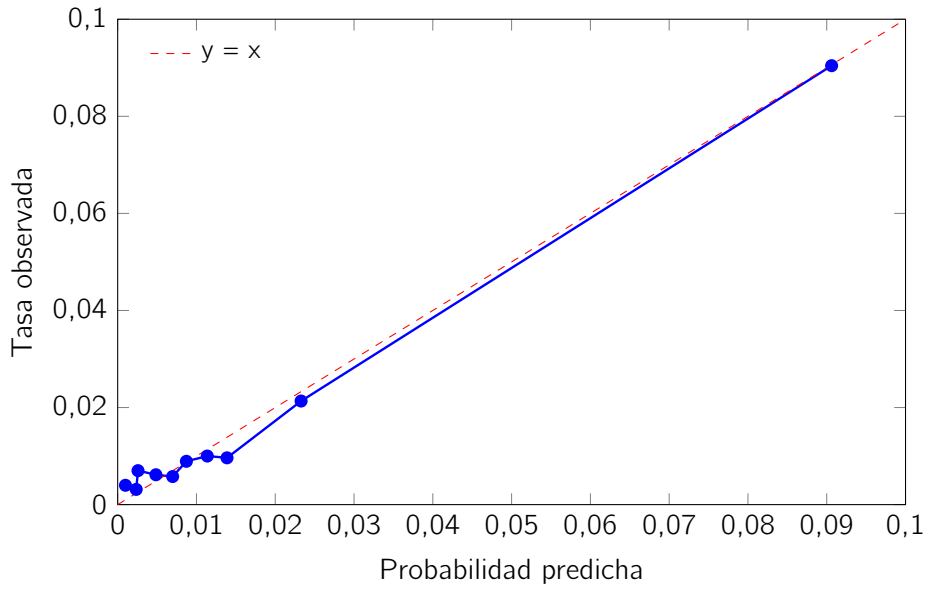

Figura 7.7: Curva de calibración por decil para el M4. Fuente: Elaboración propia.

7.3.6. Ecuación del modelo seleccionado

La ecuación que se muestra a continuación corresponde al modelo de regresión logística que obtuvo el mejor desempeño predictivo en la etapa de modelación (M4). En ella se expresan los coeficientes

estimados para cada variable, los cuales indican cómo se modifica el logit de la probabilidad de fraude según el comportamiento del consumo, la zona geográfica asociada al cliente y las características del medidor instalado. Esta representación permite interpretar directamente el efecto de cada predictor sobre el riesgo estimado:

$$\begin{aligned}\text{logit}(p) = \; & -5,71 + 0,18\,\text{Z mín.}^{(t)} + 0,18\,\text{Entropía SAX}^{(t)} \\ & + 0,16\,\text{LS} \mid \text{Zona Colchagua} - 1,22\,\text{LS} \mid \text{Zona Concepción Arauco} \\ & - 0,04\,\text{LS} \mid \text{Zona Litoral Maule} + 0,47\,\text{LS} \mid \text{Zona Maule} \\ & + 0,49\,\text{LS} \mid \text{Zona Ñuble} + 0,67\,\text{LS} \mid \text{Zona O'Higgins} \\ & - 0,62\,\text{LS} \mid \text{Otras zonas} + 0,90\,\text{AS} \mid \text{Años 2000} \\ & + 2,41\,\text{AS} \mid \text{Años 2010} + 0,53\,\text{AS} \mid \text{Años 90} \\ & + 1,64\,\text{CMS} \mid \text{Otras clases} - 1,63\,\text{MS} \mid \text{Elster-Ex-Tavira} \\ & - 0,60\,\text{MS} \mid \text{Lautaro-Sensus} - 0,30\,\text{MS} \mid \text{Marca no vigente} \\ & - 0,19\,\text{MS} \mid \text{Otras marcas} - 0,50\,\text{MS} \mid \text{Tavira-Iberconta}.\end{aligned}$$

Abreviaciones: Localidad simplificada (LS) | Año simplificado (AS) | Clase metrológica simplificada (CMS) | Marca simplificada (MS).

7.3.7. Prueba de Moran's I y dependencia espacial

Al aplicar la prueba de Moran's I con k = 10, los resultados mostraron que el modelo GLM presentaba una correlación espacial significativa en los residuos, con un p-value ≈ 0, lo cual rechaza la hipótesis nula que plantea la ausencia de autocorrelación espacial. Este resultado indicó que los residuos del modelo están espacialmente correlacionados, es decir, el modelo no ha capturado completamente la estructura espacial subyacente y podría beneficiarse de la inclusión de términos espaciales adicionales.

7.3.8. Comparación de AIC y BIC

Tabla 7.33: Comparación de modelos de fraude según criterios AIC y BIC.
Fuente: Elaboración propia.

Modelo	AIC	BIC
M1	3.612,47	3.873,36
M2	3.593,12	3.772,49
M3	3.593,69	3.756,75
M4	3.604,11	3.759,02
M5	3.702,91	3.833,36
M6	4.166,33	4.280,47

La comparación de los modelos según los criterios de información AIC y BIC (Tabla 7.33) mostró que los modelos M2 y M3 presentaron los valores más bajos en ambos indicadores, lo que indicó un mejor ajuste estadístico en términos de parsimonia y verosimilitud. En particular, el modelo M3 registró el menor BIC (3.756,75), mientras que el M2 alcanzó el AIC más reducido (3.593,12). Sin embargo, las diferencias entre ambos modelos fueron marginales (ΔAIC $\approx$ 0,6; ΔBIC $\approx$ 15,7), por lo que se consideró que ambos ofrecieron un desempeño estadísticamente comparable.

El modelo M4 obtuvo valores ligeramente superiores (AIC = 3.604,11; BIC = 3.759,02), lo que representó un incremento menor respecto al modelo M3 (ΔAIC $\approx$ 10; ΔBIC $\approx$ 2,3). Estas diferencias se mantuvieron dentro de los márgenes aceptables para la comparación de modelos, lo que permitió considerar al M4 como una alternativa estadísticamente competitiva. En contraste, los modelos M5 y M6 presentaron incrementos notables en ambos criterios, reflejando un menor ajuste global y una mayor complejidad no compensada por mejoras en la capacidad explicativa.

A diferencia de los modelos con mejor desempeño según AIC y BIC, el modelo M4 cumplió con los principales supuestos de la regresión logística, incluyendo la linealidad del logit, la ausencia de multicolinealidad y un comportamiento adecuado en la calibración. Aunque su desempeño en AIC y BIC fue levemente inferior, este modelo ofreció una mayor estabilidad de los coeficientes y una coherencia interpretativa más sólida con la estructura del fenómeno de fraude. Por estas razones, se priorizó la selección del modelo M4 como el modelo final, al representar un equilibrio adecuado entre ajuste, parsimonia e interpretabilidad.

7.3.9. Análisis de la variación en dependencia espacial con k vecinos cercanos

Tabla 7.34: M4: Comparación de valores *p* obtenidos para diferentes vecinos *k* en los modelos GLM y GAM. Fuente: Elaboración propia.

k vecinos	p-value GLM	p-value GAM
20	≈ 0	≈ 0
40	0,06	0,05
60	0,01	0,01
80	≈ 0	0,01

De acuerdo con la Tabla 7.34, en todos los casos, los valores p fueron bajos (menores a 0,06), lo que indicó la presencia de autocorrelación espacial remanente en los residuos. No obstante, se observó que el modelo GAM mantuvo consistentemente valores p ligeramente mayores que el GLM, evidenciando una menor dependencia espacial residual. Esto confirmó que el GAM logró atenuar la autocorrelación espacial de manera más efectiva, aunque en ambos modelos la dependencia no se eliminó por completo. En el contexto del presente trabajo de grado, esta persistencia de autocorrelación espacial implica que la suposición de independencia de los errores no se cumple totalmente, lo cual afecta la validez formal de los modelos, aunque no invalida su utilidad predictiva. Por tal razón, y considerando simultáneamente los criterios AIC, BIC, la calibración y la capacidad discriminativa, el modelo resultante corresponde al Modelo 4 (M4), cuya estructura logit incorpora las variables técnicas del medidor, medidas de complejidad de las series de consumo y grupos de consumo que mostraron significancia estadística y estabilidad en sus coeficientes. La ecuación final es la siguiente:

$$
\begin{aligned}
\text{logit}(p) = \; & -5,71 + 0,18\,\text{Z mín.}^{(t)} + 0,18\,\text{Entropía SAX}^{(t)} \\
& + 0,16\,\text{LS} \mid \text{Zona Colchagua} - 1,22\,\text{LS} \mid \text{Zona Concepción Arauco} \\
& - 0,04\,\text{LS} \mid \text{Zona Litoral Maule} + 0,47\,\text{LS} \mid \text{Zona Maule} \\
& + 0,49\,\text{LS} \mid \text{Zona Ñuble} + 0,67\,\text{LS} \mid \text{Zona O'Higgins} \\
& - 0,62\,\text{LS} \mid \text{Otras zonas} + 0,90\,\text{AS} \mid \text{Años 2000} \\
& + 2,41\,\text{AS} \mid \text{Años 2010} + 0,53\,\text{AS} \mid \text{Años 90} \\
& + 1,64\,\text{CMS} \mid \text{Otras clases} - 1,63\,\text{MS} \mid \text{Elster-Ex-Tavira} \\
& - 0,60\,\text{MS} \mid \text{Lautaro-Sensus} - 0,30\,\text{MS} \mid \text{Marca no vigente} \\
& - 0,19\,\text{MS} \mid \text{Otras marcas} - 0,50\,\text{MS} \mid \text{Tavira-Iberconta.} \\
& + f(\text{long, lat}).
\end{aligned}
$$

Abreviaciones: Localidad simplificada (LS) | Año simplificado (AS) | Clase metrológica simplificada (CMS) | Marca simplificada (MS).

7.4. PROPUESTA FINAL: METODOLOGÍA OPERATIVA FRAUDE

Con base en los hallazgos del estudio, se plantea una metodología replicable y actualizable que las empresas sanitarias pueden utilizar para predecir fraude y priorizar inspecciones. Esta propuesta sintetiza las mejores prácticas identificadas, los modelos con mejor desempeño y las variables mínimas necesarias para garantizar la capacidad predictiva del sistema.

1. VARIABLES REQUERIDAS

A partir de la revisión de resultados, se concluye que el desempeño depende principalmente de dos grandes grupos de variables: (i) características del medidor y (ii) características derivadas de la serie de consumo mensual. A continuación, se presentan las variables imprescindibles para correr actualizaciones periódicas:

1.1. Variables estructurales del medidor (estáticas):

Estas variables cambian poco en el tiempo y provienen de las bases comerciales y metrológicas:

- *Año simplificado*
- *Clase metrológica simplificada*
- *Diámetro simplificado*
- *Marca simplificada*
- *Transmisión simplificada*
- *Ruedas simplificadas*
- *Localidad simplificada y ubicación georreferenciada*

Estas variables fueron altamente significativas en la regresión logística y relevantes en los modelos de árboles.

1.2. Variables derivadas de la serie de consumo (dinámicas):

Son las variables que permiten capturar patrones sospechosos. El estudio evidenció que las más predictivas fueron:

Variables estadísticas:

- Consumo mínimo, máximo, media y mediana.
- Desviación estándar del consumo mensual.
- N° de atípicos moderados y N° atípicos extremos.
- Curtosis y asimetría.
- Secuencias de ceros.
- Delta brusco.

Variables de complejidad y patrones:

- Entropía SAX.
- Cambios de símbolo.
- Lempel–Ziv Complexity (LZC) 9 bins.
- Time Series Length Factor (TSLF).

2. METODOLOGÍA PROPUESTA PARA EL MODELO

La empresa puede seguir este procedimiento estándar para refrescar el modelo y mantener un sistema operativo de detección de fraude:

Paso 1: Actualización de datos y preprocesamiento

- Obtener la base de consumos mensuales de los últimos 12–24 meses para cada cliente o la periodicidad definida por la compañía.
- Actualizar la información del medidor (cambios, reemplazos, reclasificaciones).
- Depurar datos: Eliminar consumos negativos, imputar faltantes según política definida (o descartar clientes con trazas insuficientes).
- Georreferenciar direcciones nuevas o modificadas.

Paso 2: Extracción automática de características

A partir de las series mensuales, generar todas las variables estadísticas, derivaciones de las series de consumo, transformaciones (SAX, LZC, TSLF, etc.) mencionadas en este trabajo.

Paso 3: Entrenamiento y selección de modelos

De acuerdo con la evidencia empírica del estudio:

- **Modelo principal recomendado:** Random Forest con enfoque de negocio | (Mejor equilibrio entre desempeño técnico y utilidad económica) | Ganancia promedio: 2,31 M aprox., mROI 64,3 % aprox.
- **Modelo secundario:** Regresión logística como modelo interpretable base, para validación, auditoría o escenarios donde se necesite transparencia regulatoria.
- **Modelos no recomendados como base operativa:** DBSCAN, debido a sensibilidades cercanas a cero para fraude y poca separación de clases.

Paso 4: Enfoque de negocio y priorización de inspecciones

Con base en el modelo calibrado:

- Obtener puntajes de probabilidad de fraude para cada cliente.
- Ordenar descendentemente.
- Aplicar el límite operativo (por ejemplo, número de inspecciones disponibles en el mes).
- Evaluar ganancia mediante la matriz de costos real del cliente (tal como se definió en el trabajo).

Este paso permite que el algoritmo deje de ser un ejercicio técnico y pase a generar ahorro operativo real.

Paso 5: Validación continua y retroalimentación
Cada periodo definido por la empresa (por ejemplo, trimestral):

- Reentrenar el modelo con etiquetas nuevas provenientes de inspecciones.
- Evaluar variaciones del AUC-ROC, AUC-PR, *F1-Score*, *Brier Score* y calibración.
- Revisar si hay dependencia espacial significativa (prueba de Moran's I), como en el estudio.
- Detectar cambios estructurales (por ejemplo, reemplazos masivos de medidores).
- Si las métricas se deterioran, actualizar políticas de generación de variables.

8 RESULTADOS: MODELOS ANOMALÍAS

8.1. MODELACIÓN SUPERVISADA ANOMALÍAS

8.1.1. Modelación Random Forest vs. XGBoost sin enfoque de negocio (anomalías)

- **Random Forest:**

Random Forest: Búsqueda de hiperparámetros sin enfoque de negocio (anomalías)
De acuerdo con la Tabla 8.1, los resultados del *Grid Search* para el modelo *Random Forest* aplicado a la detección de anomalías mostraron valores nulos ($F1 = 0{,}00$) en todos los escenarios evaluados. Este resultado no obedeció a un error de implementación, sino a la escasa cantidad de datos disponibles para el entrenamiento, lo que impidió que el modelo aprendiera patrones representativos de la clase minoritaria. En consecuencia, el algoritmo clasificó todas las observaciones como pertenecientes a la clase mayoritaria, generando métricas nulas de precisión y exhaustividad. Este comportamiento era esperable dada la limitada muestra de anomalías, y resalta la importancia de contar con un volumen suficiente y representativo de datos para que los modelos supervisados puedan generalizar adecuadamente.

Tabla 8.1: Resultados del Grid Search *Random Forest* (anomalías) sin enfoque de negocio.
Fuente: Elaboración propia.

n estimators	Max Depth	Min Split	Min Leaf	Max Features	F1 Prom.	F1 Desv. Est.	Ranking
150	10	5	2	sqrt	0,00	0,00	1
250	10	5	2	sqrt	0,00	0,00	1
150	10	10	2	sqrt	0,00	0,00	1
250	10	10	2	sqrt	0,00	0,00	1
150	10	5	4	sqrt	0,00	0,00	1
250	10	5	4	sqrt	0,00	0,00	1
150	10	10	4	sqrt	0,00	0,00	1
250	10	10	4	sqrt	0,00	0,00	1
150	20	5	2	sqrt	0,00	0,00	1
250	20	5	2	sqrt	0,00	0,00	1
150	20	10	2	sqrt	0,00	0,00	1
250	20	10	2	sqrt	0,00	0,00	1

Continuación de la Tabla 8.1

n estimators	Max Depth	Min Split	Min Leaf	Max Features	F1 Prom.	F1 Desv. Est.	Ranking
150	20	5	4	sqrt	0,00	0,00	1
250	20	5	4	sqrt	0,00	0,00	1
150	20	10	4	sqrt	0,00	0,00	1
250	20	10	4	sqrt	0,00	0,00	1
150	Ninguno	5	2	sqrt	0,00	0,00	1
250	Ninguno	5	2	sqrt	0,00	0,00	4
150	Ninguno	10	2	sqrt	0,00	0,00	4
250	Ninguno	10	2	sqrt	0,00	0,00	4
150	Ninguno	5	4	sqrt	0,00	0,00	4
250	Ninguno	5	4	sqrt	0,00	0,00	4
150	Ninguno	10	4	sqrt	0,00	0,00	4
250	Ninguno	10	4	sqrt	0,00	0,00	4

Random Forest: Métricas en cada fold sin enfoque de negocio (anomalías)

Los resultados de la Tabla 8.2 mostraron una estabilidad general en las métricas entre los diferentes *folds*, con valores de *F1-Score* que oscilaron entre 0,50 y 0,67. El calibrador *Platt* fue el más utilizado y alcanzó el mejor desempeño en el quinto *fold*, donde se obtuvo un equilibrio más alto entre *Precision* (0,75) y *Recall* (0,60).

En la mayoría de los casos, el modelo priorizó la sensibilidad, manteniendo un *Recall* igual a 1 en cuatro de los cinco *folds*, lo que indicó una tendencia a detectar la mayoría de los casos positivos, aunque con una ligera pérdida en la precisión. En conjunto, el comportamiento del *Random Forest* sin enfoque de negocio fue consistente, pero con un margen de mejora en la calibración de los umbrales para optimizar el balance entre ambas métricas.

Tabla 8.2: *Random Forest* (anomalías) sin enfoque de negocio - Métricas detalladas por fold. Fuente: Elaboración propia.

