MODELO PREDICTIVO DE LA DEMANDA DE SERVICIOS DE ACTSIS LTDA

Diana Marcela Aguirre León
Código 8975610

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Juan Carlos Martinez A.

Codirector(a)
Gerardo Mauricio Sarria M.

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 19 DE 2025

TABLA DE CONTENIDO

Pág.

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

INTRODUCCIÓN

En la industria del software, la capacidad instalada de una empresa ha demostrado ser una pieza fundamental para su competitividad y sostenibilidad a largo plazo. Sin embargo, esta capacidad se ha visto afectada por una serie de variables del entorno de negocio, que van desde la disponibilidad de talento especializado hasta las fluctuaciones económicas y las políticas regulatorias. Estos factores dinámicos impactaron significativamente en la capacidad de las empresas de software para desarrollar, implementar y mantener productos y servicios de alta calidad de manera eficiente y rentable.

En respuesta a estos desafíos, surgió la necesidad de desarrollar un modelo predictivo que permitiera a las empresas de software proyectar cómo estas variables del entorno de negocio podrían afectar su capacidad instalada. Este modelo, al proporcionar una comprensión más profunda de los factores que influyen en la capacidad operativa de una empresa de software, sirvió como una herramienta valiosa para la toma de decisiones estratégicas y operativas.

Este proyecto se centró en la construcción de dicho modelo predictivo, basándose en una investigación sobre el impacto de las variables del entorno de negocio en la demanda de servicios de software de los clientes de la empresa ACTSIS LTDA, en su contexto específico. A través de un enfoque multidisciplinario que combinó conocimientos en análisis de datos, gestión empresarial y tecnología de la información, el proyecto ofreció a ACTSIS LTDA una herramienta efectiva para anticipar los cambios en la demanda de servicios software para el Sistema Comercial – SAC, el cual representaba el pareto de los ingresos de la compañía.

La implementación de este modelo predictivo transformó la gestión interna de ACTSIS LTDA al mejorar la capacidad de anticipar y adaptarse a los cambios en la demanda, contribuyendo a su eficiencia operativa, reducción de costos y planificación informada en la asignación de recursos. En definitiva, este proyecto representó un paso hacia la innovación en la gestión de la capacidad instalada, permitiendo a ACTSIS no solo adaptarse al cambio, sino también liderarlo.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

En el panorama actual de los negocios, la tecnología de la información (TI) desempeña un papel crucial en la transformación y optimización de las operaciones empresariales. La creciente dependencia de las soluciones de software impulsa a las empresas a invertir significativamente en este ámbito, como motor de innovación y competitividad en el entorno empresarial.

De acuerdo con el estudio de mercado realizado por Equiteq, empresa global especializada en la asesoría de fusiones y adquisiciones (M&A), se proyecta que la inversión mundial en TI llegaría a los 6.2 billones de dólares para el año 2026, siendo el software el subsegmento de mayor crecimiento, con una tasa de aumento anual del 11.9% y un crecimiento interanual del 12.8% [1]. Este panorama confirma que las empresas consideraron el desarrollo de software como un factor clave en la optimización de su eficiencia operativa y la satisfacción de las crecientes demandas de los mercados.

Sin embargo, a pesar de las expectativas de crecimiento acelerado en el sector del software, se prevé que la disponibilidad de profesionales calificados en TI no lograría igualar este ritmo de crecimiento. Según el estudio de Equiteq, se anticipa que el gasto mundial en software casi se duplicaría entre 2021 y 2026, mientras que el número de desarrolladores de software a nivel global aumentaría en menos de un 25% en el mismo período [1].

La disparidad entre la creciente demanda y la oferta limitada de talento especializado provoca un desequilibrio en el mercado. Según un estudio sobre **Escasez de Talento para 2024** realizado por Manpower Group, el 79% de las empresas del sector TI enfrentaron dificultades para encontrar el personal necesario para una gestión eficaz de la demanda de servicios [1].

Colombia no es ajena a esta realidad, lo que hace que las empresas nacionales del sector enfrenten dificultades para adquirir oportunamente el recurso vital para su operación, el talento humano calificado. Es así como ACTSIS Limitada, empresa santandereana con más de 30 años de experiencia en el desarrollo e implementación de soluciones software,

experimenta ocasionalmente esta realidad, enfrentando desafíos en la gestión de la demanda de sus servicios, en parte debido a la limitada oferta de talento humano especializado, pero también es afectada por otras variables que inciden estacionalmente en la masividad de las solicitudes de sus clientes.

Los volúmenes normales de requerimientos que ACTSIS recibe y atiende, experimentan fluctuaciones significativas debido a eventos específicos, como la incorporación de nuevos clientes, cambios en el entorno regulatorio del país, actualizaciones tecnológicas de las herramientas de construcción de software y/o la implementación de nuevos productos software, entre otros.

En suma, la falta de recursos, las fluctuaciones de demanda y la ausencia de una medida estandarizada de capacidad instalada en la empresa, puede llevar a una asignación inadecuada de recursos, dificultando la oportuna atención de la demanda de servicios e impactando negativamente la satisfacción del cliente, conduciendo a incumplimientos de acuerdos de nivel de servicio (ANS) y representando sobrecostos por multas aplicadas por los clientes.

Si esta situación persiste por prolongados periodos de tiempo o se torna demasiado reiterativa, podría incluso acarrear otro tipo de dificultades operativas como sobrecarga de personal, enfermedades laborales, incremento en la rotación de personal, disminución de la calidad de los servicios, y pérdida de competitividad en el mercado.

Para abordar este problema, se hizo imprescindible la búsqueda de una herramienta que permitiera proyectar las necesidades de servicios de sus clientes de acuerdo con el comportamiento de ciertas variables estratégicas para ACTSIS. Este proyecto se centró en la creación de un modelo predictivo del comportamiento de la demanda de los servicios ofrecidos por ACTSIS, lo cual permite a la empresa anticipar la capacidad instalada requerida y asignar recursos de manera óptima, mejorando su eficiencia, fortaleciendo la planificación de proyectos, la posición competitiva de la empresa y la percepción de excelencia en el servicio al cliente.

1.2. FORMULACIÓN DEL PROBLEMA

El desafío que implicó la anticipación de la fluctuación de la demanda de servicios de ACTSIS Ltda. supuso una adecuada identificación de variables exógenas y endógenas, lo que llevó a este proyecto a responder la siguiente pregunta:

¿Qué enfoque predictivo puede mejorar la anticipación de la demanda de servicios del Sistema Comercial (SAC) de ACTSIS Ltda., considerando las dinámicas del entorno de negocio?

En ese contexto, se abordaron las siguientes cuestiones:

- ¿Cuáles son las principales variables del entorno de negocio que afectan la demanda de los servicios ofrecidos por el Sistema Comercial SAC de ACTSIS Ltda., y cómo se pueden medir y cuantificar estas variables de manera precisa?
- ¿Qué relación estadística existe entre las variables identificadas y la demanda de los servicios del Sistema Comercial SAC de ACTSIS Ltda., y cuáles son los efectos directos e indirectos de estas variables según el análisis multivariante?
- ¿Cómo se puede desarrollar un modelo predictivo que integre las variables del entorno de negocio de ACTSIS Ltda. y proyecte la demanda de los servicios del Sistema Comercial SAC en diversos escenarios?
- ¿Cuál es la precisión y aplicabilidad del modelo predictivo para proyectar la demanda de los servicios del Sistema Comercial SAC de ACTSIS Ltda?

1.2.1. Sistematización

Para definir claramente el problema de investigación, se establecieron las siguientes sub-preguntas:

- ¿Qué tan grande debería ser la base de datos y cuáles clientes o proyectos deben considerarse?
- ¿Qué métodos se utilizarán para medir y cuantificar las variables de negocio que influyen en la demanda de servicios del SAC en ACTSIS?

- ¿Qué técnicas estadísticas se aplicarán para analizar la relación entre las variables y la demanda de servicios?
- ¿Qué herramienta se utilizará para analizar las variables y construir el modelo?
- ¿Cuántas variables se requieren para que el modelo tenga suficiencia?
- ¿qué indicadores se utilizarán para comparar los modelos?
- ¿Cómo se medirá la precisión y aplicabilidad del modelo elegido, en situaciones reales de demanda?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo predictivo de la demanda de servicios de ACTSIS Ltda, para el sistema comercial SAC, a partir del análisis de variables específicas del entorno de negocio propio de la empresa.