Fold	Calibrador	Brier Score	Mejor umbral	F1 Score	Precision	Recall
1	Platt	0,173	0,242	0,588	0,416	1
2	Isotónico	0,161	0,384	0,555	0,384	1
3	Platt	0,177	0,189	0,555	0,384	1
4	Platt	0,195	0,222	0,50	0,333	1
5	Platt	0,151	0,462	0,666	0,750	0,6

Random Forest: Métricas promedio y desviación sin enfoque de negocio (anomalías)

De acuerdo con la Tabla 8.3, el desempeño promedio del modelo *Random Forest* sin enfoque de negocio mostró un equilibrio moderado entre precisión y sensibilidad. El valor medio del *F1-Score* fue de 0.573, con una desviación de 0,06, lo que indicó una variabilidad controlada en los resultados de los distintos *folds*. El *Recall* promedio alcanzó 0,92, evidenciando que el modelo detectó la mayoría de los casos positivos, aunque con una precisión promedio más baja (0,453), lo que implicó una mayor proporción de falsos positivos.

El *Brier-Score* (0,171) y la desviación asociada (0,016) reflejaron una calibración razonablemente estable, mientras que el mejor umbral de decisión (0,30) presentó cierta dispersión entre los pliegues. En conjunto, el modelo mostró un buen nivel de sensibilidad, pero aún con margen para mejorar el equilibrio entre precisión y calibración probabilística.

Tabla 8.3: *Random Forest* (anomalías) sin enfoque de negocio - Promedio y desviación estándar de las métricas. Fuente: Elaboración propia.

Métrica	Promedio	Desviación
Brier-Score	0,171	0,016
Mejor umbral	0,300	0,117
F1-Score	0,573	0,061
Precision	0,453	0,168
Recall	0,92	0,178

- **XGBoost:**

XGBoost: Búsqueda de hiperparámetros sin enfoque de negocio (anomalías)

De acuerdo con la Tabla 8.4, el proceso de búsqueda de hiperparámetros del modelo *XGBoost* sin enfoque de negocio evidenció un desempeño general bajo en la métrica *F1-Score*. Las configuraciones con mejor rendimiento correspondieron a profundidades de árbol reducidas (*Max Depth* = 3–5) y tasas de aprendizaje pequeñas (*Learning Rate* = 0,05), alcanzando un valor máximo de *F1-Score* de 0,146 con una desviación estándar de 0,121. Estas combinaciones mostraron una ligera mejora respecto a las demás, aunque el valor absoluto del desempeño continuó siendo bajo, lo que indica una capacidad limitada del modelo para distinguir entre clases en escenarios con pocos registros de anomalías.

El resto de configuraciones presentó valores de *F1-Score* inferiores a 0,09, lo que sugiere que los ajustes en la cantidad de estimadores o en los parámetros de muestreo no generaron mejoras significativas. En consecuencia, se priorizó el equilibrio entre complejidad y estabilidad del modelo, manteniendo configuraciones conservadoras en profundidad y tasa de aprendizaje.

Tabla 8.4: Resultados del Grid Search *XGBoost* (anomalías) sin enfoque de negocio. Fuente: Elaboración propia.

n estimators	Max Depth	Learning Rate	Subsample	Colsample ByTree	F1 Prom.	F1 Desv. Est.	Ranking
250	3	0,05	0,80	0,80	0,146	0,121	1
150	3	0,05	0,80	0,80	0,138	0,113	2
150	3	0,10	0,80	0,80	0,134	0,11	3
250	7	0,10	0,80	0,80	0,130	0,107	4
250	7	0,05	0,80	0,80	0,101	0,126	5
250	5	0,05	0,80	0,80	0,101	0,126	5
250	3	0,10	0,80	0,80	0,090	0,111	7
150	5	0,10	0,80	0,80	0,084	0,103	8
250	5	0,10	0,80	0,80	0,084	0,103	8
150	7	0,10	0,80	0,80	0,084	0,103	8
150	7	0,05	0,80	0,80	0,040	0,080	11
150	5	0,05	0,80	0,80	0,040	0,080	11

XGBoost: Métricas en cada fold sin enfoque de negocio (anomalías)
De acuerdo con la Tabla 8.5, el desempeño del modelo *XGBoost* sin enfoque de negocio mostró resultados consistentes entre los diferentes *folds*, con valores de *F1-Score* comprendidos entre 0,50 y 0,533. La calibración isotónica fue la más frecuente y presentó un comportamiento estable en términos de precisión y sensibilidad. En particular, el modelo logró mantener un *Recall* elevado en la mayoría de los pliegues, lo que evidenció su capacidad para identificar correctamente los casos positivos, aunque con una precisión moderada que osciló entre 0,333 y 0,667.

El *Brier-Score* promedio se mantuvo alrededor de 0,18, reflejando una calibración razonablemente adecuada de las probabilidades predichas. No obstante, los resultados indicaron que, pese a la estabilidad del modelo, su capacidad discriminativa fue limitada, es decir, el *XGBoost* requirió un ajuste más profundo o la incorporación de variables adicionales para mejorar el equilibrio entre precisión y sensibilidad en la detección de anomalías.

Tabla 8.5: *XGBoost* (anomalías) sin enfoque de negocio - Métricas detalladas por fold.
Fuente: Elaboración propia.

Fold	Calibrador	Brier-Score	Mejor umbral	F1-Score	Precisión	Recall
1	Isotónico	0,180	0,364	0,500	0,364	0,80
2	Isotónico	0,158	0,250	0,455	0,294	1
3	Isotónico	0,188	0,250	0,476	0,313	1
4	Isotónico	0,162	0,417	0,588	0,417	1
5	Isotónico	0,185	0,333	0,500	0,333	1

XGBoost: Métricas promedio y desviación sin enfoque de negocio (anomalías)
En relación con la Tabla 8.6, el modelo *XGBoost* sin enfoque de negocio presentó un desempeño moderado, con resultados estables en las métricas globales. El valor promedio del *F1-Score* fue de 0,504, con una desviación estándar de 0,075, lo cual evidenció un equilibrio razonable entre precisión y sensibilidad a lo largo de los diferentes *folds*. El *Recall* promedio alcanzó 0,96, mostrando que el modelo fue capaz de identificar una alta proporción de casos positivos, aunque con una *Precision* promedio de 0,344, reflejando una tendencia a clasificar algunos falsos positivos.

El *Brier-Score*, con un valor promedio de 0,174 y una desviación estándar de 0,013, indicó una calibración adecuada de las probabilidades, manteniendo consistencia entre las predicciones generadas y los valores observados. Sin embargo, el *mejor umbral de decisión* ($0{,}322 \pm 0{,}072$) mostró cierta variabilidad entre los diferentes pliegues, lo que sugiere una sensibilidad del modelo frente a los ajustes de corte en la clasificación. En conjunto, los resultados evidenciaron un comportamiento coherente, con una buena capacidad de detección de anomalías, aunque con un margen de mejora en la precisión de los positivos predichos.

Tabla 8.6: *XGBoost* (anomalías) sin enfoque de negocio - Promedio y desviación estándar de las métricas.
Fuente: Elaboración propia.

Métrica	Promedio	Desviación
Brier-Score	0,174	0,013

Continuación de la Tabla 8.6

Métrica	Promedio	Desviación
Mejor umbral	0,322	0,072
F1-Score	0,503	0,05
Precision	0,344	0,048
Recall	0,96	0,089

- **Comparación y discusión Random Forest vs. XGBoost sin enfoque de negocio (anomalías):**

Tabla 8.7: Comparación de métricas entre *Random Forest* (RF) y *XGBoost* (XGB) sin enfoque de negocio (anomalías). Fuente: Elaboración propia.

Métrica	RF Promedio	RF Desv. Est.	XGB Promedio	XGB Desv. Est.
Brier-Score	0,171	0,016	0,174	0,013
F1-Score	0,573	0,061	0,503	0,005
Precision	0,453	0,168	0,344	0,04
Recall	0,92	0,178	0,96	0,08

Los resultados de la Tabla 8.7 mostraron que en el escenario sin enfoque de negocio, el desempeño de los modelos *Random Forest* (RF) y *XGBoost* (XGB) presentó diferencias relevantes en cuanto a la calibración, la precisión y la sensibilidad en la detección de anomalías. El *Random Forest* obtuvo un *Brier-Score* promedio de 0,171 con una desviación estándar de 0,016, mientras que el *XGBoost* registró un valor muy similar de 0,174 con una desviación de 0,013. Estos resultados indicaron que ambos modelos mantuvieron una adecuada coherencia entre las probabilidades predichas y los valores observados, aunque con una ligera ventaja para XGB en términos de calibración.

El *F1-Score*, criterio principal de comparación, fue superior en el modelo RF (0,573) frente al XGB (0,503), lo que evidenció un mejor equilibrio entre la identificación de los casos positivos y la reducción de falsos positivos. De manera coherente, el RF alcanzó también una mayor precisión promedio (0,453) en comparación con XGB (0,344), reflejando una mayor confiabilidad en las predicciones positivas.

En contraste, el *Recall* fue marginalmente superior en el modelo XGB (0,96) frente al RF (0,92), lo que mostró una leve ventaja del XGB en la capacidad de detección de casos positivos. Sin embargo, esta ganancia en sensibilidad no compensó la disminución en precisión y F1-Score.

Dado que el criterio principal fue el *F1-Score*, el modelo seleccionado como ganador fue el *Random Forest*, el cual, con calibración y umbral óptimos, presentó un desempeño más equilibrado y robusto para la detección de anomalías sin enfoque de negocio.

Conjunto de entrenamiento: Evaluación robusta (Repeated Stratified K-Fold)

Tabla 8.8: Matriz de confusión del modelo *Random Forest* (anomalías) en conjunto de entrenamiento. Fuente: Elaboración propia.

		Predicha	
		No anomalía (0)	Anomalía (1)
Real	No anomalía (0)	61	6
	Anomalía (1)	0	25

Tabla 8.9: Reporte de clasificación del modelo *Random Forest* (anomalías) en conjunto de entrenamiento. Fuente: Elaboración propia.

Clase	**Precision**	**Recall**	**F1-Score**	**Soporte**
0	1	0,91	0,95	67
1	0,81	1	0,89	25
Accuracy			0,93	92
Macro avg	0,90	0,96	0,92	92
Weighted avg	0,95	0,93	0,94	92

Los resultados obtenidos en la Tabla 8.8 y Tabla 8.9 evidenciaron un buen desempeño del modelo *Random Forest* durante la fase de entrenamiento. En términos generales, el modelo mostró una adecuada capacidad para distinguir entre las clases, alcanzando altos niveles de precisión y sensibilidad. La matriz de confusión reflejó una correcta clasificación de la mayoría de los casos, con un número reducido de errores, mientras que el reporte de clasificación confirmó un equilibrio favorable entre las métricas de *Precision*, *Recall* y *F1-Score*.

De manera global, el modelo demostró un comportamiento estable y consistente, lo que indica que logró aprender los patrones principales del conjunto de entrenamiento sin evidenciar sobreajuste. Estos resultados permiten concluir que el *Random Forest* ofreció un rendimiento sólido y confiable dentro del contexto de la detección de anomalías.

Conjunto de prueba: Evaluación robusta (Repeated Stratified K-Fold)

Tabla 8.10: Matriz de confusión final del modelo *Random Forest* (anomalías) en conjunto de prueba. Fuente: Elaboración propia.

		Predicha	
		No anomalía (0)	Anomalía (1)
Real	No anomalía (0)	6	11
	Anomalía (1)	2	5

Tabla 8.11: Reporte de clasificación final del modelo *Random Forest* (anomalías) en conjunto de prueba. Fuente: Elaboración propia.

Clase	Precision	Recall	F1-Score	Soporte
0	0,75	0,35	0,48	17
1	0,31	0,71	0,43	7
Accuracy			0,46	24
Macro avg	0,53	0,53	0,46	24
Weighted avg	0,62	0,46	0,47	24

Los resultados del conjunto de prueba de la Tabla 8.10 y Tabla 8.11 evidenciaron un desempeño moderado del modelo *Random Forest* en la detección de anomalías. La matriz de confusión mostró que el modelo logró identificar correctamente cinco de los siete casos positivos (*Anomalía*), alcanzando un *Recall* de 0,71 en la clase minoritaria. Sin embargo, la precisión en esta clase fue baja (0,31), lo que indicó una proporción considerable de falsos positivos.

En la clase mayoritaria (*No anomalía*), el modelo presentó una alta precisión (0,75) pero un *Recall* limitado (0,35), indicando que una parte importante de los casos no fraudulentos fueron clasificados erróneamente como fraudes.

El valor global de *Accuracy* (0,46) reflejó un equilibrio inestable entre ambas clases, condicionado por el reducido tamaño del conjunto de prueba. Aun así, el modelo logró un *F1-Score* promedio de 0,46, evidenciando que, aunque existe cierta capacidad de discriminación, el desempeño se ve afectado por la limitada cantidad de observaciones y el desbalance entre clases. En síntesis, el modelo mostró potencial para detectar casos positivos, pero con sacrificio en la precisión general.

Importancia de características del modelo con mejor desempeño

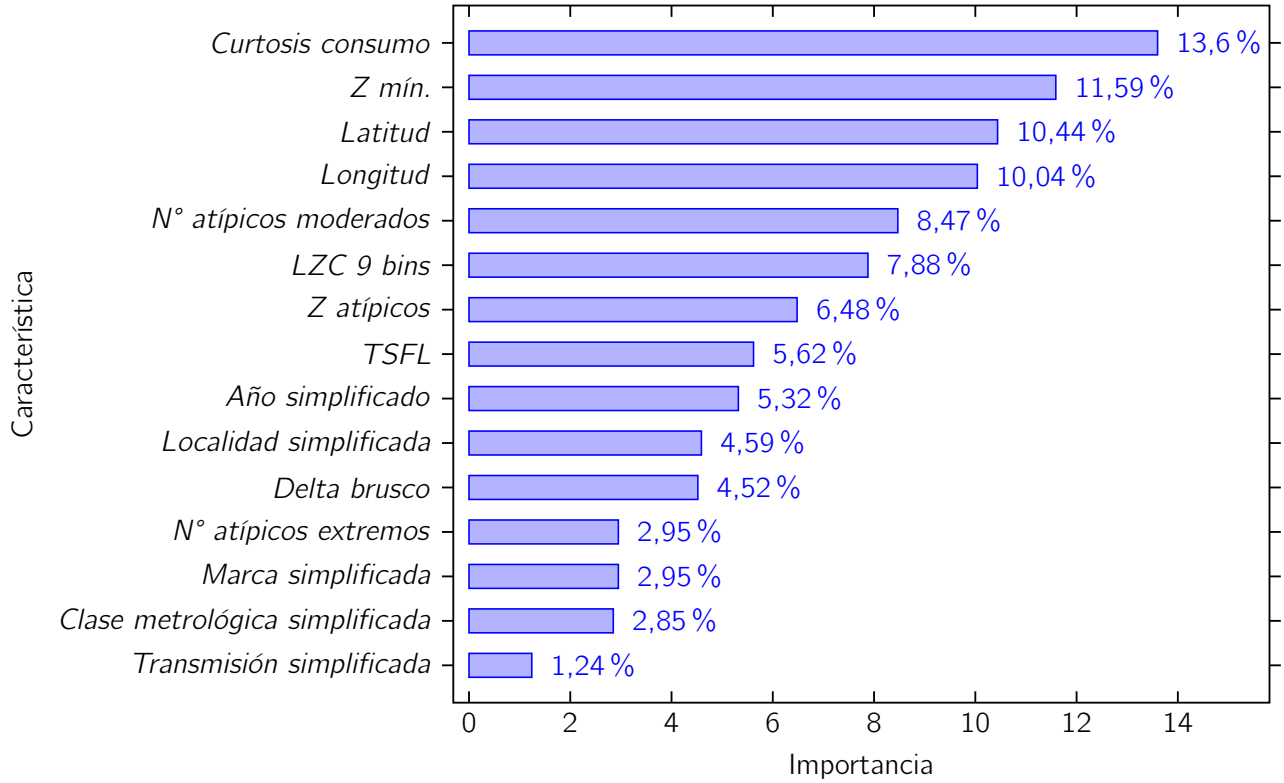

Figura 8.1: Top 15 características más importantes (%) del modelo *Random Forest* (anomalías) sin enfoque de negocio. Fuente: Elaboración propia.

Los resultados de la Figura 8.1 evidenciaron que el modelo *Random Forest* identificó un conjunto de variables con un peso considerable en la predicción de anomalías, principalmente aquellas asociadas a la distribución y variabilidad del consumo. Las características *Curtosis consumo* (13,6 %) y *Z mín.* (11,59 %) fueron las más influyentes, lo que indicó que la forma de la distribución del consumo y los valores atípicos bajos jugaron un papel clave en la diferenciación de comportamientos anómalos.

Asimismo, las variables *Latitud* (10,44 %) y *Longitud* (10,04 %) presentaron una contribución importante, lo que sugiere la existencia de patrones espaciales en la ocurrencia de anomalías. Las medidas de complejidad temporal, como *LZC 9 bins* (7,88 %) y *TSFL* (5,62 %), también aportaron información relevante sobre la dinámica de los consumos, complementando los indicadores estadísticos tradicionales.

Por otro lado, las variables categóricas, como *Marca simplificada*, *Clase metrológica simplificada* y *Transmisión simplificada*, mostraron una menor influencia individual (entre 1 % y 3 %), lo que indicó que su contribución fue marginal frente a las características cuantitativas. En conjunto, los resultados reflejaron que el modelo priorizó los patrones de comportamiento y dispersión del consumo sobre las características del medidor o su ubicación administrativa.

8.2. REGRESIÓN LOGÍSTICA ANOMALÍAS

En esta sección se emplea la nomenclatura de M1 a M5 para identificar los diferentes modelos para la detección de anomalías. La Tabla 8.12 detalla las variables incluidas en cada uno de ellos.