2.2 OBJETIVOS ESPECÍFICOS

1. Identificar y cuantificar las variables del entorno de negocio de ACTSIS Ltda, que impactan el comportamiento de la demanda de los servicios ofrecidos para el Sistema Comercial SAC.

2. Analizar la relación entre las variables identificadas y la demanda de los servicios ofrecidos por ACTSIS Ltda para el SAC, mediante técnicas estadísticas y de análisis multivariante, para entender su impacto directo e indirecto.

3. Desarrollar y evaluar modelos predictivos que integren las variables del entorno de negocio y proyecten la demanda de los servicios ofrecidos por ACTSIS Ltda para el SAC, seleccionando el modelo que ofrezca proyecciones más precisas y fiables.

4. Validar el modelo predictivo utilizando datos históricos de ACTSIS Ltda, para asegurar su precisión y aplicabilidad en la toma de decisiones estratégicas.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

Los siguientes temas se relacionan con el desarrollo de este proyecto, teniendo en cuenta que se apoya en conceptos clave de la predicción de la demanda en ingeniería de software, la elaboración de modelos predictivos desde la Ciencia de Datos, y la gestión de recursos humanos.

- **Predicción de la Demanda**

La predicción de la demanda es una técnica crucial en la gestión de negocios que permite anticipar las necesidades futuras del mercado. Según Chopra y Meindl [2], la precisión en la predicción de la demanda es fundamental para la planificación de la producción, la gestión de inventarios y la optimización de la cadena de suministro. En el contexto de ACTSIS LTDA, empresa de desarrollo de software, esta predicción se traduce en la planificación eficiente de proyectos y la asignación óptima de la mano de obra calificada.

- **Técnicas de Predicción**

Existen diversas técnicas para la predicción de la demanda, que van desde métodos cualitativos hasta cuantitativos. Entre los métodos cuantitativos se encuentran:

- Modelos de series temporales: como ARIMA (Autoregressive Integrated Moving Average), que se utilizan para analizar y predecir datos que cambian con el tiempo.
- Modelos de regresión: que analizan la relación entre variables dependientes e independientes para predecir la demanda futura.
- Algoritmos de aprendizaje automático: como los árboles de decisión, los bosques aleatorios y las redes neuronales, que pueden manejar grandes volúmenes de datos y captar patrones complejos en los mismos.

- **Modelo Predictivo desde la Ciencia de Datos**

La Ciencia de Datos es una disciplina interdisciplinaria que utiliza métodos, procesos, algoritmos y sistemas científicos para extraer conocimiento e insights de los datos. Según

Provost y Fawcett [6], la Ciencia de Datos combina habilidades de estadística, informática y conocimientos específicos del dominio para resolver problemas complejos.

- **Herramientas y Técnicas**

El desarrollo de un modelo predictivo en Ciencia de Datos generalmente implica varias etapas:

- **Recolección de datos:** obtención de datos relevantes. Para el caso de este proyecto serán principalmente internos, asociados a las solicitudes o requisiciones de clientes.
- **Limpieza y preprocesamiento de datos**: eliminación de datos inconsistentes y manejo de valores faltantes.
- **Análisis exploratorio de datos:** comprensión de las características de los datos a través de visualizaciones y estadísticas descriptivas.
- **Modelado predictivo:** selección y entrenamiento de modelos predictivos utilizando técnicas de machine learning.
- **Evaluación y validación del modelo**: aseguramiento de la precisión y generalización del modelo mediante técnicas de validación cruzada.

- **Gestión de Recursos Humanos**

La gestión eficiente de los recursos humanos es crucial para cualquier empresa de desarrollo de software. Según Ulrich [7], la capacidad de una empresa para gestionar y desarrollar su capital humano es un determinante clave de su éxito.

- **Planificación de Recursos**

La planificación de recursos en el contexto de ACTSIS LTDA implica asignar adecuadamente la mano de obra calificada a los proyectos en función de la demanda prevista. Esto no solo optimiza la utilización de los recursos, sino que también mejora la satisfacción de los empleados al evitar sobrecargas de trabajo y garantizar una distribución equilibrada de las tareas.

3.2 ANTECEDENTES

- **Referencia** "**Applied Predictive Modeling"** Kuhn, M., & Johnson, K. (2013). [2].

  Resumen: libro sobre el desarrollo de modelos predictivos aplicada a datos del mundo real. Proporciona tanto fundamentos teóricos como ejemplos prácticos y aplicaciones en R.

  El libro está dividido en cuatro partes principales: fundamentos del Modelado predictivo, modelos predictivos, evaluación de Modelos y aplicaciones y ejemplos.

  Contribución: Este libro ofrece una base sólida en técnicas de modelado predictivo que pueden ser aplicadas al análisis de la demanda de servicios. La combinación de teoría y práctica que presenta permitirá una implementación eficiente y efectiva del modelo predictivo, asegurando que se consideren todos los aspectos críticos desde la preparación de datos hasta la validación del modelo.

- **Referencia "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities"** de Mishra, D., Kumar, P., y Kumar, A. (2021). [3].

  Resumen: este artículo aborda el uso de análisis predictivo de big data en la predicción de la demanda en cadenas de suministro. Los autores exploran diversas metodologías y aplicaciones del análisis predictivo utilizando grandes volúmenes de datos para mejorar la precisión en la predicción de la demanda en entornos de cadenas de suministro. Además, identifican oportunidades de investigación futuras en este campo, enfocándose en cómo las tecnologías emergentes y las técnicas avanzadas de análisis de datos pueden transformar la gestión de la cadena de suministro mediante la anticipación y la respuesta proactiva a cambios en la demanda del mercado.

  Contribución y Diferencia: El artículo se centra en la aplicación de análisis predictivo para mejorar la anticipación de la demanda, buscando desarrollar modelos que utilicen datos históricos y variables del entorno para prever la demanda futura. Discute métodos avanzados de análisis predictivo y analiza diversas técnicas utilizadas en la predicción de la demanda. Estas metodologías pueden servir de guía para seleccionar y aplicar las técnicas más adecuadas para el modelo predictivo en ACTSIS Ltda. En cuanto a las diferencias, la principal es que el artículo aborda la demanda en cadenas de suministro.

- **Referencia**: "**Demand Forecasting Based on Machine Learning Algorithms on Customer Information: An Applied Approach."** Journal of Business Analytics, 4(3), 210-224. Yang, X., Wang, S., & Li, Y. (2021). [4]

  Resumen: Este estudio aplica algoritmos de machine learning para la predicción de la demanda utilizando información de los clientes. Los autores emplean técnicas como el árbol de decisión, bosques aleatorios y redes neuronales para analizar datos históricos de ventas y comportamiento del cliente. El enfoque se centra en la identificación de patrones en los datos de clientes y en cómo estos patrones pueden predecir con precisión la demanda futura.

  Contribución y Diferencia: El trabajo aporta una metodología práctica para el uso de algoritmos de machine learning en la predicción de la demanda, algo que es altamente relevante para este proyecto de grado. La principal diferencia radica en el enfoque del proyecto de grado, que no solo utiliza técnicas de machine learning, sino que también integra un análisis multivariante de variables específicas del entorno de negocio de ACTSIS Ltda. Además, mientras el artículo se centra en el comportamiento del cliente, el proyecto de grado se enfoca en los servicios ofrecidos por el Sistema Comercial SAC, incorporando variables adicionales del entorno de negocio que pueden influir en la demanda.

- **Referencia**: *"**Holt Winter's Method for Time Series Analysis."** A. Mukherjee.* [5]

  Resumen: El artículo presenta el método de Holt-Winters, una técnica de suavizado exponencial triple utilizada para el análisis y pronóstico de series temporales que exhiben tendencia y estacionalidad. Este método descompone la serie temporal en tres componentes principales: nivel (valor promedio), tendencia (pendiente a lo largo del tiempo) y estacionalidad (patrón cíclico repetitivo). El artículo detalla las fórmulas asociadas a cada componente y proporciona ejemplos prácticos de implementación, incluyendo el tratamiento de valores atípicos y datos faltantes.

  Contribución y Diferencia: Este recurso ofrece una explicación clara y práctica del método de Holt-Winters, complementando el marco teórico del proyecto de grado al proporcionar una técnica robusta para el pronóstico de series temporales con tendencia y estacionalidad.