Tabla 8.12: Variables incluidas en la regresión logística de anomalías (M1–M5). Fuente: Elaboración propia.

Variable	M1	M2	M3	M4	M5
SD consumo $^{(t)}$	✓				
Mediana consumo $^{(t)}$	✓	✓			
Mín.consumo $^{(t)}$	✓				
Asimetría consumo $^{(t)}$	✓	✓			
N° atípicos moderados $^{(t)}$	✓				
Delta brusco $^{(t)}$	✓				✓
Entropía SAX $^{(t)}$	✓				✓
Cambios de símbolo $^{(t)}$	✓	✓	✓	✓	
LZC 9 bins $^{(t)}$	✓	✓	✓		✓
LZC 99 bins $^{(t)}$	✓	✓	✓	✓	
TSFL $^{(t)}$	✓	✓			
Clase metrológica simplificada	✓	✓	✓	✓	✓

Nota: Las variables marcadas con (t) hacen referencia a la variable transformada y estandarizada que se aplica a la regresión logística.

8.2.1. Significancia y ajustes del modelo

Tabla 8.13: Medidas de influencia de los modelos de regresión logística de anomalías (M1–M5). Fuente: Elaboración propia.

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M1	37	0,114	0,291	0,715
M1	5	0,108	0,323	0,501
M1	19	0,0678	0,200	0,212
M1	89	0,0614	0,506	5,29
M1	43	0,0566	0,234	0,259
M1	21	0,0543	0,253	1,00
M1	73	0,0535	0,350	3,53
M1	38	0,0511	0,124	4,26
M1	86	0,0483	0,169	2,92
M1	64	0,0465	0,137	0,706
M1	42	0,0458	0,233	0,998
M1	25	0,0423	0,133	0,284
M1	67	0,0412	0,108	0,123
M1	75	0,0381	0,301	0,361
M1	47	0,0323	0,0680	0,707

Continuación de la Tabla 8.13

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M2	89	0,0884	0,267	0,160
M2	73	0,0758	0,291	0,241
M2	21	0,0738	0,208	0,263
M2	43	0,0678	0,205	0,182
M2	81	0,0422	0,203	0,209
M2	8	0,0420	0,262	0,204
M2	19	0,0367	0,130	0,242
M2	47	0,0342	0,0436	0,136
M2	22	0,0334	0,0505	0,158
M2	1	0,0333	0,196	0,190
M2	42	0,0330	0,127	0,187
M2	67	0,0294	0,0884	0,116
M2	88	0,0277	0,0566	0,164
M2	25	0,0269	0,0875	0,169
M2	62	0,0265	0,0931	0,162
M3	89	0,1310	0,231	0,135
M3	43	0,0972	0,172	0,207
M3	73	0,0728	0,223	0,122
M3	8	0,0603	0,232	0,205
M3	81	0,0507	0,112	0,151
M3	1	0,0486	0,163	0,176
M3	22	0,0439	0,0342	0,120
M3	21	0,0347	0,118	0,109
M3	67	0,0316	0,0708	0,137
M3	19	0,0298	0,0869	0,175
M3	47	0,0266	0,0297	0,0653
M3	6	0,0239	0,0332	0,0577
M3	38	0,0226	0,0682	0,125
M3	42	0,0219	0,0450	0,0955
M3	88	0,0209	0,0330	0,0605
M4	89	0,1710	0,230	0,136
M4	73	0,0860	0,151	0,156
M4	81	0,0506	0,107	0,112
M4	21	0,0458	0,106	0,107
M4	1	0,0424	0,149	0,161
M4	22	0,0380	0,0307	0,0885
M4	19	0,0289	0,0813	0,160
M4	6	0,0259	0,0325	0,0640
M4	67	0,0253	0,0333	0,0695
M4	88	0,0252	0,0331	0,0692
M4	71	0,0249	0,0738	0,154
M4	70	0,0231	0,114	0,0891
M4	44	0,0215	0,0202	0,0769

Continuación de la Tabla 8.13

Modelo	Index	Cook's D	Hat	MaxAbsDFBeta
M4	66	0,0212	0,0274	0,0510
M4	27	0,0211	0,0694	0,143
M5	38	0,0725	0,112	3,43
M5	67	0,0684	0,0539	0,166
M5	43	0,0559	0,150	1,39
M5	37	0,0412	0,0842	0,825
M5	92	0,0333	0,286	3,35
M5	35	0,0317	0,0732	0,414
M5	86	0,0297	0,0552	1,33
M5	25	0,0295	0,0284	0,0710
M5	22	0,0281	0,0422	1,79
M5	1	0,0274	0,0908	0,413
M5	73	0,0273	0,178	1,31
M5	24	0,0251	0,0682	0,141
M5	89	0,0237	0,157	0,531
M5	47	0,0205	0,0312	0,157
M5	88	0,0187	0,0214	0,0652

8.2.2. GVIF ajustado

Tabla 8.14: GVIF ajustado para los M1 a M5 de anomalías.
Fuente: Elaboración propia.

Variable	M1	M2	M3	M4	M5
SD consumo $^{(t)}$	1,42				
Mediana consumo $^{(t)}$	1,98	1,89			
Mín.consumo $^{(t)}$	1,28				
Asimetría consumo $^{(t)}$	2,09	1,98			
N° atípicos moderados $^{(t)}$	1,26				
Delta brusco $^{(t)}$	1,23				1,05
Entropía SAX $^{(t)}$	1,47				1,05
Cambios de símbolo $^{(t)}$	1,31	1,50	1,21	1,20	
LZC 9 bins $^{(t)}$	1,78	1,56	1,37		1,13
LZC 99 bins $^{(t)}$	1,83	1,76	1,56	1,22	
TSFL $^{(t)}$	2,11	2,02			
Clase metrológica simplificada	1,16	1,08	1,02	1,02	1,02

Nota: Las variables marcadas con (t) hacen referencia a la variable transformada y estandarizada que se aplica a la regresión logística.

La Tabla 8.14 envidenció que en todos los casos los valores se mantuvieron por debajo del umbral de referencia igual a 5, por lo que no se identificaron problemas relevantes de colinealidad. Las variables asociadas al consumo, como la desviación estándar, la mediana, el mínimo, la asimetría y la

entropía SAX, registraron GVIF ajustados en un rango entre 1,2 y 2,1, lo que correspondió a niveles moderados y aceptables de relación con otros predictores.

Las variables categóricas, en particular la clase metrológica simplificada, presentaron valores próximos a 1, lo que indicó baja redundancia y aporte informativo complementario. En conjunto, los resultados respaldaron la idoneidad del conjunto de variables y de los procedimientos de preprocesamiento para la estimación de los modelos de regresión logística, favoreciendo su estabilidad e interpretabilidad.

8.2.3. Métricas de ajuste (prueba de Hosmer-Lemeshow)

- **Hipótesis nula** (H_0)**:** El modelo se ajusta bien a los datos. Es decir, no hay diferencia significativa entre los valores observados y los esperados, y el modelo tiene una buena calibración.
- **Hipótesis alternativa** (H_1)**:** El modelo no se ajusta bien a los datos. Es decir, existe una diferencia significativa entre los valores observados y los esperados, y el modelo no tiene una buena calibración.

Tabla 8.15: Indicadores de ajuste y prueba de bondad para los modelos de anomalías. Fuente: Elaboración propia.

Modelo	llh	llhNull	McFadden	r_{ML}	r_{CU}	p-value	Decisión (H_0)
M1	-49,20	-55,40	0,113	0,125	0,180	0,784	No rechazar H_0
M2	-51,10	-55,40	0,078	0,088	0,127	0,400	No rechazar H_0
M3	-52,10	-55,40	0,060	0,068	0,099	0,173	No rechazar H_0
M4	-52,60	-55,40	0,051	0,058	0,084	0,349	No rechazar H_0
M5	-50,50	-55,40	0,088	0,099	0,143	0,088	No rechazar H_0

Los indicadores de ajuste de la Tabla 8.15 mostraron que todos los modelos presentaron valores de log-verosimilitud negativos, coherentes con la naturaleza de la estimación por máxima verosimilitud. El coeficiente de McFadden osciló entre 0,05 y 0,11, lo que correspondió a un nivel de ajuste aceptable para modelos de regresión logística aplicados a fenómenos con clases desbalanceadas. Los coeficientes r_{ML} y r_{CU} mantuvieron valores bajos, lo que indicó que la proporción de varianza explicada por los predictores fue limitada, aunque suficiente para capturar patrones relevantes en los datos.

El valor p de la prueba de Hosmer–Lemeshow fue superior a 0,05 en todos los casos, por lo que no se rechazó la hipótesis nula. Esto implicó que no existieron diferencias significativas entre los valores observados y los esperados, y por tanto, los modelos presentaron una *calibración adecuada*.

8.2.4. Métricas de evaluación

Tabla 8.16: Métricas de desempeño por modelo de anomalías en validación cruzada y prueba. Fuente: Elaboración propia.

Métrica	M1	M2	M3	M4	M5
F2 CV entrenamiento	0,64	0,65	0,66	0,66	0,65
Spec. CV entrenamiento	No aplica	No aplica	No aplica	No aplica	No aplica
Recall CV entrenamiento	No aplica	No aplica	No aplica	No aplica	No aplica

Continuación de la Tabla 8.16

Métrica	M1	M2	M3	M4	M5
Prec. CV entrenamiento	No aplica	No aplica	No aplica	No aplica	No aplica
Threshold CV	0,02	0,03	0,07	0,10	0,08
F2 prueba	0,55	0,65	0,65	0,65	0,55
F1 prueba	0,37	No aplica	No aplica	No aplica	0,37
Recall	0,83	No aplica	No aplica	No aplica	0,83
Precision	0,23	No aplica	No aplica	No aplica	0,23
Specificity	0,00	No aplica	No aplica	No aplica	0,00
Balanced Accuracy	0,41	No aplica	No aplica	No aplica	0,41
AUC-ROC	0,75	0,50	0,51	0,40	0,77
AUC-PR	0,20	0,25	0,39	0,21	0,20
Brier-Score	0,30	0,23	0,21	0,22	0,27
LogLoss	2,30	0,66	0,62	0,65	2,24

De acuerdo con la Tabla 8.16,las métricas de desempeño obtenidas para los modelos de anomalías evidenciaron comportamientos consistentes entre las distintas configuraciones evaluadas. Durante la validación cruzada, los valores del estadístico *F2* oscilaron entre 0,64 y 0,66, lo que reflejó un rendimiento estable del proceso de entrenamiento. El umbral de decisión se mantuvo en valores bajos, entre 0,02 y 0,10, coherente con la prevalencia reducida de casos positivos en el conjunto de datos.

En la fase de prueba, el M3 y el M4 alcanzaron los valores más altos de *F2* (0,65), lo que indicó una adecuada capacidad para equilibrar la sensibilidad y la precisión. Sin embargo, el M1 se destacó por presentar un *Recall* superior (0,83), aún cuando su precisión fue más baja (0,23), lo que indicó una mayor cobertura de casos anómalos a costa de un mayor número de falsos positivos.

El *AUC-ROC* se mantuvo en niveles moderados, entre 0,40 y 0,75, lo que evidenció una capacidad discriminativa aceptable considerando el tamaño de la muestra. De forma complementaria, los valores del *Brier-Score* y el *LogLoss* indicaron un ajuste probabilístico adecuado, aunque con cierto margen de mejora en la calibración de las probabilidades. En conjunto, los resultados mostraron un desempeño estable entre los modelos, con un compromiso razonable entre sensibilidad y precisión en la identificación de anomalías.

Tabla 8.17: Indicadores de desempeño y calibración por modelo de anomalías en los conjuntos de entrenamiento y prueba. Fuente: Elaboración propia.

Métrica	M1	M2	M3	M4	M5
Entrenamiento					
AUC-ROC	0,73	0,72	0,70	0,69	0,70
AUC-PR	0,44	0,42	0,41	0,43	0,46
Brier	0,17	0,18	0,18	0,18	0,17
LogLoss	0,52	0,54	0,55	0,55	0,51
Hosmer–Lemeshow	0,78	0,39	0,17	0,34	0,08
Intercepto	0	0	0	0	0
Pendiente	1	1	1	1	1

Continuación de la Tabla 8.17

Métrica	M1	M2	M3	M4	M5
Prueba					
AUC-ROC	0,75	0,50	0,51	0,40	0,77
AUC-PR	0,20	0,25	0,39	0,21	0,18
Brier	0,30	0,23	0,21	0,22	0,29
LogLoss	1,36	0,66	0,62	0,65	1,36
Hosmer–Lemeshow	0,00	0,35	0,44	0,31	0,00
Intercepto	-0,44	-0,52	-0,34	-0,34	-0,23
Pendiente	-1,19	0,00	0,18	-0,08	-1,58

De acuerdo con la Tabla 8.17 se evidenció que en entrenamiento, los valores de *AUC-ROC* se mantuvieron entre 0,69 y 0,73, lo que reflejó una capacidad moderada de discriminación entre observaciones normales y anómalas. De manera complementaria, los valores de *AUC-PR* oscilaron entre 0,41 y 0,46, coherentes con la baja prevalencia de casos positivos en la base de datos. Las métricas de *Brier Score* y *LogLoss* mostraron valores bajos y consistentes, lo que indicó un adecuado ajuste probabilístico y ausencia de sobreajuste durante el entrenamiento.

En la etapa de prueba, el Modelo 5 alcanzó el mayor *AUC-ROC* (0,77), evidenciando una mejor capacidad de discriminación respecto a los demás. Sin embargo, el incremento en el *LogLoss* y los valores del estadístico de Hosmer–Lemeshow cercanos a cero señalaron una pérdida de calibración, especialmente en los modelos con pendiente negativa. Esto indicó que las probabilidades predichas tendieron a subestimar la ocurrencia de anomalías en los deciles superiores. En conjunto, los resultados mostraron que, aunque todos los modelos conservaron un nivel aceptable de discriminación, su calibración fue sensible a las variaciones del conjunto de prueba, siendo el Modelo 5 el más equilibrado en términos de ajuste y capacidad predictiva.

8.2.5. Modelo seleccionado

Los resultados de la evaluación de supuestos mostraron que ninguno de los modelos propuestos cumplió con el criterio de linealidad en el logit, condición fundamental para la validez de la regresión logística. Esta falta de linealidad implica que la relación entre los predictores continuos y la probabilidad del evento no puede representarse adecuadamente mediante una función logit, afectando la interpretación y la estabilidad de los coeficientes estimados. En consecuencia, aunque algunos modelos presentaron métricas aceptables de ajuste global y calibración, estos resultados no pueden considerarse confiables, ya que se obtuvieron sobre una estructura que viola un supuesto básico del modelo. Por esta razón, no fue posible seleccionar ningún modelo como válido, dado que hacerlo supondría aceptar estimaciones distorsionadas y conclusiones que no reflejan el comportamiento real de los datos.

8.2.6. Prueba de Moran's I y dependencia especial

Dado que ninguno de los modelos evaluados cumplió con el supuesto de linealidad en el logit, no fue posible seleccionar un modelo final válido para el análisis. En consecuencia, no se realizó la prueba de dependencia espacial (Moran's I), ya que su aplicación requiere la existencia de residuos provenientes de un modelo estadísticamente consistente. Realizar dicha prueba bajo estas condiciones habría implicado

evaluar la autocorrelación espacial de estimaciones inválidas, lo que no resultaría metodológicamente adecuado.

8.2.7. Comparación de AIC y BIC

Tabla 8.18: Comparación de modelos de anomalías según criterios AIC y BIC.
Fuente: Elaboración propia.

Modelo	AIC	BIC
M1	124,32	157,39
M2	118,20	138,54
M3	114,20	126,91
M4	113,23	123,40
M5	111,09	123,81

La comparación de los valores de AIC y BIC (Tabla 8.18) evidenció que el modelo M5 presentó los valores más bajos en ambos criterios, lo que indicó un mejor ajuste relativo. Sin embargo, dado que ninguno de los modelos cumplió con el supuesto de linealidad en el logit, estos resultados se reportan únicamente con fines descriptivos y no como criterio de selección de modelo.

8.2.8. Análisis de la variación en dependencia espacial con k vecinos cercanos

El análisis de la variación en la dependencia espacial con k vecinos cercanos se basa en la evaluación de la autocorrelación de los residuos del modelo. Dado que este procedimiento sólo es pertinente cuando los residuos provienen de un modelo estadísticamente consistente, no se realizó su estimación en esta etapa. Ejecutarlo bajo condiciones de incumplimiento de los supuestos habría implicado analizar patrones espaciales sobre errores no válidos, lo que habría comprometido la solidez metodológica de los resultados.

8.3. PROPUESTA FINAL: METODOLOGÍA OPERATIVA ANOMALÍAS

Los resultados del presente trabajo evidencian que la predicción de anomalías técnicas presenta desafíos significativos, especialmente asociados a la baja resolución temporal de los datos y la limitada disponibilidad de registros de las anomalías. Aun así, se identifica un conjunto claro de variables, métodos y pasos que permiten a una empresa sanitaria actualizar y mejorar periódicamente su sistema de detección de anomalías, así como una ruta concreta para fortalecer la calidad de la información y, con ello, la capacidad predictiva del modelo.

1. VARIABLES REQUERIDAS

La predicción de anomalías técnicas presenta un nivel de dificultad mayor que la predicción de fraude, debido a que:

- Las anomalías dependen menos del comportamiento del cliente y más de fallas operativas, hidráulicas o metrológicas.
- La base es pequeña y con pocos registros.

- La lectura mensual manual no captura anomalías transitorias, lo que el documento resalta como una limitación estructural del problema.

Aun así, con los resultados obtenidos, es posible definir un método y un conjunto claro de variables indispensables.