4 IDENTIFICACIÓN DE VARIABLES DE NEGOCIO.

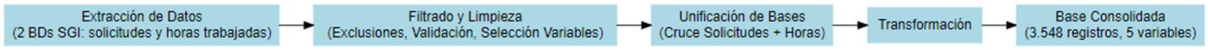

Ilustración 1 Proceso de obtención y consolidación de los datos

La ***ilustración 1*** representa de manera esquemática el proceso desarrollado para la obtención, depuración y consolidación de los datos que sirvieron de insumo para la elaboración de modelos. Este flujo parte de la extracción de datos históricos almacenados en dos bases de datos alimentadas por el Sistema de Gestión de Información (SGI) de ACTSIS Ltda., herramienta *in house* que centraliza tanto las solicitudes de servicio como los reportes de horas trabajadas por cada colaborador.

En una primera etapa, se realizó la **extracción de datos brutos** provenientes de las dos bases principales: la de solicitudes de clientes (V_REQUERIMIENTOS) y la de horas trabajadas (REQ_ACTIVIDADES). Posteriormente, en la fase de **filtrado y limpieza**, se eliminaron registros atípicos, como proyectos en fase de implantación, actividades administrativas y registros sin trazabilidad adecuada. Esta etapa fue crucial para garantizar la calidad de los datos y la validez del análisis posterior.

La tercera fase correspondió a la **unificación de las dos bases** mediante llaves de emparejamiento por fecha, cliente y tipo de solicitud, utilizando como llave el identificador único de cada solicitud (número de requerimiento), lo que permitió obtener una vista consolidada de las solicitudes atendidas y el tiempo asociado a cada una. Luego se aplicaron transformaciones a las variables, como la conversión de fechas a componentes temporales, y la codificación de variables categóricas como factores, adecuándolas para su uso en modelos de aprendizaje automático.

Inicialmente, se contempló el uso de un modelo de series de tiempo, dado que la variable objetivo —las horas trabajadas— presentaba una estructura temporal mensual entre 2022 y 2024. Sin embargo, tras realizar pruebas preliminares, los resultados obtenidos no mostraron un desempeño predictivo satisfactorio. Además, se identificó que la demanda de servicios en ACTSIS está influenciada por múltiples factores exógenos, como el tipo de requerimiento o el cliente, los cuales no podían ser adecuadamente modelados mediante técnicas univariadas tradicionales. En consecuencia, se optó por aplicar modelos de aprendizaje automático supervisado (como Random Forest y XGBoost), los cuales permitieron incorporar otras variables predictoras.

Como resultado de este proceso, se generó una base consolidada definitiva compuesta por 3.548 registros mensuales y cinco atributos clave: cliente, tipo de requerimiento, mes, cantidad de solicitudes trabajadas y horas trabajadas. Esta base fue la utilizada para el desarrollo y evaluación de los modelos.

4.1 Recopilación y Preparación de Datos

El desarrollo de este proyecto implicó la recopilación de datos históricos sobre los servicios del Sistema de Administración Comercial (SAC), brindados por ACTSIS LTDA a sus diversos clientes. Estos datos fueron obtenidos directamente de la base de datos de la empresa, que cuenta con cerca de 20 años de experiencia consolidada en su sistema de gestión de información - SGI (***Ilustración 2***). Este sistema es utilizado por la compañía para centralizar los requerimientos de los clientes y para el reporte diario de las actividades de cada uno de sus empleados. A continuación, se describe el proceso de recopilación y preparación de los datos.

Ilustración 2 Sistema Gestión Información ACTSIS LTDA

4.2 Fuente de datos

Los datos recolectados conforman una serie temporal con información diaria sobre los servicios prestados por ACTSIS, abarcando un período de tres años, desde enero de 2022 hasta diciembre de 2024. Los registros fueron extraídos de dos bases de datos fundamentales: REQ_ACTIVIDADES, que contiene el detalle de las horas trabajadas por requerimiento (con 17 columnas), y V_REQUERIMIENTOS, que almacena la información relacionada con las solicitudes

generadas por los clientes (con 254 columnas), Como se muestra en la ***Ilustración 3.***

Los datos fueron organizados por fecha, cliente y tipo de servicio, permitiendo su posterior depuración, transformación y consolidación. Este proceso fue esencial para analizar la evolución de la demanda de servicios a lo largo del tiempo e identificar patrones históricos relevantes para el desarrollo de los modelos predictivos.

Ilustración 3 Dimensión de las bases de datos de actividades y de solicitudes – ACTSIS Ltda

Las variables recolectadas para cada registro fueron:

- **Número requerimiento:** Identificador único de cada solicitud del cliente
- **Cliente**: Empresa a la cual se le prestó el servicio (incluye en sí mismo a ACTSIS, debido que existe una dedicación interna al desarrollo de plan de producto y atención de mantenimientos correctivos).
- **Tipo de Proyecto**: identificador del proyecto asociado al requerimiento, lo que permite agrupar y gestionar los requerimientos dentro de un contexto de proyecto específico. Un mismo contrato puede tener más de un proyecto activo con características diferentes.
- **Año**: Año en el que se creó la solicitud de servicio.
- **Mes**: Mes en el que se creó la solicitud de servicio.
- **Tipo de servicio**: Clasificación a alto nivel del servicio prestado (desarrollo específico, soporte, mantenimiento, garantía, entre otros).
- **Descripción del Alcance:** Detalle de la solicitud realizada por el cliente.
- **Módulo:** Módulo del sistema al cual está asociada la solicitud (ejemplo: Facturación, Recaudo, Cartera, etc)
- **Horas trabajadas**: Cantidad de horas dedicadas por parte del personal de ACTSIS para resolver cada solicitud, en cada mes. Un mismo requerimiento puede ser atendido y tener horas dedicadas en más de un periodo. Esta última constituye la variable a predecir.

4.3 Limpieza de datos

Durante el proceso de recolección y análisis preliminar de los datos, se identificaron algunos registros que debían ser excluidos de las bases de datos debido a que correspondían a casos atípicos o irrelevantes para la predicción de la demanda. Algunas de las principales exclusiones fueron las siguientes:

- **Proyectos en fase de implantación del sistema**: Estos proyectos presentaban características distintas debido a la naturaleza del proceso y los servicios prestados durante la implementación, lo que los hacía inapropiados para el análisis de la demanda habitual.
- **Tiempo de requerimientos administrativos:** Este tipo de actividades no contribuye directamente a la demanda de los servicios que se buscan predecir, por lo que fue excluido del análisis.
- **Solicitudes atendidas por personal ACTSIS en sitio**: Los registros de solicitudes gestionadas por personal dedicado exclusivamente a un cliente no impactan la atención que se realiza desde la sede central (fábrica), por lo que fueron excluidos también.
- **Atributo MÓDULO**: Aunque este atributo podría haber sido significativo para el análisis y desarrollo del modelo, se decidió excluirlo debido a que la información no era confiable. Su normalización implicaría un costo elevado, ya que requeriría una revisión exhaustiva de cada registro para realizar imputaciones precisas.

Estas exclusiones fueron realizadas con el fin de mejorar la calidad de los datos y garantizar la precisión del modelo predictivo.

4.4 Elección de la herramienta

Para el desarrollo de este proyecto, se eligió **RStudio** como la herramienta principal de análisis y modelado de datos. RStudio ofrece un entorno robusto y especializado en la manipulación, visualización y análisis de datos, siendo ampliamente utilizado en entornos académicos y profesionales para la implementación de proyectos de series de tiempo y modelos predictivos.

Las razones detrás de esta elección incluyen su capacidad para manejar grandes volúmenes de datos, su extensa biblioteca de paquetes especializados en estadística y aprendizaje automático, y su versatilidad para la visualización avanzada de datos. RStudio también permite una

integración eficaz con otros entornos y herramientas de análisis, lo que facilita la personalización y optimización de los modelos para satisfacer los objetivos específicos de este proyecto.

5 ANÁLISIS EXPLORATORIO

5.1 Análisis Exploratorio de Cantidad de Solicitudes

El análisis exploratorio preliminar de la base de datos de **solicitudes** se realizó sobre un total de 16.190 registros únicos, generados por 12 empresas clientes diferentes. Se elaboraron visualizaciones para examinar los volúmenes de solicitudes por cliente, evidenciando variaciones tanto entre clientes como a lo largo de los meses.

Los resultados muestran que el 80% del volumen total de solicitudes proviene de solo cinco empresas, entre ellas ACTSIS, que también gestiona internamente la evolución del producto y la atención de garantías.

Asimismo, se identificaron patrones estacionales a lo largo del año, lo que indica que ciertos periodos registran una menor cantidad de solicitudes. Por último, se observó que el servicio de soporte concentra la mayor demanda en términos de requerimientos generados.