1.1. Variables estructurales del medidor (estáticas):

Son requeridas en cada actualización del modelo y provienen del sistema comercial o metrológico:

- *Año simplificado*
- *Clase metrológica simplificada*
- *Diámetro simplificado*
- *Marca simplificada*
- *Transmisión simplificada*
- *Ruedas simplificadas*
- *Localidad simplificada y ubicación georreferenciada*

1.2. Variables derivadas de la serie de consumo (dinámicas):

Son las variables que permiten capturar patrones sospechosos. El estudio evidenció que las más predictivas fueron:

- *Z mín.*
- *Z atípicos.*
- *Lempel–Ziv Complexity (LZC) 9 bins.*
- *Time Series Length Factor (TSLF).*
- *N° atípicos moderados.*
- *N° atípicos extremos.*
- *Curtosis consumo.*
- *Delta brusco.*

2. METODOLOGÍA PROPUESTA PARA EL MODELO DE ANOMALÍAS

La siguiente metodología se basa exclusivamente en los resultados y conclusiones obtenidos en el análisis de anomalías del estudio. Su propósito es ofrecer un procedimiento estándar que permita a la empresa refrescar, evaluar y mejorar periódicamente su modelo de detección de anomalías, considerando las limitaciones reales identificadas en el trabajo y las variables que demostraron mayor relevancia predictiva.

Paso 1: Actualización de datos y preprocesamiento

- Obtener la base de consumos mensuales de los últimos 12–24 meses por cliente, o la periodicidad definida por la compañía.
- Actualizar la información estructural del medidor: año de fabricación o instalación, clase metrológica, diámetro, marca, transmision y reemplazos recientes.
- Depurar datos: Eliminar consumos negativos o inconsistentes, identificar y codificar explícitamente casos con errores de digitación.

- Aplicar la imputación en consumos faltantes.
- Georreferenciar direcciones nuevas o corregidas.

Paso 2: Extracción automática de características
A partir de las series mensuales se deben generar las variables que demostraron ser relevantes en la predicción de anomalías según este trabajo:

Estadísticos derivados del consumo:

- Consumo mínimo estandarizado (*Z mín.*).
- Desviación estándar mensual.
- *Curtosis* y *asimetría*.
- *N° atípicos moderados*.
- *N° atípicos extremos*.

Variables de complejidad y forma de la serie:

- *Entropía SAX*, altamente relevante en los resultados.
- *Cambios de símbolo*, asociados a variaciones abruptas.
- Lempel–Ziv Complexity (*LZC 9 y 99 bins*).
- Time Series Length Factor (*TSLF*).

Variables espaciales:

- *Latitud* y *Longitud* del cliente, que mostraron patrones espaciales no aleatorios en los modelos basados en árboles.

Estas variables ayudaron a diferenciar anomalías de consumos normales.

Paso 3: Entrenamiento y selección de modelos
De acuerdo con los resultados del estudio y considerando tanto desempeño técnico como interpretabilidad los modelos recomendados para anomalías son:

- **Modelos recomendados:**
  - **Random Forest**: útil como modelo complementario para capturar interacciones y evaluar importancia de variables.
  - **Regresión Logística (potencialmente aplicable):** aunque no fue posible ajustar un modelo de regresión logística para anomalías con la información disponible en este estudio debido a la baja frecuencia del evento y a las limitaciones derivadas de la lectura mensual del consumo. Este tipo de modelo podría resultar útil en escenarios donde la empresa disponga de una base de datos más robusta. En particular, si se amplía el número de anomalías confirmadas, la regresión logística podría emplearse como un modelo interpretable que permita identificar predictores significativos y complementar la detección temprana de anomalías.

- **Modelos no recomendados como base operativa:**

- **DBSCAN**: no fue adecuado para anomalías debido a su baja sensibilidad y dificultad para separar la clase minoritaria (limitación mencionada en el documento).

Paso 4: Uso operativo y priorización técnica
Con base en las probabilidades de anomalía generadas por los modelos supervisados:

- Obtener la probabilidad estimada de anomalía para cada medidor.
- Ordenar los clientes según dicho puntaje.
- Establecer un umbral operativo basado en la capacidad real de atención técnica (por ejemplo, top X % de anomalía).
- Priorizar inspecciones preventivas o revisiones técnicas en los casos más críticos.

El documento enfatiza que, dada la limitación de la lectura mensual, la utilidad del modelo en este momento es principalmente como *herramienta de alerta temprana* y apoyo para priorización técnica.

Paso 5: Validación continua y retroalimentación
Cada periodo definido por la empresa (por ejemplo, trimestral):

- Reentrenar los modelos con nuevas etiquetas provenientes de revisiones técnicas realizadas.
- Evaluar variaciones del AUC-ROC, AUC-PR, *F1-Score*, *Brier Score* y calibración.
- Revisar la estabilidad espacial de las anomalías detectadas, conforme a resultados previos del estudio.
- Identificar si existen cambios estructurales en los datos (por ejemplo, reemplazos masivos de medidores o variaciones en la red que puedan afectar la señal del consumo).
- Ajustar políticas de generación de variables si las métricas muestran deterioro consistente.

9 CONCLUSIONES

Se concluye que el modelo desarrollado es efectivo y financieramente sostenible para la detección de fraudes, logrando transformar los datos históricos de facturación en un activo estratégico. La investigación determinó que el algoritmo *Random Forest* con ponderación de clases es la solución definitiva, superando a técnicas tradicionales y garantizando la viabilidad económica:

- **Impacto económico:** Al operar con una capacidad de inspección del 1 % de la cartera (bolsón de riesgo), el modelo genera una ganancia neta proyectada de $1.760.000 CLP por lote y un ROI del 78,57 %, superando el umbral de rentabilidad operativa.
- **Gestión de recursos:** La integración de matrices de costos en la función de optimización permitió priorizar inspecciones con alto retorno, cumpliendo con la necesidad de gestionar eficazmente los recursos de la empresa.

Conclusión respecto al objetivo específico 1

Se concluye que la metodología propuesta logró sistematizar el procesamiento de datos y la modelación predictiva con éxito, validando empíricamente qué técnicas son viables para la operación y cuáles no:

1. Superioridad del Random Forest (RF)

La comparación exhaustiva de algoritmos determinó que el *Random Forest* es el modelo más competente para este dominio.

- **En fraude:** Superó a XGBoost tanto en métricas técnicas (F1-Score de 0,315 vs. 0,302) como en desempeño económico (mROI del 64,3 % frente al 59,9 % de XGBoost). Su arquitectura de ensamblaje demostró ser más robusta frente al ruido y al desbalance extremo de clases (1,65 % de prevalencia), ofreciendo una calibración de probabilidades más estable.
- **En anomalías:** El *Random Forest* también se posicionó como líder, alcanzando un F1-Score promedio de 0,573, superior al 0,503 de XGBoost, lo que confirma su capacidad para generalizar mejor ante patrones de consumo irregulares.

2. Ineficacia de los enfoques alternativos

El estudio aportó evidencia crítica para descartar métodos tradicionales y no supervisados en este contexto específico:

- **Regresión logística:** En el caso de anomalías, a pesar de su interpretabilidad, se demostró su inviabilidad técnica debido al incumplimiento sistemático de los supuestos de linealidad en el logit y bondad de ajuste. Esto confirma que la relación entre las variables de consumo y anomalías es altamente no lineal y compleja.
- **DBSCAN:** El enfoque no supervisado, aunque conceptualmente atractivo por no requerir etiquetas, falló en la práctica operativa. No logró separar eficazmente la clase minoritaria de fraude debido a la superposición de densidades con los consumos normales, resultando en una sensibilidad cercana a cero para la detección de fraude.

3. Validación de la metodología integral

Se validó que el éxito del modelo depende de una metodología de preprocesamiento rigurosa que incluya limpieza, imputación, georreferenciación y, crucialmente, la extracción de características temporales complejas. La combinación de este tratamiento de datos con una estrategia de validación cruzada estratificada y calibración permitió mitigar el sesgo del desbalance, entregando una herramienta predictiva que no solo es estadísticamente válida, sino económicamente rentable para la empresa.

Conclusión respecto al objetivo específico 2

Se concluye que la capacidad predictiva del modelo no reside en variables demográficas generales, sino en la ingeniería de características. La identificación de las variables críticas fue producto del cálculo cuantitativo de la importancia de características del modeo *Random Forest*, el cual reveló una dualidad fundamental en los fenómenos estudiados:

- **Para el fraude:** Las variables más influyentes resultaron ser la *Clase metrológica simplificada* (17,44 %) y el *Año simplificado* (12,96 %). Esto concluye que el fraude no es aleatorio, sino que explota la vulnerabilidad física de medidores antiguos o de tecnologías específicas, concentrándose además en bolsones geográficos.
- **Para las anomalías:** La jerarquía de importancia cambió drásticamente, priorizando la Curtosis del Consumo (13,6 %) y el Z-mínimo (11,59 %). Esto indica que las fallas técnicas (no intencionales) se manifiestan mediante la deformación de la distribución y la introducción de ruido, independientemente de la marca del medidor.

Propuesta de variables críticas

Basado en lo anterior, la empresa debe construir obligatoriamente las siguientes 15 variables para replicar el modelo:

- **Variables metrológicas:** *Clase metrológica simplificada* y *Año simplificado*.
- **Variables espaciales:** *Latitud*, *Longitud* y *Localidad simplificada*.
- **Comportamiento de consumo:** *Z-mín.* y *Curtosis consumo* como estadísticas robustas y variables de complejidad temporal tales como *Entropía SAX*, *Complejidad de Lempel-Ziv* (LZC) y *Time Series Length Factor* (TSLF).

Conclusión respecto al objetivo específico 3

La evaluación en la muestra real reveló una dicotomía en el rendimiento según el tipo de irregularidad, exponiendo las limitaciones intrínsecas de la base de datos:

- **Detección de fraude:** El modelo es altamente efectivo para *rankear* prioridades de inspección (Top-K), permitiendo inspecciones dirigidas con alta probabilidad de éxito.
- **Detección de anomalías técnicas:** Para las anomalías (fallas no intencionales), el enfoque actual es insuficiente (Precisión $< 35\,\%$). Se concluye que la baja frecuencia de muestreo (lectura manual mensual) impide detectar anomalías transitorias con precisión.
- **Recomendación final:** Se recomienda a la empresa utilizar este modelo como herramienta de fraude y solo como sistema de alerta temprana (revisión manual) para anomalías técnicas, reconociendo que el desbalance extremo impone un techo técnico que impide la automatización completa sin verificación en campo.

10 TRABAJO FUTURO

El presente estudio deja abiertas varias líneas de investigación que podrían fortalecer la capacidad predictiva y operativa del sistema de detección de fraude. En primer lugar, la incorporación de datos con mayor resolución temporal, provenientes de medidores digitales o telemedidos, permitiría modelar consumos horarios o diarios y mejorar la detección de anomalías transitorias que no pueden identificarse con registros mensuales. Asimismo, futuros desarrollos podrían integrar modelos espaciales explícitos o técnicas de aprendizaje profundo, con el fin de capturar dependencias territoriales y patrones no lineales que exceden las capacidades de los modelos utilizados en este trabajo.

Desde la perspectiva operativa, una línea relevante consiste en articular los puntajes de riesgo del modelo con herramientas de planificación de inspecciones, optimización de rutas y sistemas de alerta temprana. Finalmente, la validación externa del modelo en otras regiones o configuraciones de parque de medidores permitiría evaluar su transferibilidad y robustez, ampliando su utilidad como herramienta estratégica para la gestión de activos y la reducción de pérdidas comerciales.

11 ANEXOS

11.1. Correlación de variables: Mapa de calor

- **Fraude:**

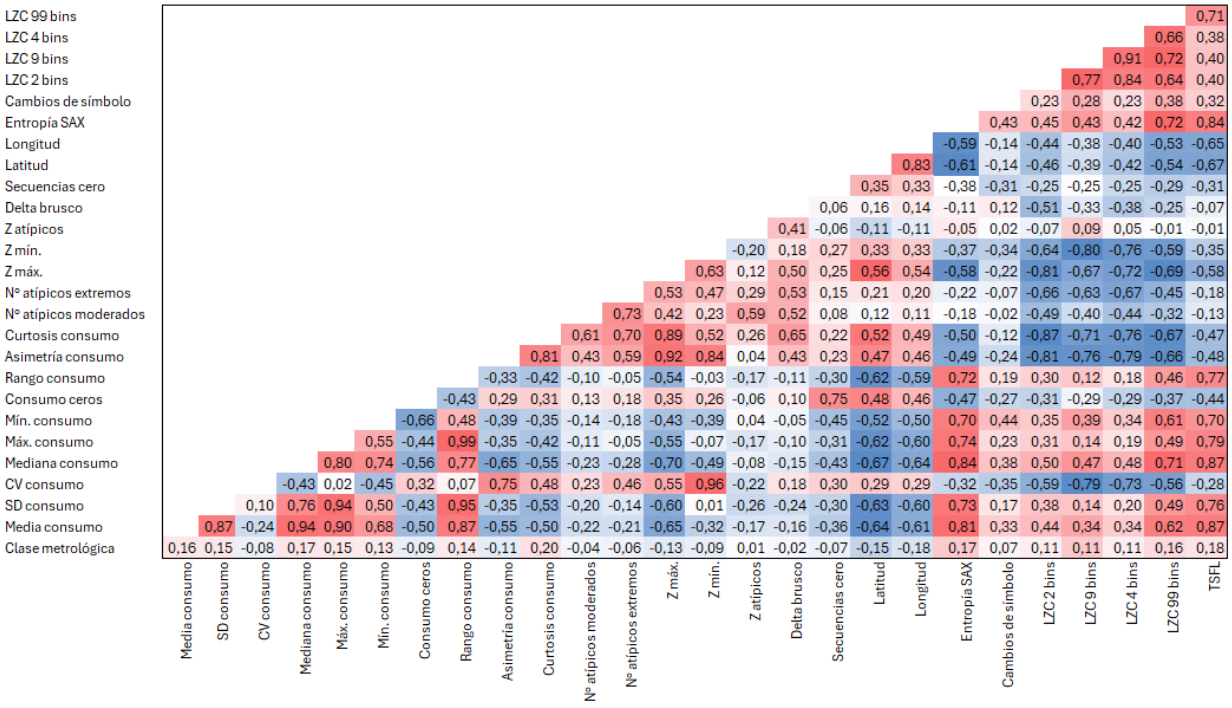

Figura 11.1: Mapa de calor correlaciones para la detección de fraude
Fuente: Elaboración propia.

- **Anomalías:**

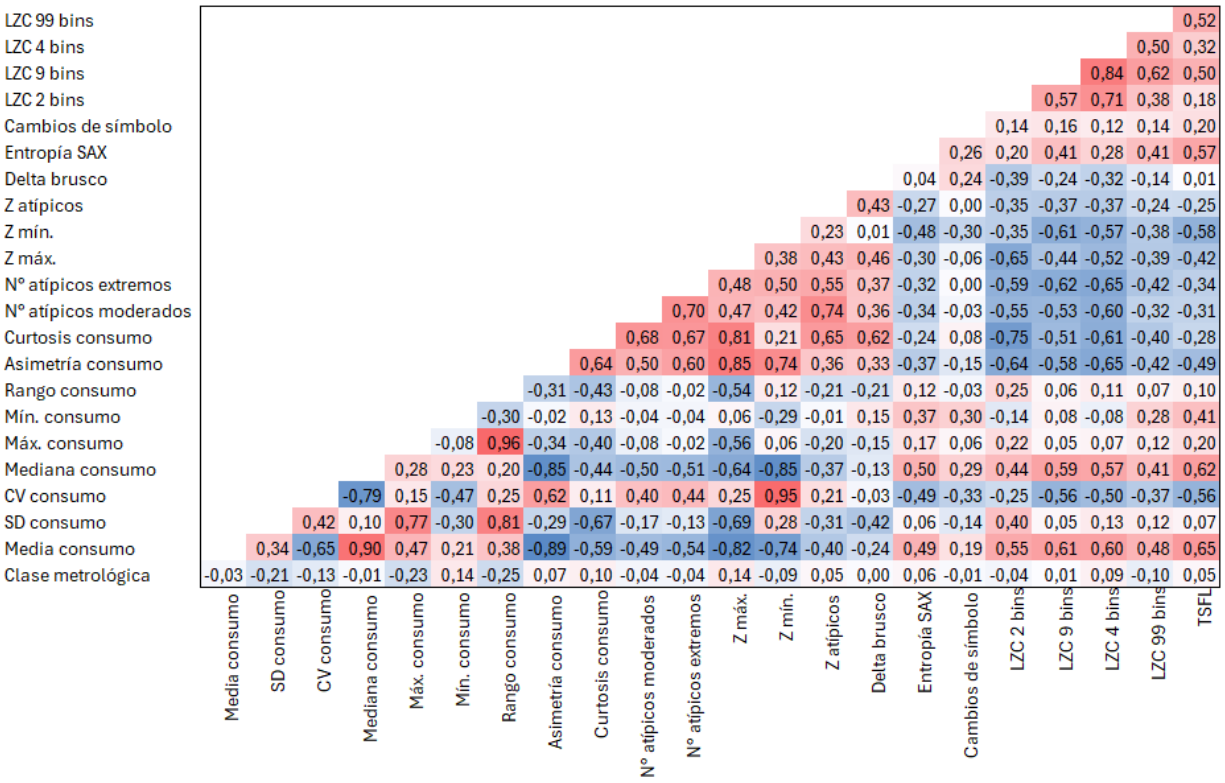

Figura 11.2: Mapa de calor correlaciones para la detección de anomalías
Fuente: Elaboración propia.