5.1.1 Tendencia por MES y AÑO

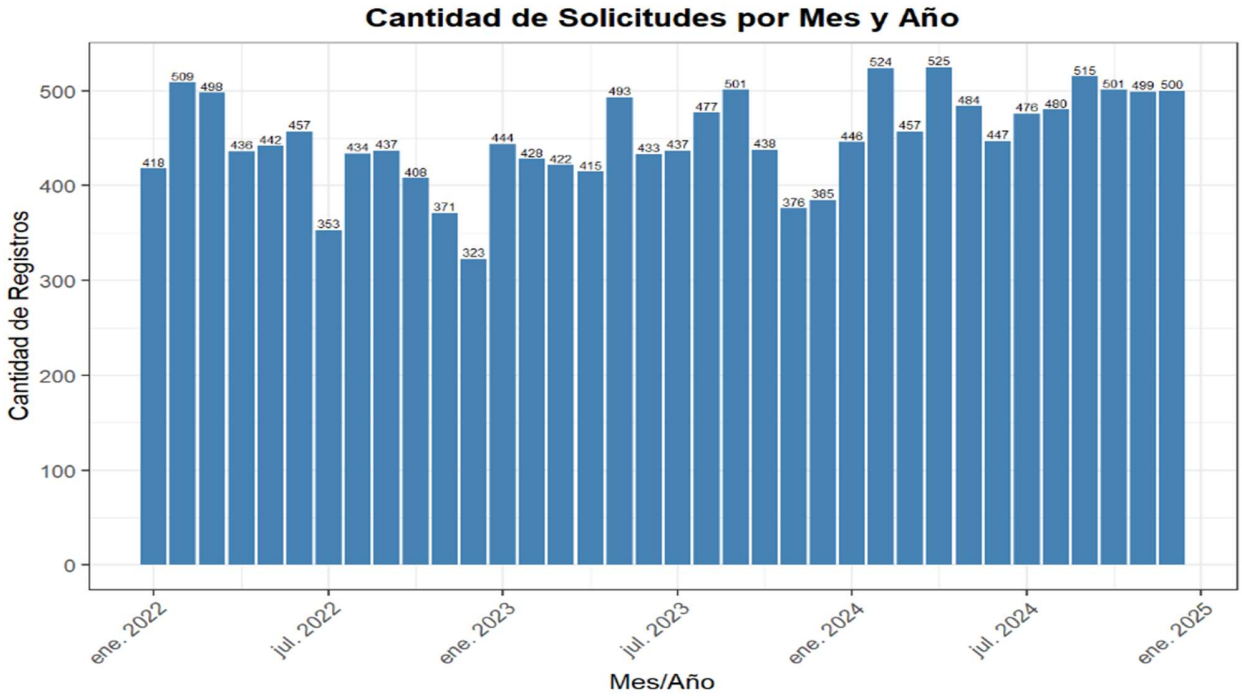

Ilustración 4. Cantidad de solicitudes mensuales ACTSIS entre enero de 2022 y enero de 2025.

La **Ilustración 4** muestra la evolución mensual del número de solicitudes recibidas por ACTSIS a lo largo del periodo comprendido entre enero de 2022 y enero de 2025. En ella se evidencia un crecimiento sostenido en la demanda del servicio, con una aparente estabilización en niveles más altos, a partir de 2024, rondando los 500 requerimientos solicitados por mes. Año a año, se identifican meses específicos en los que la cantidad de solicitudes disminuye drásticamente, lo que podría atribuirse a factores estacionales, tales como periodos vacacionales o menor actividad empresarial en ciertos meses.

Para identificar patrones estacionales en la cantidad de solicitudes, se calcularon los promedios mensuales sin diferenciar el año, los cuales se evidencian en la **tabla 1**. Esto permite observar cómo varía el número de registros en función del mes.

Mes	2022	2023	2024	Promedio Mensual
Enero	418	444	446	436.0
Febrero	509	428	524	487.0
Marzo	498	422	457	459.0
Abril	436	415	525	458.7
Mayo	442	493	484	473.0
Junio	457	433	447	445.7
Julio	353	437	476	422.0
Agosto	434	477	480	463.7
Septiembre	437	501	515	484.3
Octubre	408	438	501	449.0
Noviembre	371	376	499	415.3
Diciembre	323	386	500	403.0

Tabla 1.Promedio de requerimientos por mes, solicitados en ACTSIS entre 2022 y 2024.

La **ilustración 5** muestra la variabilidad en la cantidad de solicitudes recibidas a lo largo del año, lo que permite identificar meses con mayor o menor actividad. Parece haber una estacionalidad en el último trimestre de cada año, donde a partir de octubre el volumen de requerimientos solicitados comienza a disminuir progresivamente hasta diciembre, retomando la tendencia alcista en enero y febrero. El año 2024 mantuvo un atípico alto número de solicitudes en el último trimestre, por factores atípicos de regulaciones CREG que se presentaron en ese periodo.

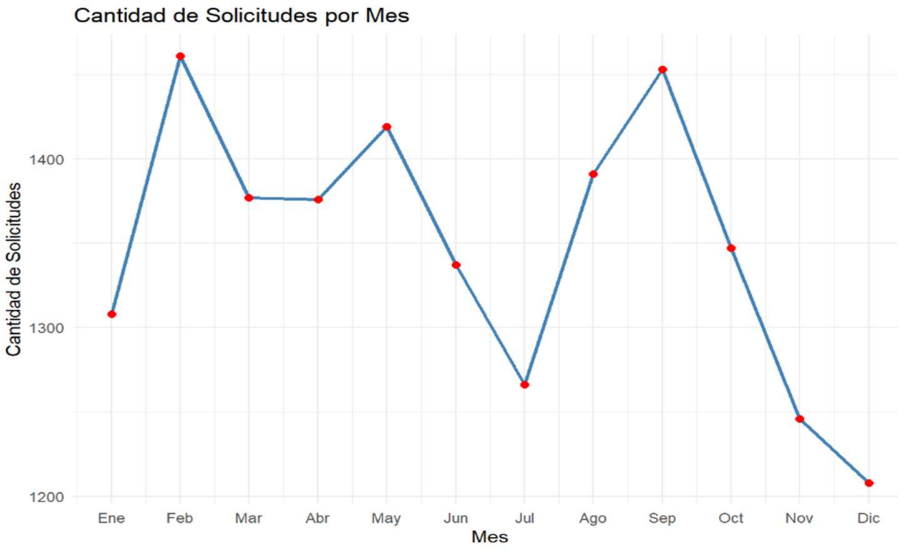

Ilustración 5. Tendencia mensual de la cantidad de solicitudes recibidas durante el año

5.1.2 Tendencia por Cliente

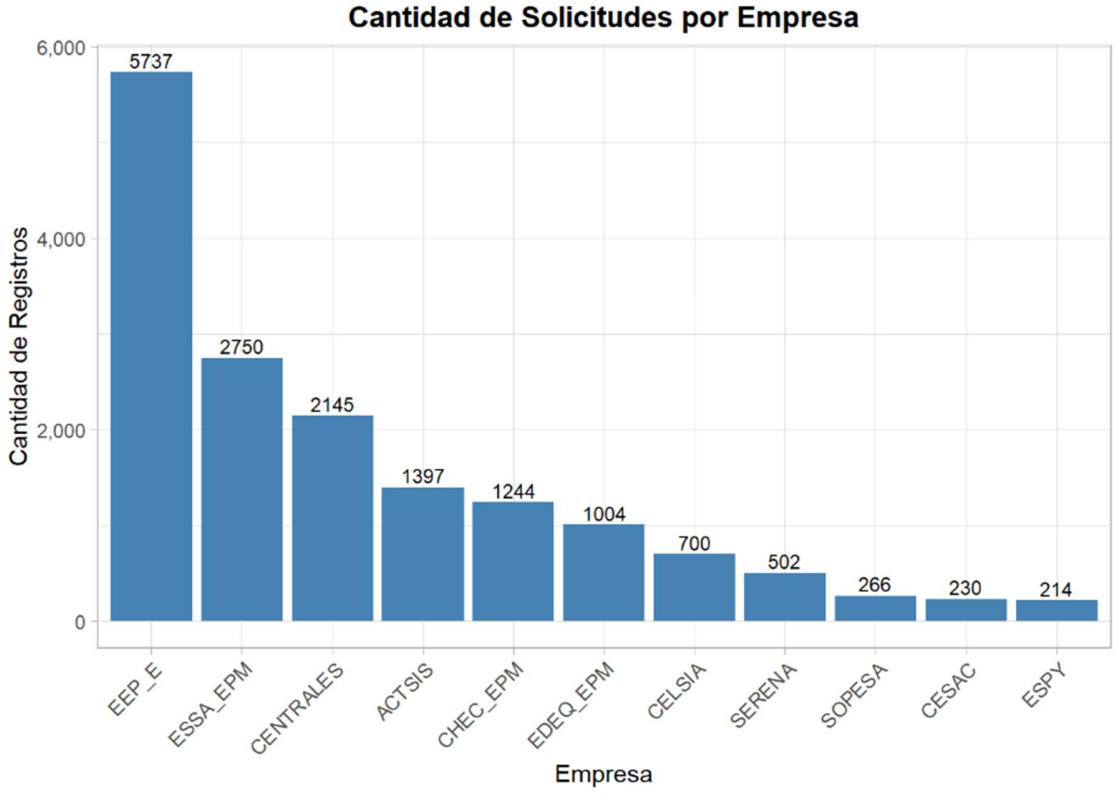

Ilustración 6. Cantidad de solicitudes recibidas por empresa, entre 2022 y 2024

La **Ilustración 6** presenta la distribución de solicitudes recibidas por empresa durante el periodo 2022-2024, evidenciando una alta concentración en cuatro organizaciones: EEP, ESSA, CENTRALES y ACTSIS. Este patrón sugiere que estas entidades representan los principales demandantes del servicio, posiblemente debido a su tamaño, nivel de actividad o dependencia operativa de los sistemas gestionados.

La variabilidad en la cantidad de solicitudes gestionadas por cada empresa puede estar influenciada por diversos factores, entre los que destacan:

- **Tamaño de la empresa:** Las organizaciones con un mayor volumen de solicitudes podrían atender una base de clientes más amplia o gestionar operaciones de mayor complejidad.
- **Relación contractual:** Algunas empresas pueden mantener acuerdos estratégicos que les otorgan un soporte prioritario, lo que impacta en la frecuencia y volumen de solicitudes.
- **Naturaleza de los requerimientos:** La cantidad de solicitudes también puede depender del tipo y la complejidad de los servicios demandados, lo que refleja necesidades operativas específicas de cada empresa. Este es el caso de EEP, cuya dinámica de gestión operativa requiere un alto flujo de requerimientos de soporte de menor magnitud.

Empresa	Cantidad de Solicitudes	Porcentaje (%)
EEP_E	5,737	35.4
ESSA_EPM	2,750	17.0
CENTRALES	2,145	13.2
ACTSIS	1,397	8.6
CHEC_EPM	1,244	7.7
EDEQ_EPM	1,004	6.2
CELSIA	700	4.3
SERENA	502	3.1
SOPESA	266	1.6
CESAC	230	1.4
ESPY	214	1.3
ENERCA	1	0.0

Tabla 2. Distribución de solicitudes creadas por empresa cliente, entre 2022 y 2024

Comprender esta concentración resulta clave para la toma de decisiones operativas y estratégicas, ya que permite priorizar recursos, ajustar capacidades de atención y anticipar la demanda futura de manera más precisa.

5.1.3 Tendencia por Tipo de Servicio

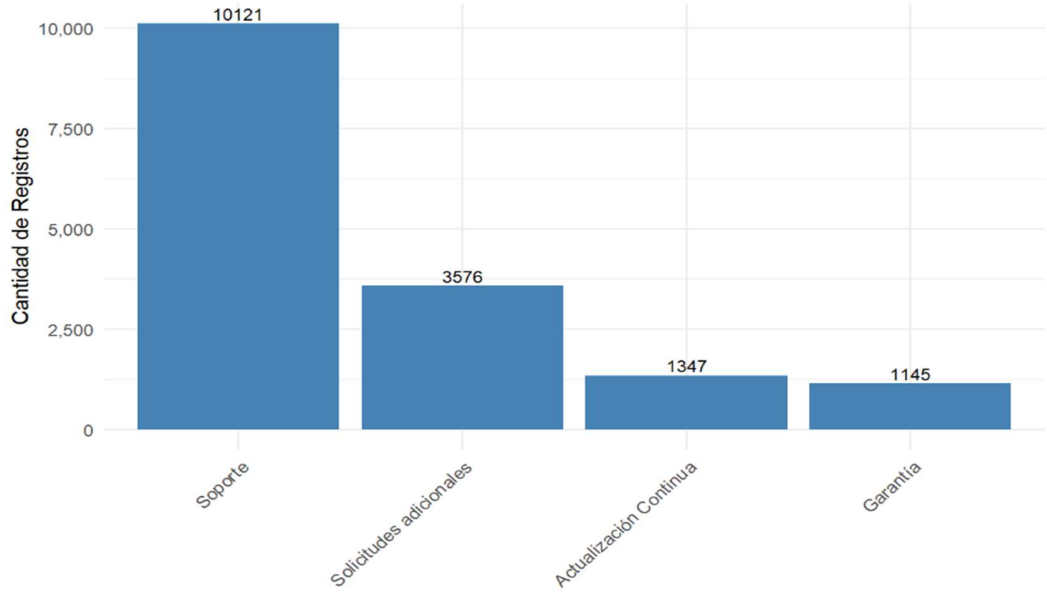

Ilustración 7 Cantidad de solicitudes recibidas por tipo de servicio, entre 2022 y 2024

En la **Ilustración 7**, se observa que el **servicio de soporte** encabeza con diferencia la cantidad de solicitudes recibidas, con un total de 10121 registros, lo que representa más del 50 % del total general. Este predominio sugiere una alta demanda sostenida en temas de soporte técnico, probablemente asociada a la participación activa de EEP durante el periodo analizado.

5.3 Análisis Exploratorio de Horas Trabajadas

Para el análisis exploratorio de las horas trabajadas asociadas a cada solicitud, se consideró un total de **17,496** registros de actividades correspondientes al período comprendido entre enero de 2022 y diciembre de 2024.

Este conjunto de datos proporciona una base sólida para comprender cómo se distribuye el trabajo en las diferentes solicitudes a lo largo del tiempo. El análisis se abordó desde múltiples perspectivas con el fin de identificar patrones relevantes, comportamientos atípicos, tendencias temporales y diferencias según variables como el mes, el tipo de requerimiento, la empresa prestadora del servicio y la carga de trabajo por colaborador.

Al examinar estos datos, se observan variaciones significativas en las horas dedicadas por solicitud, lo que sugiere la existencia de múltiples factores que inciden en la duración del tiempo de atención. Estas diferencias podrían estar relacionadas tanto con la naturaleza técnica de las tareas como con aspectos operativos y logísticos asociados a cada empresa o región.

Este análisis exploratorio constituye una etapa clave en la comprensión del comportamiento general del sistema de atención, y sirve como insumo fundamental para la posterior construcción de modelos predictivos más robustos y precisos.