11.2. Curvas GAM

Modelación de fraude:

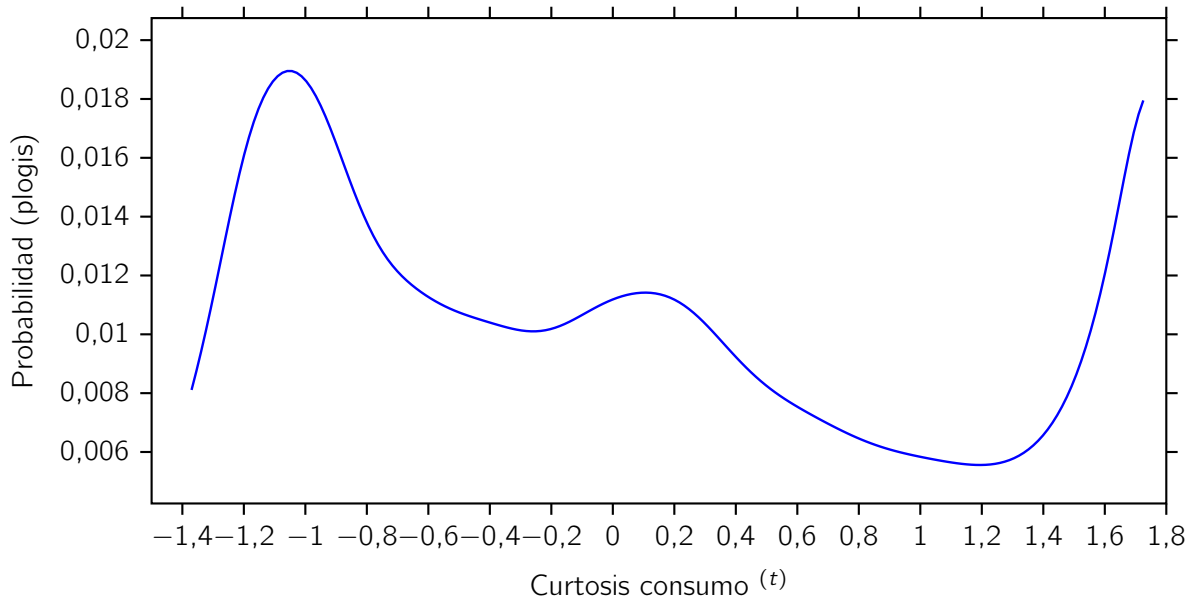

Figura 11.3: Curva GAM fraude: Curtosis consumo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

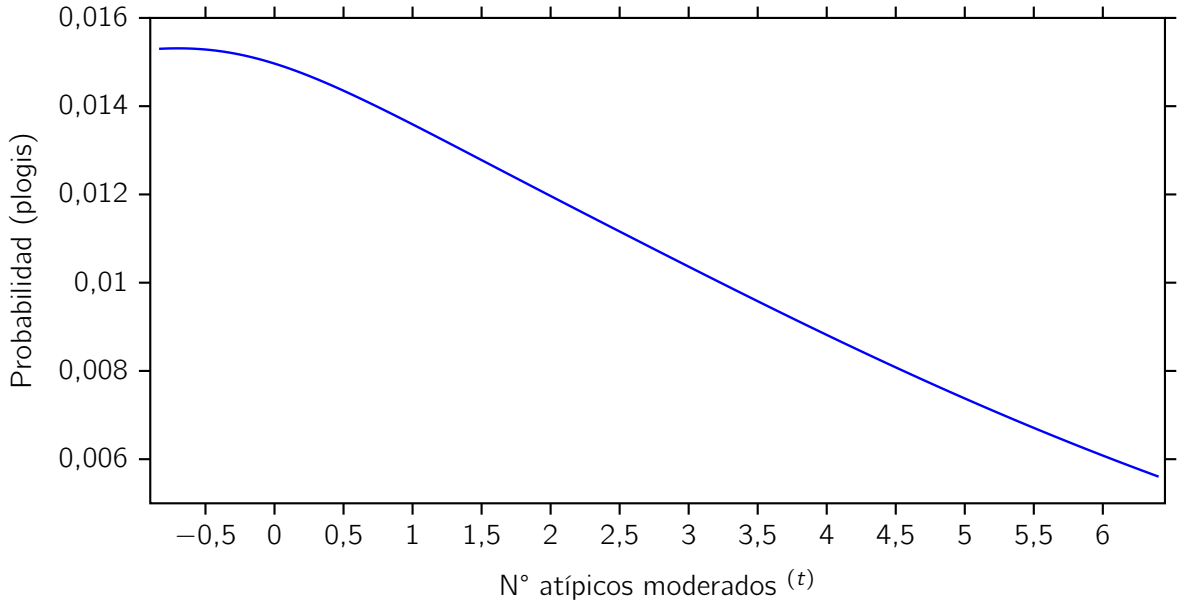

Figura 11.4: Curva GAM fraude: N° atípicos moderados $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

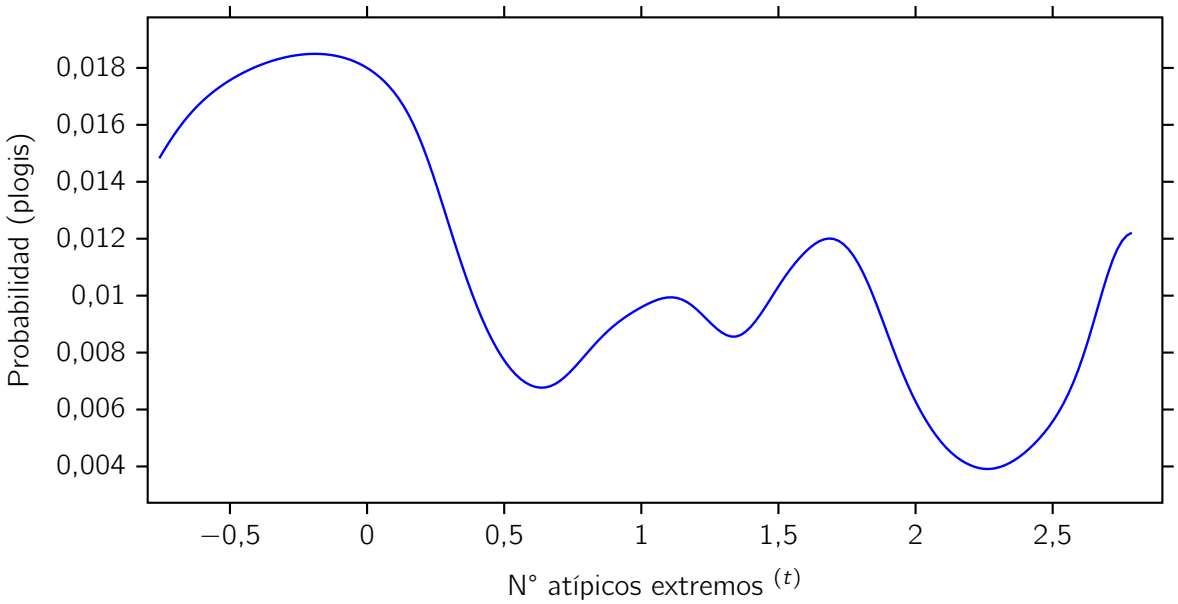

Figura 11.5: Curva GAM fraude: N° atípicos extremos $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

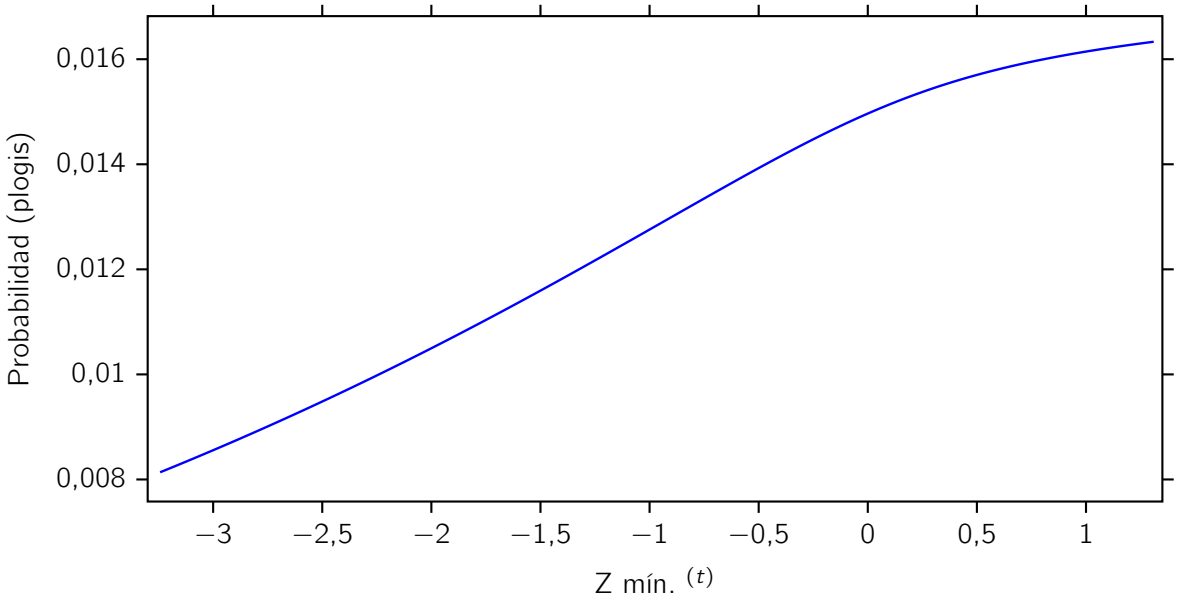

Figura 11.6: Curva GAM fraude: Z mín $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

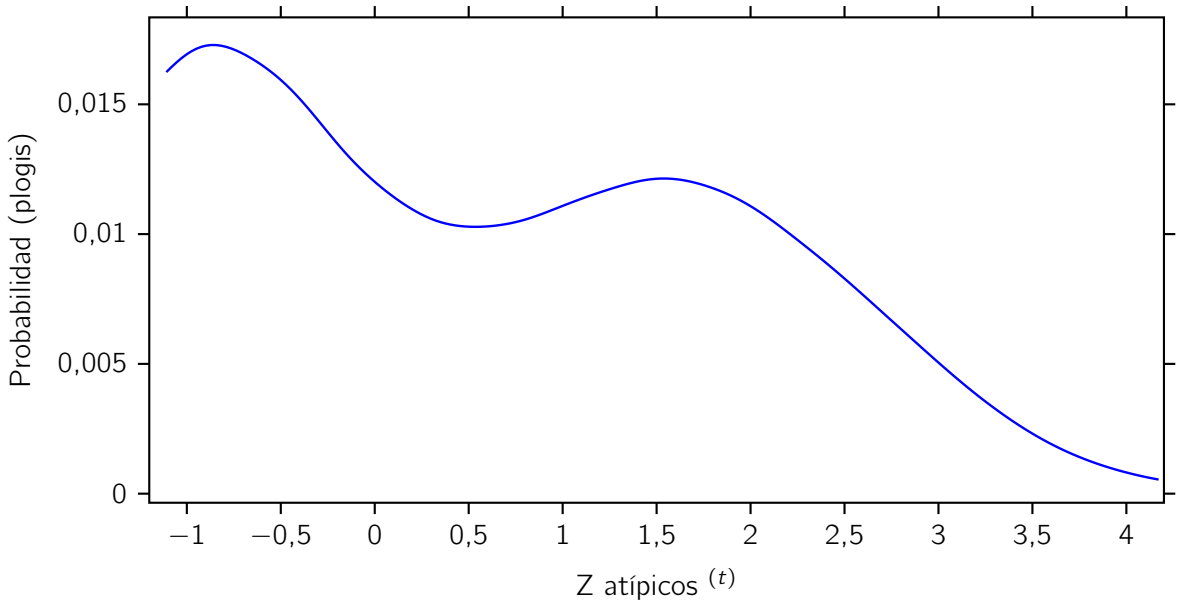

Figura 11.7: Curva GAM fraude: Z atípicos $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

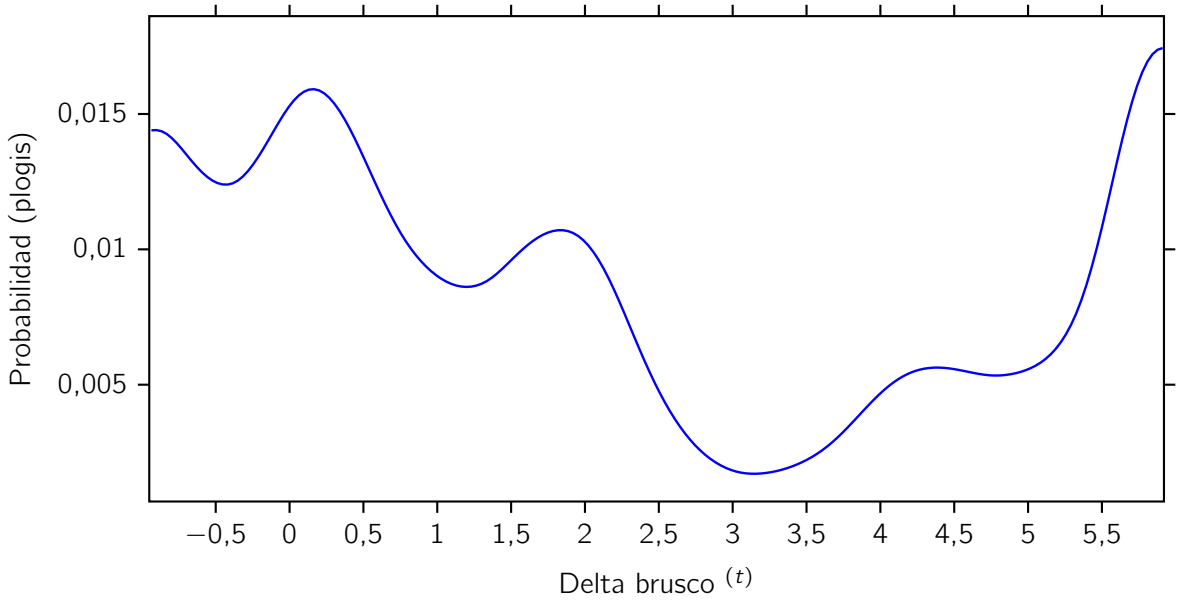

Figura 11.8: Curva GAM fraude: Delta brusco $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

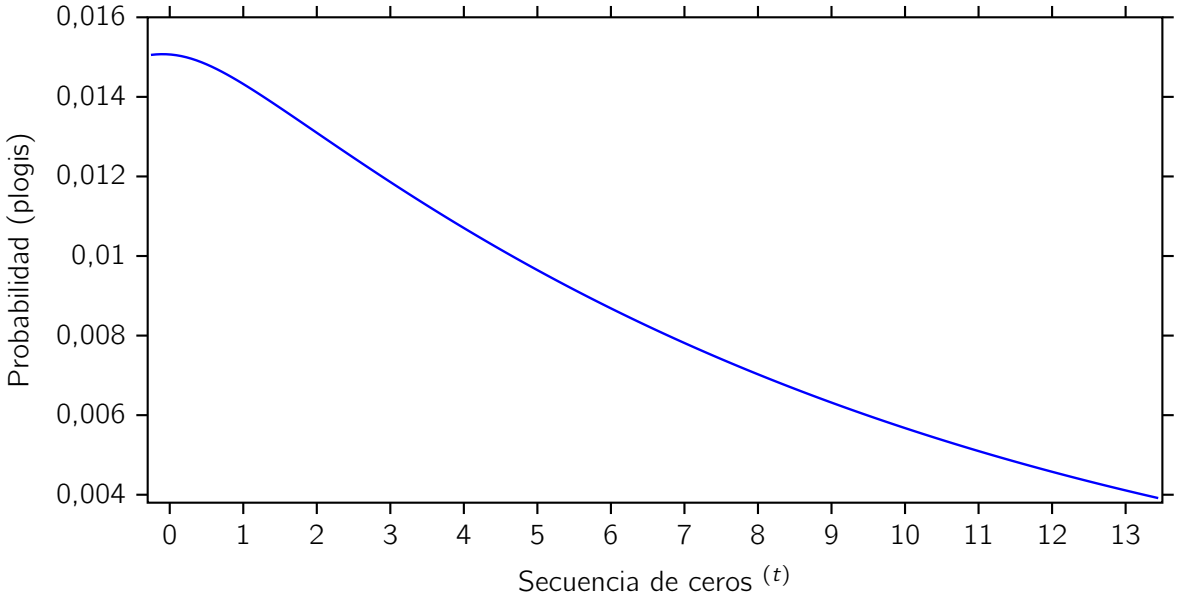

Figura 11.9: Curva GAM fraude: Secuencia de ceros $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

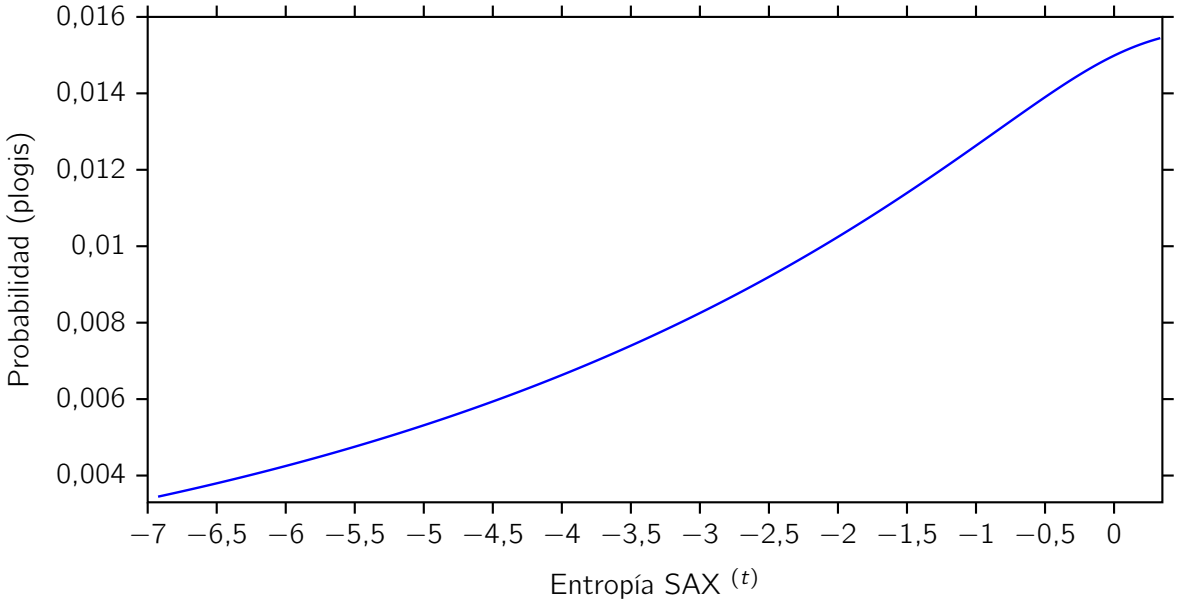

Figura 11.10: Curva GAM fraude: Entropía SAX $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