5.3.1 Tendencia por Mes y Año

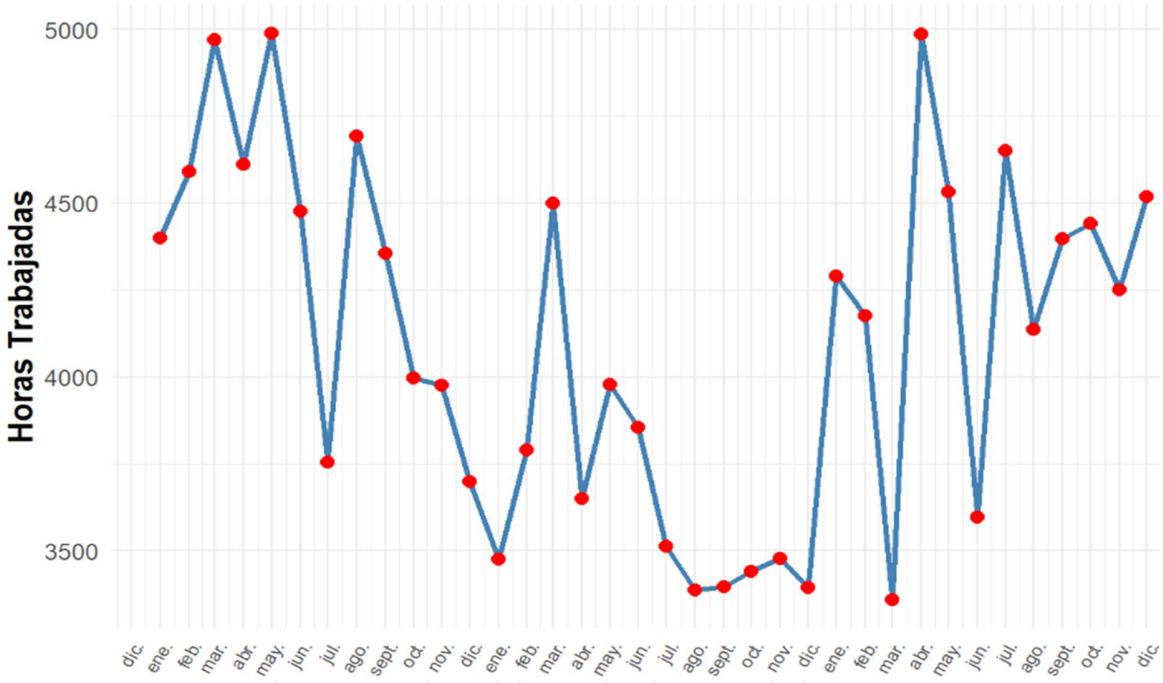

Ilustración 8 tendencia de horas trabajadas por mes desde 2022 a 2024

La **ilustración 8** presenta la evolución de las horas trabajadas por mes, en el sistema comercial de ACTSIS Ltda., durante un periodo de análisis continuo de tres años. En el gráfico se observa una considerable variabilidad mensual, con valores que oscilan entre aproximadamente 3.300 y 5.000 horas trabajadas por mes. Esta fluctuación evidencia la existencia de ciclos de alta y baja intensidad laboral que pueden estar asociados tanto a factores internos de la organización como a dinámicas externas de mercado.

Se identifican varios picos de actividad, particularmente en los meses de febrero, abril y marzo,

donde se alcanzan los valores más altos de horas trabajadas, superando en algunos casos las 4.900 horas. Estos periodos podrían coincidir con temporadas de alta demanda o cierres de trimestre que requieren mayor dedicación operativa. Por otro lado, los meses de diciembre, agosto y marzo (de un segundo año) muestran caídas pronunciadas por debajo de las 3.600 horas, lo cual puede deberse a periodos vacacionales, reducción de operaciones o ajustes administrativos.

El comportamiento general sugiere cierta estacionalidad, especialmente con tendencias repetidas de disminución en diciembre, lo cual es consistente con el cierre del año fiscal y las vacaciones colectivas. Asimismo, se observa una recuperación progresiva en los últimos meses analizados, que podría interpretarse como una estabilización o reactivación de la actividad laboral.

En términos de gestión operativa, esta información es clave para la planificación de recursos humanos y logísticos, ya que permite anticipar los periodos de mayor carga laboral y optimizar la asignación del personal disponible. Una estrategia basada en este análisis podría contribuir a una mayor eficiencia operativa, reducción de costos y mejora en la calidad del servicio ofrecido.

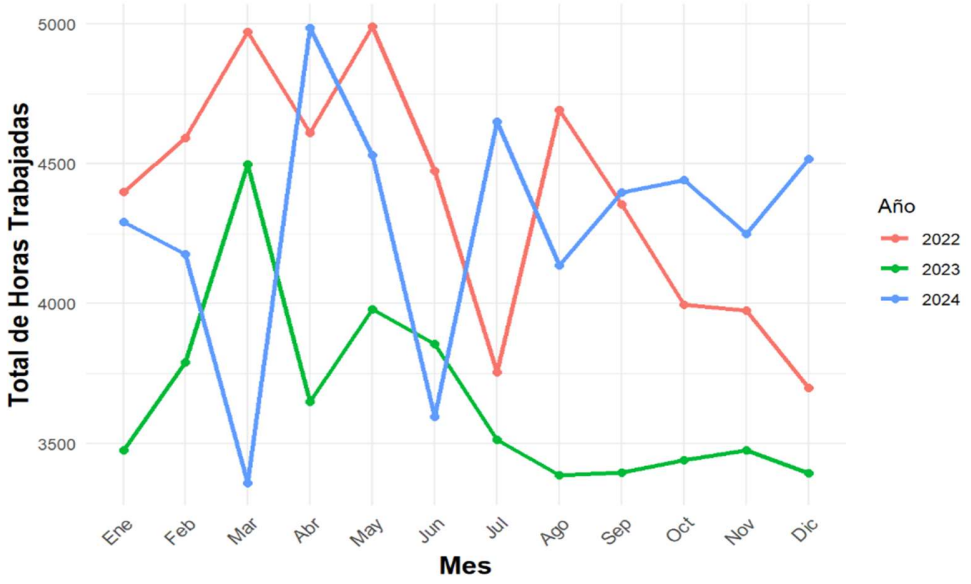

Ilustración 9 tendencias anuales de horas trabajadas mensualmente entre 2022 y 2024

La **ilustración 9** muestra la evolución mensual del total de horas trabajadas durante los años 2022, 2023 y 2024, permitiendo identificar patrones estacionales y comparativos relevantes entre los tres periodos.

En el año 2022 se observa un comportamiento claramente estacional, con un aumento progresivo

desde enero hasta marzo. Posteriormente, se evidencia una disminución casi sostenida en la carga laboral hasta diciembre.

El año 2023, presenta el mismo crecimiento estacional en el primer trimestre del año, siendo marzo el mes con mayor número de horas trabajadas en el año. A partir de abril se produce una disminución progresiva casi constante. Es de resaltar que en esta vigencia, los promedios de horas trabajadas fueron considerablemente más bajos que en los otros dos años analizados.

Por su parte, el año 2024 si bien presenta algunas estacionalidades similares al 2022, evidencia una tendencia más fluctuante en el primer y último trimestre del año. Sin embargo esto está justificado en una serie de regulaciones de FACTURACIÓN ELECTRÓNICA, que modificaron el comportamiento normal de la demanda en esos periodos.

En términos generales, los meses de marzo y abril se perfilan como los de mayor carga laboral de forma consistente en los tres años, lo que podría estar relacionado con ciclos naturales de desarrollo de proyectos. Por otro lado, los meses de julio y agosto, y el último trimestre del año, tienden a registrar una disminución en la carga de trabajo, lo cual podría asociarse a periodos vacacionales o a una menor demanda operativa.

Este análisis permite comprender mejor los patrones temporales del volumen de trabajo en ACTSIS LTDA y constituye un insumo importante para la planificación futura de recursos humanos, así como para la mejora de la eficiencia operativa.

5.3.2 Tendencia por Empresa

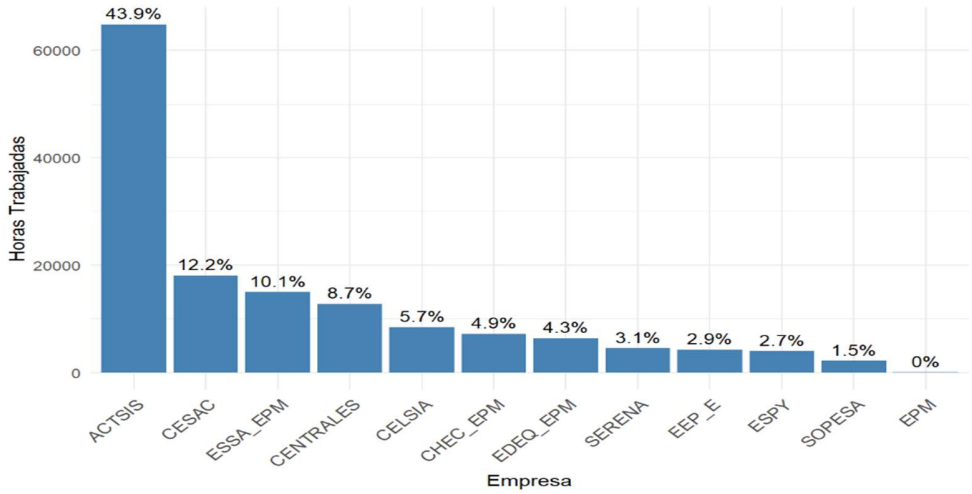

Ilustración 10 horas trabajadas por empresa cliente, entre 2022 y 2024

La **ilustración 10** presenta la distribución del total de horas trabajadas por empresa, permitiendo identificar la participación relativa de cada cliente en la carga operativa del sistema durante el periodo analizado.

Se destaca que ACTSIS concentra el mayor volumen de horas trabajadas, con un 43.9% del total, lo que indica el tiempo dedicado a la evolución del producto (roadmap) y la atención del producto no conforme (garantías).

En segundo lugar, se ubican CESAC (12.2%) y ESSA_EPM (10.1%), seguidos por CENTRALES (8.7%) y CELSIA (5.7%). Estas empresas, aunque con menor participación representan actores importantes en la distribución del trabajo y permiten diversificar en cierta medida la carga operativa.