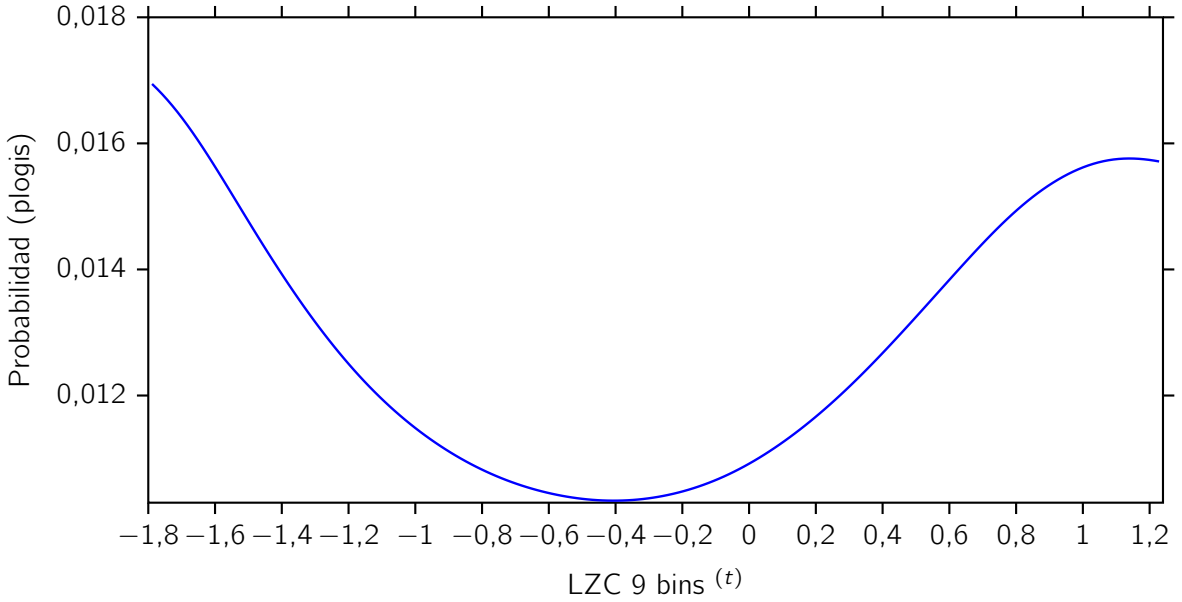

Figura 11.11: Curva GAM fraude: LZC 9 bins $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

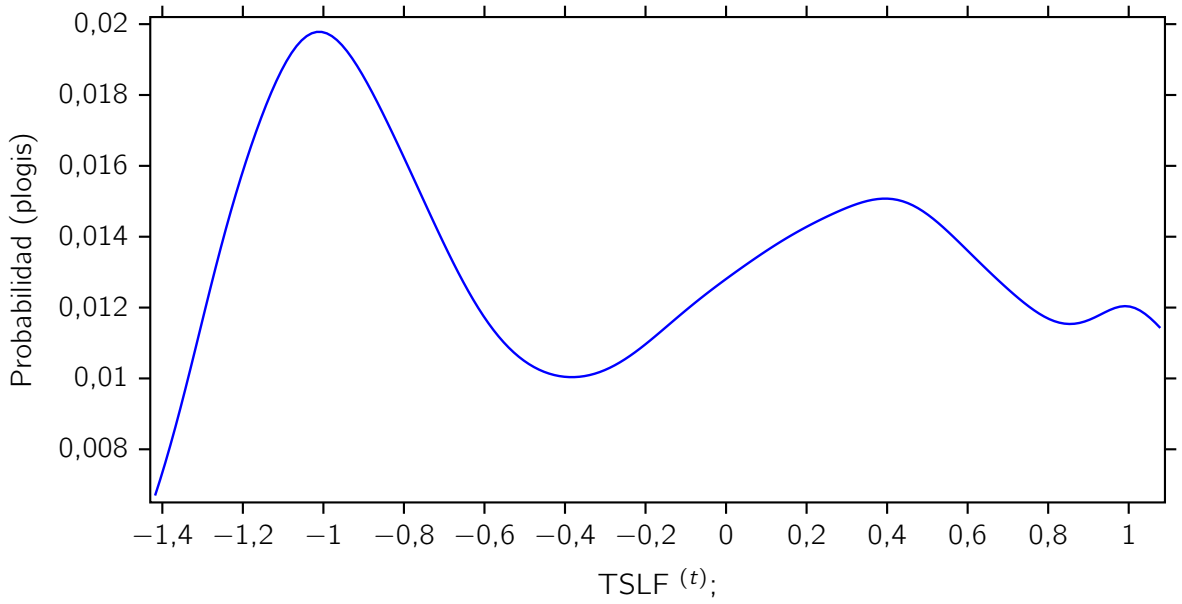

Figura 11.12: Curva GAM fraude: TSLF $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

Modelación de anomalías:

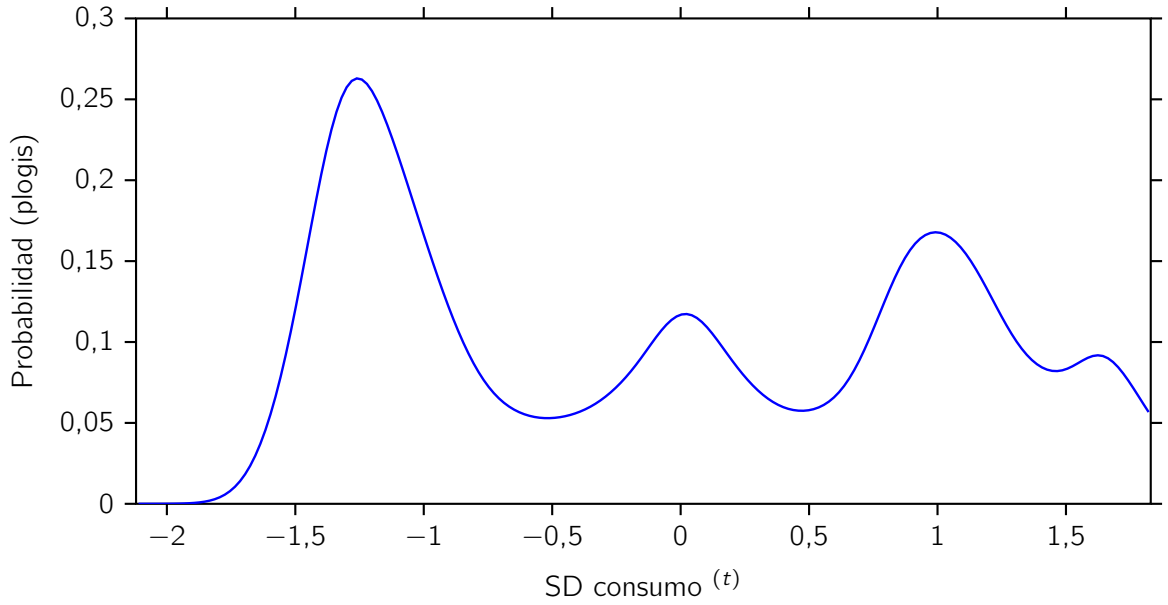

Figura 11.13: Curva GAM anomalías: SD consumo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

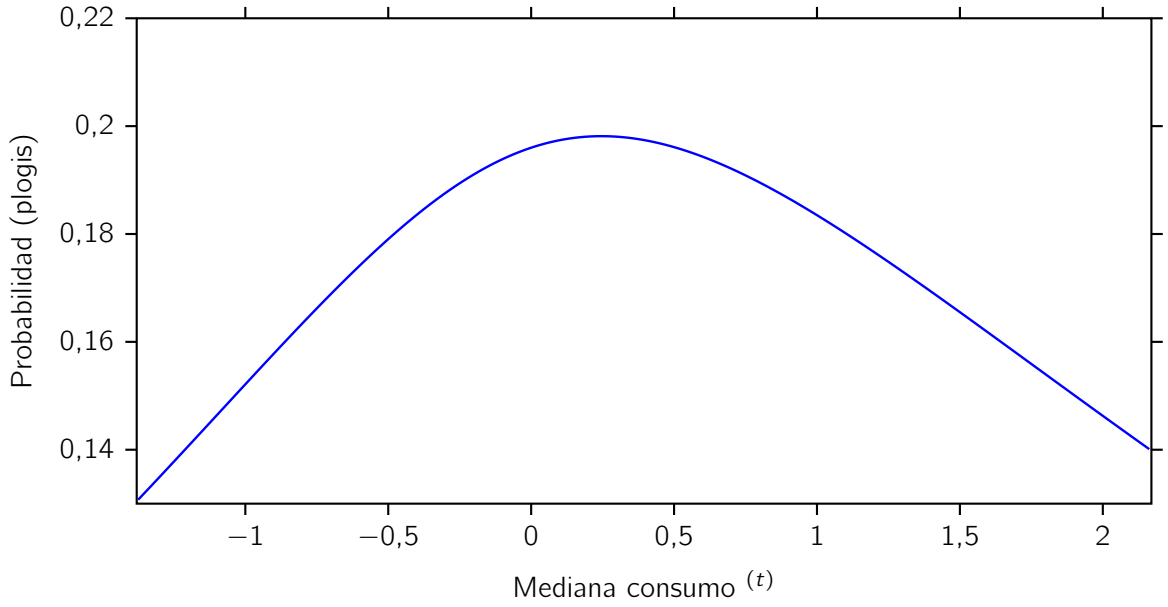

Figura 11.14: Curva GAM anomalías: Mediana consumo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

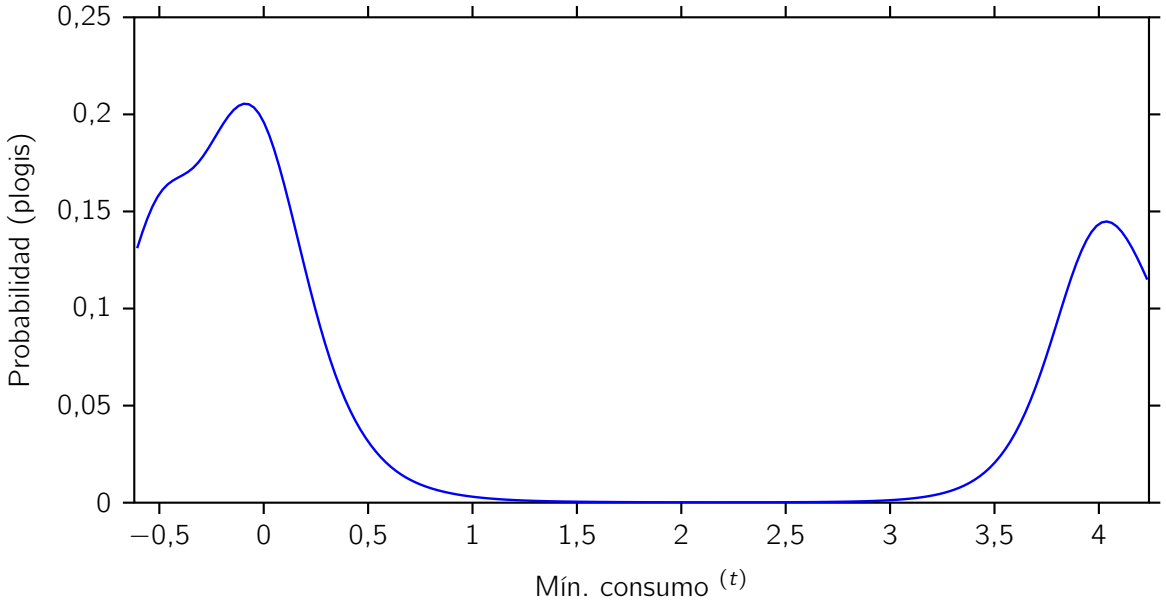

Figura 11.15: Curva GAM anomalías: Mín. consumo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

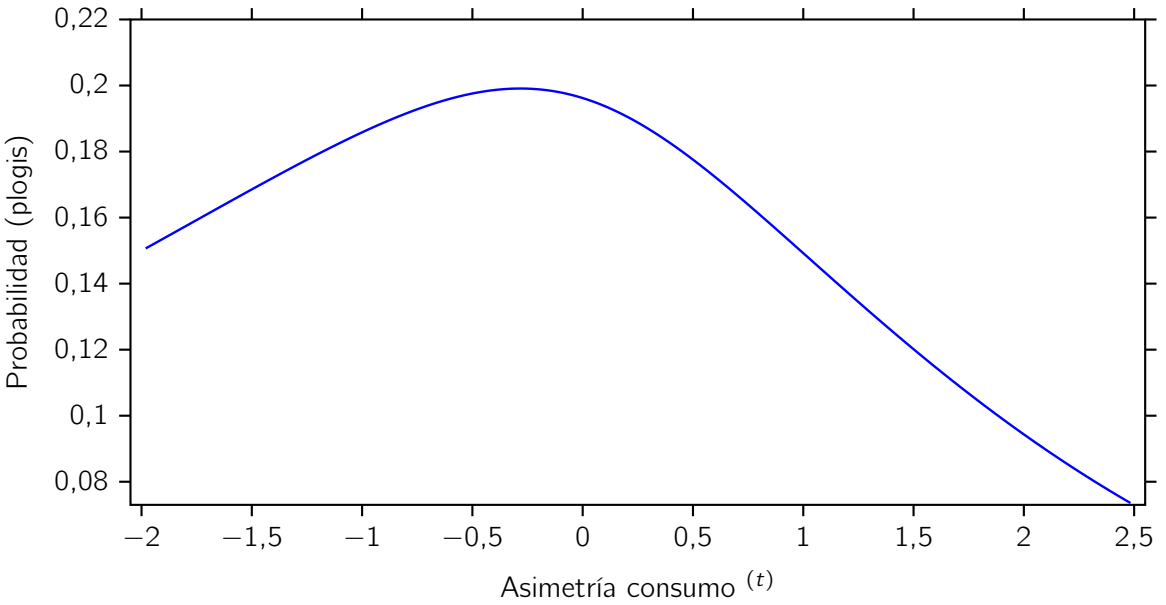

Figura 11.16: Curva GAM anomalías: Asimetría consumo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

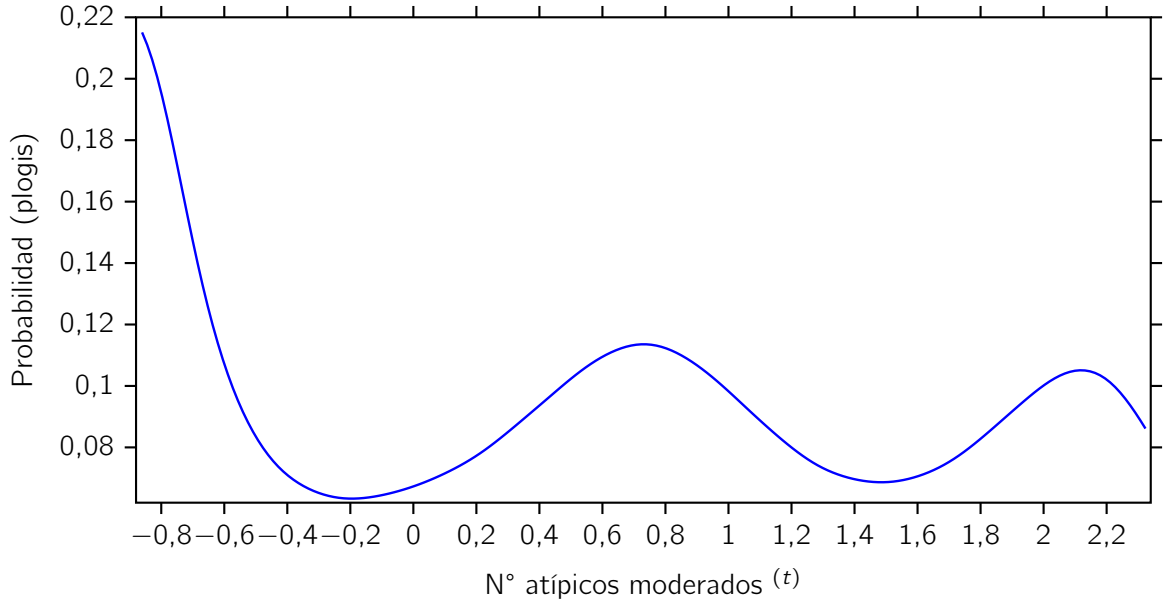

Figura 11.17: Curva GAM anomalías: N° atípicos moderados $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