El resto de las empresas, como CHEC_EPM (4.9%), EDEQ_EPM (4.3%) y SERENA (3.1%), presentan una participación moderada, mientras que otras como SOPESA, ESPY y EPM tienen una incidencia mínima (inferior al 3%) o incluso nula en el caso de EPM, lo que se debe en parte al modelo de contratación que tienen esas empresas.

Este análisis resulta fundamental para comprender el perfil de uso del sistema comercial SAC, ya que permite enfocar los esfuerzos de mantenimiento, escalabilidad y soporte técnico hacia los usuarios con mayor impacto operativo. Asimismo, la información puede ser útil para la toma de decisiones en términos de segmentación de clientes, optimización de recursos y análisis de riesgos relacionados con la dependencia de ciertos actores clave.

5.3.3 Tendencia por Tipo de Servicio

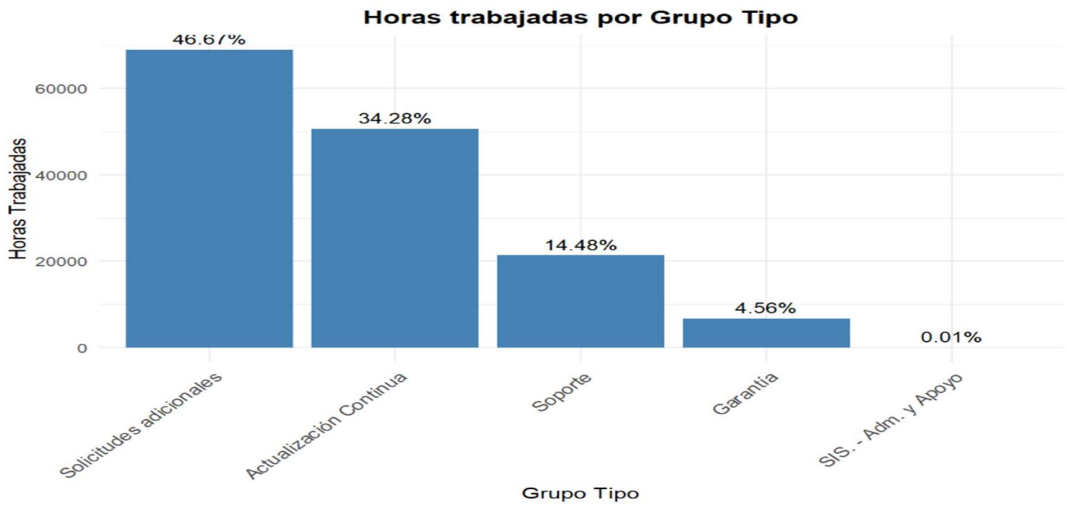

Ilustración 11 horas trabajadas por tipo de solicitud, entre 2022 y 2024

La **ilustración 11** evidencia la distribución de las horas trabajadas según el tipo de requerimiento atendido, categorizada en cinco grupos: Solicitudes Adicionales, Actualización Continua, Soporte, Garantía, y Sistemas – Administración y Apoyo.

Se observa que el grupo "Solicitudes Adicionales" representa la mayor proporción del tiempo trabajado, con un 46.67%. Este hallazgo indica que casi la mitad del esfuerzo operativo del equipo técnico está orientado a atender requerimientos fuera del alcance original del sistema, lo que puede reflejar un dinamismo significativo en las necesidades del cliente o una evolución constante en los requerimientos del negocio.

Le sigue el grupo "Actualización Continua", con un 34.28%, lo cual confirma el compromiso con la mejora progresiva del sistema a través de actividades de mantenimiento evolutivo, implementación de nuevas funcionalidades y adaptación a cambios normativos o tecnológicos. Esta proporción refuerza la importancia de mantener el sistema actualizado como una prioridad estratégica.

El grupo "Soporte" representa un 14.48% del total, lo cual es coherente con la necesidad constante de asistencia operativa, resolución de incidencias y acompañamiento al usuario. Aunque es una proporción menor, sigue siendo significativa en términos de recursos asignados.

Por otro lado, los grupos "Garantía" (4.56%) y "Sistemas – Administración y Apoyo" (0.01%) tienen una participación mucho más reducida. El escaso porcentaje de actividades asociadas a "Garantía" podría interpretarse positivamente, como indicio de una baja tasa de errores post-despliegue. En cuanto al grupo de "Sistemas – Administración y Apoyo", su mínima participación sugiere que las tareas administrativas o de soporte interno no consumen recursos significativos dentro del equipo analizado.

Este análisis permite identificar los principales focos de esfuerzo del equipo técnico y orienta la toma de decisiones hacia una mejor asignación de recursos, priorización de actividades y posibles oportunidades de automatización o mejora continua.

3.2 Consolidación de bases y Selección de variables para el modelo

3.2.1 Variables seleccionadas

Realizada la exploración de las dos bases de datos, para el desarrollo del modelo predictivo, se procede con la consolidación de estas para constituir una base definitiva con los atributos a trabajar en el modelo para capturar la dinámica de la demanda de servicios.

- **Fecha (mes):** Variable de fecha que permite establecer la temporalidad de los registros y facilita el análisis de tendencias y estacionalidades a lo largo del tiempo.
- **Empresa Cliente:** Variable categórica que identifica a cada cliente. Al segmentar la demanda por empresa, se pueden identificar patrones específicos y comportamientos de solicitud que varían según el cliente.
- **Tipo de solicitud (requerimiento):** Variable categórica que clasifica los diferentes tipos de servicios prestados. Esta información es esencial para entender las variaciones en la demanda según el tipo de servicio que se está solicitando.
- **Solicitudes trabajadas:** cantidad de solicitudes trabajadas por ACTSIS.
- **Horas trabajadas:** horas de atención dedicadas a cada solicitud.

3.2.2 Base definitiva

Como resultado del proceso de integración y limpieza de las fuentes de datos disponibles, se consolidó una base definitiva conformada por **3.548 registros** y **5 atributos**. Esta base resume la información histórica mensual de atención de solicitudes por tipo de requerimiento y cliente, y contiene tanto variables categóricas como numéricas, lo que permite una representación adecuada de la dinámica de la demanda de servicios.

La estructura resultante permitió una segmentación significativa por cliente y categoría de solicitud, así como un análisis temporal de volúmenes y tiempos de atención, sirviendo como insumo principal para el entrenamiento y validación del modelo predictivo, asegurando coherencia, trazabilidad y relevancia analítica en el proceso de modelado.

MES_ATENCION	CLIENTE	TIPO_DE_RQ	SOLICITUDES_TRABAJADAS	HORAS_TRABAJADAS
2022-02-01	ACTSIS	ACT	85	1342.4
2022-02-01	ACTSIS	ADS	1	15.7
2022-02-01	ACTSIS	CAP	9	396.3
2022-02-01	ACTSIS	GT3	6	42.2
2022-02-01	ACTSIS	INS	8	39.1
2022-02-01	ACTSIS	LEY	3	9.2
2022-02-01	ACTSIS	PRU	7	164.0
2022-02-01	ACTSIS	SP1	1	12.1
2022-02-01	CELSIA	ACT	1	2.1
2022-02-01	CELSIA	ADD	9	60.5
2022-02-01	CELSIA	AFN	2	57.9

Tabla 3. Muestra de la base de datos consolidada

Para la construcción de los modelos predictivos, se definieron tanto la variable objetivo como las variables predictoras de la siguiente manera:

- **Variable dependiente (objetivo):**
  - **HORAS_TRABAJADAS**: Representa la cantidad total de horas dedicadas mensualmente a la atención de solicitudes. Esta variable es continua y constituye el valor que se desea predecir a partir de las demás variables disponibles.

- **Variables independientes (predictoras):**
  - **MES_ATENCION**: Corresponde a la fecha del mes de atención, que puede ser transformada en una variable temporal (por ejemplo, como un índice cronológico o utilizando componentes de fecha como el mes o el año).
  - **CLIENTE**: Variable categórica que identifica a la empresa a la cual se le brindó atención en el mes.
  - **TIPO_DE_RQ**: Variable categórica que clasifica el tipo de requerimiento o solicitud atendida.
  - **SOLICITUDES_TRABAJADAS**: Variable numérica que indica la cantidad de requerimientos gestionados durante el período correspondiente.

Esta selección de variables permitió capturar tanto aspectos cuantitativos como cualitativos del comportamiento mensual de trabajo, facilitando así una predicción más robusta de las horas requeridas para atender los distintos tipos de solicitudes.