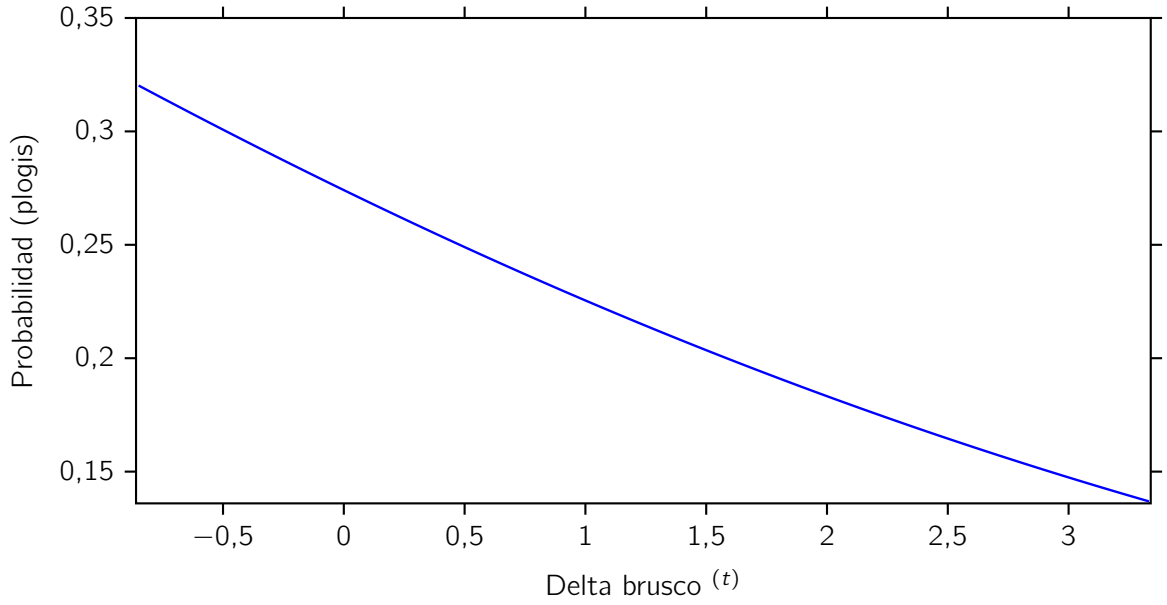

Figura 11.18: Curva GAM anomalías: Delta brusco $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

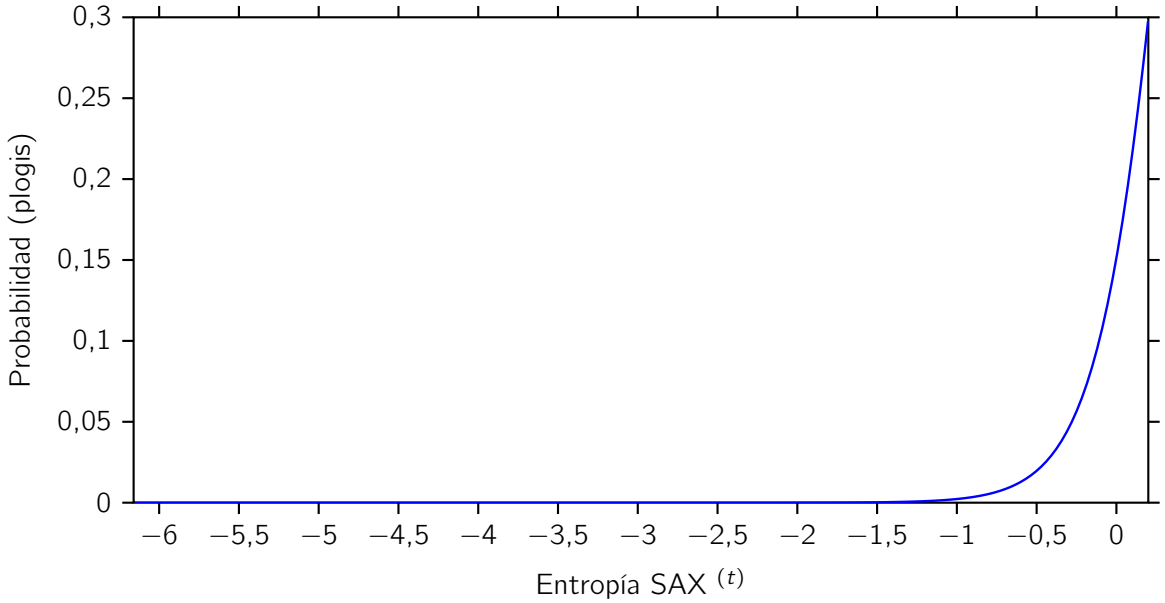

Figura 11.19: Curva GAM anomalías: Entropía SAX $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

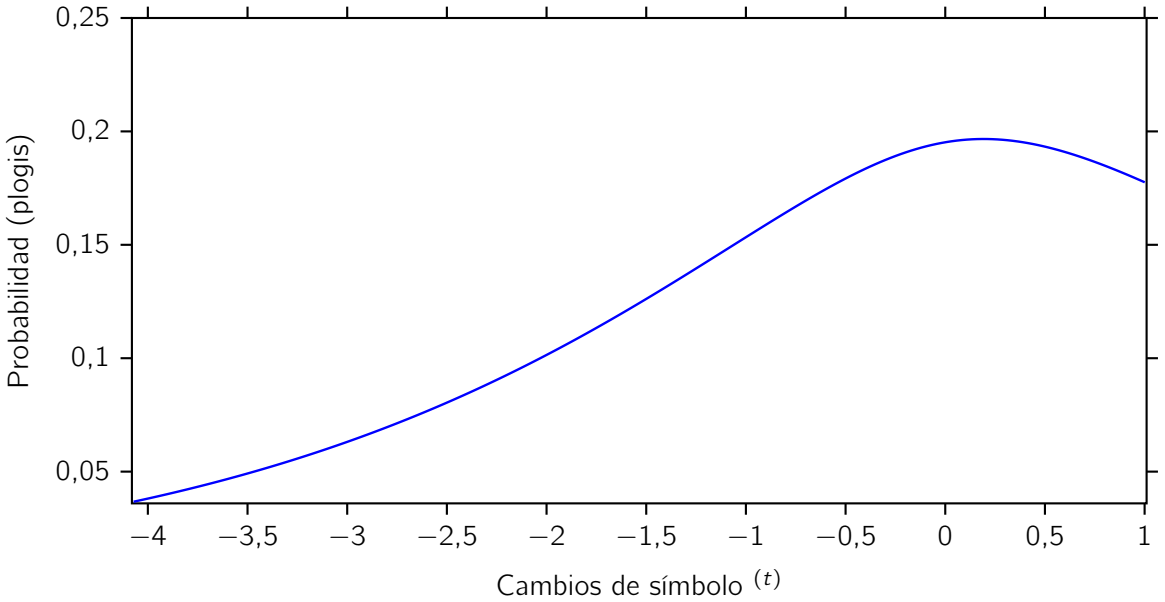

Figura 11.20: Curva GAM anomalías: Cambios de símbolo $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

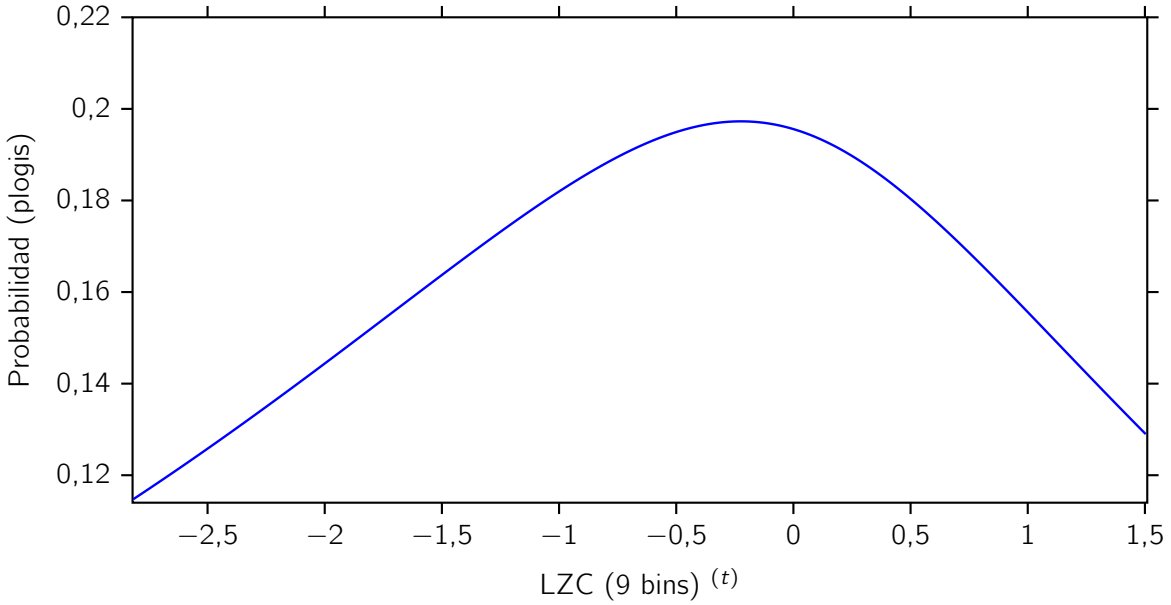

Figura 11.21: Curva GAM anomalías: LZC 9 bins $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

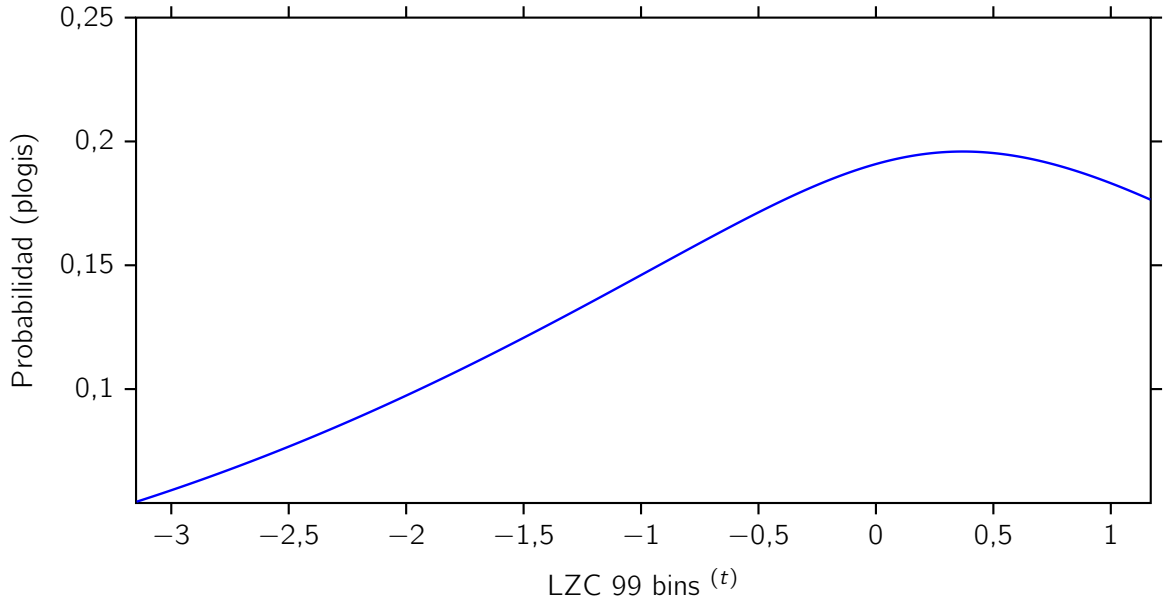

Figura 11.22: Curva GAM anomalías: LZC 99 bins $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

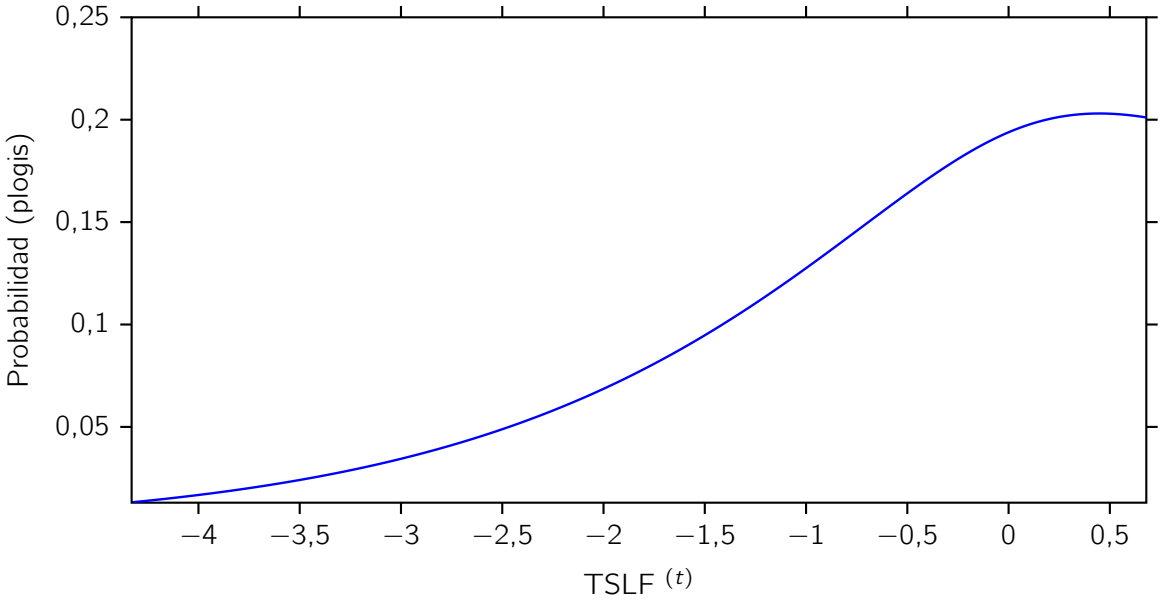

Figura 11.23: Curva GAM anomalías: TSLF $^{(t)}$ vs. Probabilidad (plogis).
Fuente: Elaboración propia.

Bibliografía

[1] The Clinic. «Un Tercio del Agua Producida por Sanitarias en Chile se Pierde.» Accedido: 2024-10-29. (2024), dirección: https://www.theclinic.cl/2024/07/01/un-tercio-del-agua-producida-por-sanitarias-en-chile-se-pierden/.

[2] O. R. Medrano, «Retos y Oportunidades para una Gestión Eficiente de los Servicios de Agua Potable, Saneamiento y Electricidad en República Dominicana,» *Acta Universitaria*, vol. 29, e2364, 2019. doi: 10.15174/au.2019.2364.

[3] Banco Mundial. 2021, «El Agua en Chile: Elemento de Desarrollo y Resiliencia,» Banco Mundial, Washington, DC.

[4] La Tercera. «Así ha Cambiado el Consumo de Agua Potable en los Últimos 25 Años en el País.» Accedido: 2023-10-29. (2023), dirección: https://www.latercera.com/que-pasa/noticia/asi-ha-cambiado-el-consumo-de-agua-potable-en-los-ultimos-25-anos-en-el-pais/KTCXNSLTRBAERLKEUVRG7M2LFQ/#.

[5] La Tribuna. «Crisis Hídrica: Chile Presenta la Sequía más Prolongada de los Últimos Tiempos.» Accedido: 2024-10-29. (2023), dirección: https://www.latribuna.cl/medio-ambiente/2023/06/03/crisis-hidrica-chile-presenta-la-sequia-mas-prolongada-de-los-ultimos-tiempos.html.

[6] International Organization of Legal Metrology (OIML), «Water Meters for Cold Potable Water and Hot Water. Part 1: Metrological and Technical Requirements,» International Organization of Legal Metrology, Paris, France, inf. téc., 2013, OIML R 49-1:2013 (E). dirección: https://www.oiml.org.

[7] iAgua, «¿Qué es un Contador de Agua y Cuántos Tipos Hay?» Accedido: 2024-18-11, 2024. dirección: https://www.iagua.es/respuestas/que-es-contador-agua-y-cuantos-tipos-hay.

[8] Cicasa, «Cómo Leer el Medidor de Agua,» Accedido: 2024-18-11, 2024. dirección: https://cicasa.com/como-leer-el-medidor-de-agua.

[9] K. Rojas, «Análisis de Series de Tiempo.» Bookdown, 2024, Accedido: 2024-21-11. dirección: https://bookdown.org/keilor_rojas/CienciaDatos/an%C3%A1lisis-de-series-de-tiempo.html.

[10] D. García, J. Quevedo, V. Puig et al., «Detection of Fraud and Data Reconstruction in Water Networks Using Principal Component Analysis and Structured Residuals,» *IFAC-PapersOnLine*, vol. 48, n.º 21, págs. 220-225, 2015. doi: 10.1016/j.ifacol.2015.09.523.

[11] S. Carrasco-Jiménez, M. Carrera-Gómez y P. Martínez-Santos, «Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches,» *Universitat Politècnica de Catalunya, School of Civil Engineering*, págs. 1-72, 2021, Trabajo de fin de máster, Máster Universitario en Ciencia y Tecnología del Agua. dirección: https://upcommons.upc.edu/bitstream/handle/2117/328782/DetectionofAnomalousPatternsinWaterConsumption-2.pdf.

[12] J. C. Carrasco-Jiménez, F. Baldaro y F. Cucchietti, «Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches,» en *Intelligent Systems Conference (IntelliSys)*, vol. 1, Springer, 2021, págs. 19-33. doi: 10.1007/978-3-030-55180-3_2.

[13] S. R. Mounce y J. B. Boxall, «Implementation of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows,» *Water Science and Technology: Water Supply*, vol. 11, n.º 5, págs. 629-636, 2011. doi: 10.2166/ws.2011.072.

[14] F. Kaveh-Yazdy y S. Zarifzadeh, «Water Meter Replacement Recommendation for Municipal Water Distribution Networks Using Ensemble Outlier Detection Methods,» *Journal of Artificial Intelligence and Data Mining*, vol. 9, n.º 4, págs. 425-438, 2021. doi: 10.22044/JADM.2021.10672.2202.

[15] F. Arregui, E. Cabrera y R. Cobacho, «Performance Analysis of Water Metering: From Measurement to Management.» London: IWA Publishing, 2018. doi: 10.2166/9781780407874.

[16] M. Bouzada, C. Rodrigues, R. Nogueira y J. Vieira, «Smart Water Metering: A Review of Data-Driven Applications and Challenges,» *Water*, vol. 12, n.º 11, pág. 3123, 2020. doi: 10.3390/w12113123.

[17] B. del Congreso Nacional de Chile, «Código Penal: Artículo 489° bis,» https://www.bcn.cl/leychile, Disponible en línea. Accedido: 2024-11-13, 2024.

[18] G. de Chile, «Ley General de Servicios Sanitarios, Artículo 36,» Accedido: 2024-11-17, 2024. dirección: https://www.bcn.cl/leychile.

[19] G. de Chile, «Ley General de Servicios Sanitarios, Artículo 47A,» Accedido: 2024-11-17, 2024. dirección: https://www.bcn.cl/leychile.

[20] C. Cordeiro, A. Borges y M. R. Ramos, «A Strategy to Assess Water Meter Performance,» *Journal of Water Resources Planning and Management*, vol. 148, n.º 2, págs. 05021027-1-05021027-11, 2022. doi: 10.1061/(ASCE)WR.1943-5452.0001492.

[21] A. Cravero y S. Sepúlveda, «Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio,» en *Workshop Internacional EIG2009*, Departamento de Ingeniería de Sistemas, Universidad de La Frontera, Temuco, Chile, Diciembre de 2009. dirección: https://www.researchgate.net/publication/221419375_Aplicacion_de_Mineria_de_Datos_para_la_Deteccion_de_Anomalias_Un_Caso_de_Estudio.

[22] S. B. Kotsiantis, «Supervised Machine Learning: A Review of Classification Techniques,» *Informatica*, vol. 31, n.º 3, págs. 249-268, 2007. dirección: https://datajobs.com/data-science-repo/Supervised-Learning-%5BSB-Kotsiantis%5D.pdf.

[23] Encord, «Train-Test-Validation Split: How to And Best Practices,» Accedido: 2024-11-21, 2023. dirección: https://encord.com/blog/train-val-test-split/.

[24] E. Rahm y H. H. Do, «Data Cleaning: Problems and Current Approaches,» *IEEE Data Engineering Bulletin*, vol. 23, n.º 4, págs. 3-13, 2000. dirección: https://www.researchgate.net/publication/220282831_Data_Cleaning_Problems_and_Current_Approaches.

[25] R. J. A. Little y D. B. Rubin, «Statistical Analysis with Missing Data,» 3rd. John Wiley & Sons, 2019. doi: 10.1002/9781119482260.

[26] A. Scrivano, «Fraud Detection Pipeline Using Machine Learning: Methods, Applications, and Future Directions,» *Unpublished manuscript, Politecnico di Milano, DEIB Department*, 2024.

[27] I. Khongsrabut y K. Waiyamai, «Outliers Detection in Time Series Data : Case study: Provincial Waterworks Authority,» en *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, 2019, págs. 234-238. doi: 10.1109/ECTI-NCON.2019.8692257.

[28] J. W. Tukey, «Exploratory Data Analysis.» Reading, MA: Addison-Wesley, 1977.

[29] S. Nofal, A. Alfarrarjeh y A. A. Jabal, «A use case of anomaly detection for identifying unusual water consumption in Jordan,» *Water Supply*, vol. 22, n.º 1, págs. 1131-1140, 2021. doi: 10.2166/ws.2021.210. dirección: https://doi.org/10.2166/ws.2021.210.

[30] R. Chambers, W. Cleveland, B. Kleiner y P. Tukey, «Robust automatic methods for outlier and error detection,» *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 167, n.º 3, págs. 323-339, 2004. dirección: https://www.istat.it/wp-content/uploads/2014/05/Chambers_et_al-2004_Journal-of-the-Royal-Statistical-Society-Series-A-Statistics-in-Society.pdf.

[31] D. B. Rubin, «Inference and Missing Data,» *Biometrika*, vol. 63, n.º 3, págs. 581-592, 1976. doi: 10.1093/biomet/63.3.581.

[32] C. K. Enders, «Applied Missing Data Analysis.» Guilford Press, 2010.

[33] D. W. Goldberg, M. G. Cockburn, F. W. Lurmann y B. Ritz, «A Geocoding Best Practices Guide,» *International Journal of Health Geographics*, vol. 10, n.º 1, pág. 35, 2011. doi: 10.1186/1476-072X-10-35.

[34] P. A. Longley, M. F. Goodchild, D. J. Maguire y D. W. Rhind, «Geographic Information Science and Systems,» 4th. Hoboken, New Jersey: Wiley, 2015, isbn: 978-1-118-67695-0.

[35] M. F. Goodchild, «Geographic Information Systems and Science: Today and Tomorrow,» *Annals of GIS*, vol. 15, n.º 1, págs. 3-9, 2009. doi: 10.1080/19475680903250715.

[36] F. Harvey, «The Social Construction of Geographic Information Systems.» New York, USA: Routledge, 2008, isbn: 978-0-415-77346-9.

[37] B. W. Silverman, «Density Estimation for Statistics and Data Analysis.» London: Chapman y Hall, 1986. doi: 10.1007/978-1-4899-3324-9.

[38] D. W. Scott, «Multivariate Density Estimation: Theory, Practice, and Visualization,» 2nd. John Wiley & Sons, 2015. doi: 10.1002/9781118575574.

[39] C. Brunsdon y L. Comber, «An Introduction to R for Spatial Analysis and Mapping,» 3rd. London: SAGE Publications, 2022, isbn: 9781529773364.

[40] R. H. Shumway y D. S. Stoffer, «Time Series Analysis and Its Applications: With R Examples,» 4th. New York: Springer, 2017, isbn: 978-3-319-52452-8. doi: 10.1007/978-3-319-52452-8.

[41] G. E. P. Box, G. M. Jenkins, G. C. Reinsel y G. M. Ljung, «Time Series Analysis: Forecasting and Control,» 5th. Hoboken, NJ: John Wiley & Sons, 2016, isbn: 978-1-118-67502-1. doi: 10.1002/9781118619193.

[42] R. J. Hyndman y G. Athanasopoulos, «Forecasting: Principles and Practice,» 2nd. Melbourne, Australia: OTexts, 2018. dirección: https://otexts.com/fpp2/.

[43] R. Ghamkhar, T. Chonavel, A. Chebira y E. Lee, «DBScan Based Approach for Detection of Abnormal Consumption in Water Meter Data,» *Water*, vol. 15, n.º 6, pág. 1101, 2023. doi: 10.3390/w15061101.

[44] J. Lin, E. Keogh, L. Wei y S. Lonardi, «Experiencing SAX: A Novel Symbolic Representation of Time Series,» *Data Mining and Knowledge Discovery*, vol. 15, n.º 2, págs. 107-144, 2007. doi: 10.1007/s10618-007-0064-z.

[45] L. Yan, X. Wu y J. Xiao, «An Improved Time Series Symbolic Representation Based on Multiple Features and Vector Frequency Difference,» *Journal of Computer and Communications*, vol. 10, n.º 6, págs. 44-62, 2022. doi: 10.4236/jcc.2022.106005.

[46] W. W. S. Wei, «Time Series Analysis: Univariate and Multivariate Methods,» 2nd. Boston, MA: Pearson Education, 2006, isbn: 978-0-321-16531-4.

[47] M. Ghamkhar et al., «An unsupervised anomaly detection method for water consumption time series using DBSCAN and complexity metrics,» *Journal of Water Supply: Research and Technology—AQUA*, vol. 72, n.º 6, págs. 827-842, 2023. doi: 10.2166/aqua.2023.127. dirección: https://doi.org/10.2166/aqua.2023.127.

[48] T. Hastie, R. Tibshirani y J. Friedman, «The Elements of Statistical Learning: Data Mining, Inference, and Prediction,» 2nd. New York: Springer, 2009, isbn: 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7.

[49] R. Kohavi, «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,» en *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, Montreal, Quebec, Canada: Morgan Kaufmann, 1995, págs. 1137-1143.

[50] A. Fernández, S. García y F. Herrera, «Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution,» *Progress in Artificial Intelligence*, vol. 7, n.º 4, págs. 1-12, 2018. doi: 10.1007/s13748-018-0157-4.

[51] N. Japkowicz y S. Stephen, «The Class Imbalance Problem: A Systematic Study,» *Intelligent Data Analysis*, vol. 6, n.º 5, págs. 429-449, 2002. doi: 10.3233/IDA-2002-6504.

[52] T. Hastie, R. Tibshirani y J. Friedman, «The Elements of Statistical Learning: Data Mining, Inference, and Prediction,» 2nd. New York: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort et al., «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[54] J. Bergstra e Y. Bengio, «Random Search for Hyper-Parameter Optimization,» *Journal of Machine Learning Research*, vol. 13, n.º 10, págs. 281-305, 2012. dirección: https://www.jmlr.org/papers/v13/bergstra12a.html.

[55] J. C. Platt, «Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,» en *Advances in Large Margin Classifiers*, Cambridge, MA: MIT Press, 1999, págs. 61-74.

[56] A. Niculescu-Mizil y R. Caruana, «Predicting Good Probabilities with Supervised Learning,» en *Proceedings of the 22nd International Conference on Machine Learning (ICML-05)*, ACM, 2005, págs. 625-632. doi: 10.1145/1102351.1102430.

[57] T. Fawcett, «An Introduction to ROC Analysis,» *Pattern Recognition Letters*, vol. 27, n.º 8, págs. 861-874, 2006. doi: 10.1016/j.patrec.2005.10.010.

[58] D. M. W. Powers, «Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation,» *Journal of Machine Learning Technologies*, vol. 2, n.º 1, págs. 37-63, 2011.

[59] M. Sokolova y G. Lapalme, «A Systematic Analysis of Performance Measures for Classification Tasks,» *Information Processing and Management*, vol. 45, n.º 4, págs. 427-437, 2009. doi: 10.1016/j.ipm.2009.03.002.

[60] L. Breiman, «Random Forests,» *Machine Learning*, vol. 45, págs. 5-32, 2001. doi: 10.1023/A:1010933404324.

[61] J. H. Friedman, «Greedy function approximation: A gradient boosting machine,» *Annals of Statistics*, vol. 29, n.º 5, págs. 1189-1232, 2001. doi: 10.1214/aos/1013203451.

[62] T. Chen y C. Guestrin, «XGBoost: A scalable tree boosting system,» en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, págs. 785-794. doi: 10.1145/2939672.2939785. dirección: https://arxiv.org/pdf/1603.02754v3.pdf.

[63] M. Ester, H.-P. Kriegel, J. Sander y X. Xu, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,» en *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, USA: AAAI Press, 1996, págs. 226-231. dirección: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

[64] D. W. Hosmer, S. Lemeshow y R. X. Sturdivant, «Applied Logistic Regression,» 3rd. New York: John Wiley & Sons, 2013.

[65] S. Menard, «Applied Logistic Regression Analysis,» 2nd. Thousand Oaks, CA: Sage Publications, 2002.

[66] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford y A. R. Feinstein, «A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis,» *Journal of Clinical Epidemiology*, vol. 49, n.º 12, págs. 1373-1379, 1996. doi: 10.1016/S0895-4356(96)00236-3.

[67] C.-Y. J. Peng, K. L. Lee y G. M. Ingersoll, «An Introduction to Logistic Regression Analysis and Reporting,» *The Journal of Educational Research*, vol. 96, n.º 1, págs. 3-14, 2002.

[68] L. Anselin, «Local Indicators of Spatial Association—LISA,» *Geographical Analysis*, vol. 27, n.º 2, págs. 93-115, 1995. doi: 10.1111/j.1538-4632.1995.tb00338.x.

[69] A. D. Cliff y J. K. Ord, «Spatial Processes: Models and Applications.» London: Pion Ltd, 1981.

[70] T. J. Hastie y R. J. Tibshirani, «Generalized Additive Models.» London: Chapman y Hall, 1990. doi: 10.1201/9780203753781.

[71] S. N. Wood, «Generalized Additive Models: An Introduction with R,» 2nd. Boca Raton, FL: Chapman y Hall/CRC, 2017, isbn: 9781498728331.

[72] H. Akaike, «A new look at the statistical model identification,» *IEEE Transactions on Automatic Control*, vol. 19, n.º 6, págs. 716-723, 1974. doi: 10.1109/TAC.1974.1100705.

[73] G. Schwarz, «Estimating the Dimension of a Model,» *The Annals of Statistics*, vol. 6, n.º 2, págs. 461-464, 1978. doi: 10.1214/aos/1176344136.

[74] R. Liemberger y A. Wyatt, «Quantifying the global non-revenue water problem,» *Water Supply*, vol. 19, n.º 3, págs. 831-837, 2019. doi: 10.2166/ws.2018.129.

[75] T. M. Al-Washali, S. K. Sharma y M. D. Kennedy, «Methods of Assessment of Water Losses in Water Supply Systems: A Review,» *Water Resources Management*, vol. 30, n.º 14, págs. 4985-5001, 2016. doi: 10.1007/s11269-016-1503-7.

[76] J. P. Detroz y A. T. da Silva, «Fraud Detection in Water Meters Using Pattern Recognition Techniques,» en *Proceedings of the 32nd Annual ACM Symposium on Applied Computing (SAC '17)*, ACM, 2017, págs. 930-935. doi: 10.1145/3019612.3019634.

[77] F. H. Troncoso Espinosa, P. G. Fuentes Figueroa e I. R. Belmar Arriagada, «Predicción de fraudes en el consumo de agua potable mediante el uso de minería de datos,» *Universidad, Ciencia y Tecnología*, vol. 24, n.º 104, págs. 58-66, 2020. dirección: https://www.researchgate.net/publication/346144224_PREDICCION_DE_FRAUDES_EN_EL_CONSUMO_DE_AGUA_POTABLE_MEDIANTE_EL_USO_DE_MINERIA_DE_DATOS.

[78] Q. A. Al-Radaideh y M. M. Al-Zoubi, «A Data Mining Based Model for Detection of Fraudulent Behaviour in Water Consumption,» en *2018 9th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2018, págs. 48-53. doi: 10.1109/IACS.2018.8355440.

[79] K. P. Teodoro da Silva, A. Kalbusch y E. Henning, «Detection of unauthorized consumption in water supply systems: A case study using logistic regression,» *Utilities Policy*, vol. 84, pág. 101 647, 2023, issn: 0957-1787. doi: 10.1016/j.jup.2023.101647. dirección: https://doi.org/10.1016/j.jup.2023.101647.

[80] M. R. Stramari, A. Kalbusch y E. Henning, «Random forest for the detection of unauthorized consumption in water supply systems: a case study in Southern Brazil,» *Urban Water Journal*, vol. 20, n.º 3, págs. 325-334, 2023. doi: 10.1080/1573062X.2022.2155856.

[81] O. Kainz, U. König, G. Lichtenegger y G. Stettinger, «Non-standard situation detection in smart water metering,» *Open Computer Science*, vol. 11, n.º 1, págs. 259-272, 2021. doi: 10.1515/comp-2020-0190.

[82] M. N. Kanyama et al., «AI-Driven Anomaly Detection in Smart Water Metering Systems Using Ensemble Learning,» *Water*, vol. 17, n.º 13, pág. 1933, 2025. doi: 10.3390/w17131933.

[83] S. Nofal, S. Al-Jabi y M. Al-Qudah, «A use case of anomaly detection for identifying unusual water consumption,» *Water Supply*, vol. 22, n.º 6, págs. 6413-6426, 2022. doi: 10.2166/ws.2021.210.

[84] A. Bagnall, J. Lines, A. Bostrom, J. Large y E. Keogh, «The Great Time Series Classification Bake Off: A Review and Experimental Evaluation of Recent Algorithmic Advances,» *Data Mining and Knowledge Discovery*, vol. 31, n.º 3, págs. 606-660, 2017. doi: 10.1007/s10618-016-0483-9.

[85] F. Troncoso y N. Fernández, «Limpieza, corrección y geocodificación de grandes bases de direcciones utilizando minería de texto,» *Universidad, Ciencia y Tecnología*, vol. 25, n.º 109, págs. 80-87, 2021, issn: 2542-3401/ 1316-4821. doi: 10.47460/uct.v25i109.451. dirección: https://doi.org/10.47460/uct.v25i109.451.

[86] F. Provost y T. Fawcett, «Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.» Sebastopol, CA: O'Reilly Media, 2013, isbn: 978-1-449-36132-7.

[87] C. Elkan, «The Foundations of Cost-Sensitive Learning,» en *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, Seattle, WA: Morgan Kaufmann Publishers, 2001, págs. 973-978.

[88] SASIPA SpA, «Tarifas del servicio de agua potable y alcantarillado 2025,» Cargo por verificación de medidor: 15 351 CLP, 2025. dirección: https://www.sasipa.cl/tarifas-agua-potable (visitado 11-06-2025).

[89] D. G. de Aguas, «Resolución de marzo 2023: multa de 482,6 UTM (30 138 370 CLP) a Essbio por no medir e informar caudales extraídos,» 2023. dirección: https://dga.mop.gob.cl/noticias/multa-essbio-482-utm (visitado 11-06-2025).

[90] B. del Congreso Nacional de Chile, «Código Penal de la República de Chile – Texto actualizado arts. 459–461,» 2021. dirección: https://www.bcn.cl/leychile/navegar?idNorma=1984 (visitado 11-06-2025).

[91] Nueva Atacama S.A., «Boletín de gestión operativa 2023: 231 hallazgos en 6 021 visitas,» 2023. dirección: https://www.nuevaatacama.cl (visitado 11-06-2025).

[92] Essbio S.A., «Hoja tarifaria región del Biobío y Ñuble – Octubre 2024,» Cargo por verificación de medidor: 28 980 CLP, 2024. dirección: https://www.essbio.cl/nosotros/regulacion/nuestras-tarifas (visitado 11-06-2025).

[93] Calendarr, «Estaciones del año en Chile: cuáles son, cuándo cambian y características,» https://www.calendarr.com/chile/estaciones-ano/, Consultado el 30 de septiembre de 2025, 2025.

[94] C. Chen, A. Liaw y L. Breiman, «Using random forest to learn imbalanced data,» University of California, Berkeley, inf. téc., 2004. dirección: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.