3.2.2.1 Evaluación de la relación de variables

Con el fin de comprender mejor los factores que influyen en el tiempo de trabajo dedicado a las solicitudes, se evaluó la relación entre la variable dependiente HORAS_TRABAJADAS y las variables independientes TIPO_DE_RQ, CLIENTE y SOLICITUDES_TRABAJADAS.

3.2.2.2 Comprobación de normalidad y elección de prueba estadística

Inicialmente se aplicó un análisis de varianza (ANOVA) para determinar si existían diferencias significativas en HORAS_TRABAJADAS en función de las variables categóricas TIPO_DE_RQ y CLIENTE. No obstante, se evaluó la normalidad de los residuos mediante la prueba de Shapiro-Wilk, la cual arrojó los siguientes resultados:

- Para TIPO_DE_RQ: W = 0.45567, p-value < 2.2e-16
- Para CLIENTE: W = 0.45932, p-value < 2.2e-16

Ambos valores p indican que los residuos no siguen una distribución normal, incumpliendo uno de los supuestos fundamentales del ANOVA. En consecuencia, se optó por utilizar la prueba no paramétrica de Kruskal-Wallis, la cual no requiere el supuesto de normalidad.

3.2.2.3 Resultados del test de Kruskal-Wallis

- HORAS_TRABAJADAS ~ TIPO_DE_RQ: $\chi^2(24) = 873.66$, $p < 2.2e\text{-}16$

  Este resultado indica que existen diferencias significativas en las horas trabajadas según el tipo de requerimiento. Es decir, no todos los tipos de solicitudes demandan la misma cantidad de esfuerzo.

- HORAS_TRABAJADAS ~ CLIENTE: $\chi^2(11) = 548.01$, $p < 2.2e\text{-}16$
  Esto sugiere que algunos clientes requieren significativamente más o menos horas de trabajo que otros.

Ambos resultados permiten concluir que tanto TIPO_DE_RQ como CLIENTE son variables significativamente asociadas a las HORAS_TRABAJADAS.

3.2.2.4 Asociación entre variables categóricas

Para evaluar la fuerza de asociación entre CLIENTE y TIPO_DE_RQ (ambas variables categóricas), se utilizó el coeficiente V de Cramer, obteniéndose un valor de 0.2267. Este resultado indica una asociación moderada: existe cierta relación entre el tipo de requerimiento y el cliente que lo solicita, pero no es lo suficientemente fuerte como para considerar que una variable depende completamente de la otra.

3.2.2.5 Correlación entre variables numéricas

Se analizó la relación entre las variables numéricas SOLICITUDES_TRABAJADAS y HORAS_TRABAJADAS, mediante una matriz de correlación correspondiente a la **Ilustración 12**.

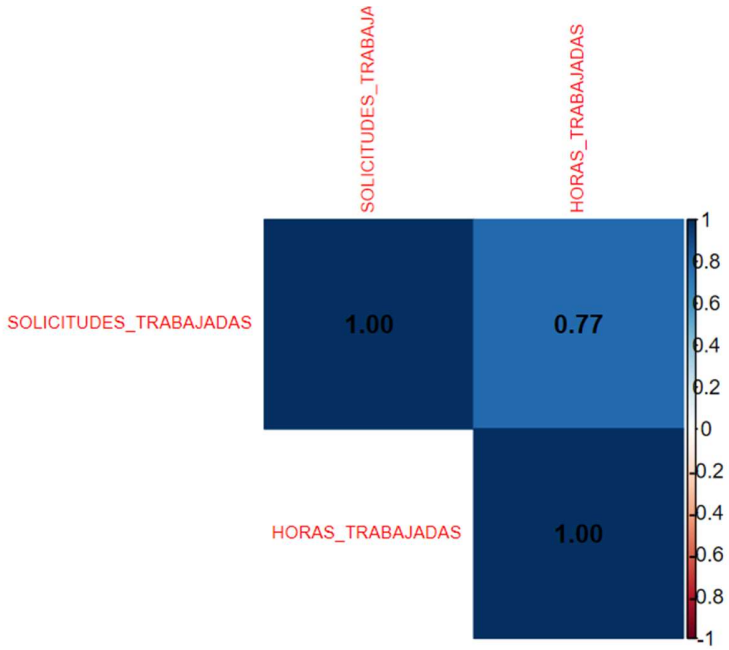

Ilustración 12.correlación entre solicitudes trabajadas y horas trabajadas

La matriz reveló una **correlación positiva fuerte (0.77) entre ambas variables**, lo cual indica que a medida que aumenta el número de solicitudes, también se incrementan las horas trabajadas. Sin embargo, esta relación no es perfecta, lo que sugiere:

- Algunas solicitudes pueden requerir más o menos tiempo que otras.
- Existe variabilidad en la duración de cada solicitud, lo cual es coherente con entornos de trabajo reales.

Esta correlación respalda conceptualmente el análisis, ya que un mayor volumen de trabajo suele

implicar una mayor demanda de tiempo. No obstante, el hecho de que la correlación no sea perfecta también resalta la importancia de considerar otras variables explicativas —como **CLIENTE** y **TIPO_DE_RQ**— en los modelos predictivos posteriores.

3.2.2.6 Verificación de valores nulos

Como parte del proceso de preparación de los datos, se evaluó la existencia de valores faltantes en las variables relevantes. Los resultados de esta verificación se presentan en la **Tabla 4**.

Variable	Valores nulos
MES_ATENCION	0
CLIENTE	0
TIPO_DE_RQ	0
SOLICITUDES_TRABAJADAS	0
HORAS_TRABAJADAS	0

Tabla 4. Verificación de valores nulos en las variables relevantes del conjunto de datos

Esto indica que los datos están completos y no requieren imputación, lo que fortalece la calidad de las conclusiones obtenidas.

3.2.3 División de los datos

Se separaron los datos en un conjunto de entrenamiento y un conjunto de prueba.

- Dimensión del conjunto de entrenamiento: 2839 6
- Dimensión del conjunto de prueba: 709 6

Se trabajó con un conjunto de datos consolidado compuesto por 2839 observaciones y 4 variables predictoras, las cuales incluyen tanto variables numéricas como categóricas, estas últimas debidamente codificadas como factores. La partición de los datos se realizó en un 80% para entrenamiento y un 20% para prueba.

6 DESARROLLO Y EVALUACIÓN DE MODELOS PREDICTIVOS

6.1 Modelos candidatos

Aunque el conjunto de datos analizado inicialmente presentaba una estructura temporal (registro mensual de horas trabajadas), se decidió no aplicar un modelo de series de tiempo clásico como ARIMA o Holt-Winters. Esta decisión se sustentó en varias razones técnicas y de contexto:

1. **Multicausalidad del fenómeno**:

   La demanda de servicios en ACTSIS Ltda. no depende exclusivamente de la evolución temporal, sino de múltiples factores categóricos y numéricos como el cliente, el tipo de requerimiento, la cantidad de solicitudes mensuales, entre otros. Los modelos clásicos de series de tiempo no capturan bien estas relaciones multivariadas.

2. **Presencia de estacionalidades atípicas e influencias externas**:

   Si bien se identificaron ciertos patrones estacionales, también se evidenciaron alteraciones provocadas por eventos regulatorios o contractuales. Esto reduce la eficacia de un modelo puramente basado en la periodicidad de la serie.

3. **Mayor capacidad explicativa y predictiva del aprendizaje automático**: Los modelos como Random Forest o XGBoost permiten incorporar variables exógenas (como tipo de servicio o cliente) y capturar relaciones no lineales entre estas y la variable objetivo, mejorando la precisión de la predicción.

4. **Validación empírica**:

   Durante la etapa exploratoria, se probó un enfoque preliminar de predicción con series de tiempo (no documentado en esta versión), pero los errores obtenidos fueron mayores que los alcanzados con los modelos de machine learning, lo que ratificó la decisión del cambio de enfoque.

Por tanto, si bien el análisis partió de un enfoque de series temporales, las características del problema y los objetivos del proyecto justificaron una transición metodológica hacia el aprendizaje automático multivariable supervisado.

Dado lo anterior y con el objetivo de predecir la variable **horas trabajadas por solicitud**, se consideraron distintos enfoques de aprendizaje supervisado, seleccionados por su capacidad

para manejar relaciones complejas entre múltiples variables. Los modelos candidatos fueron los siguientes:

- **Random Forest**: Este algoritmo de ensamble basado en árboles de decisión fue elegido por su capacidad para modelar relaciones no lineales, su robustez ante el sobreajuste y su buen desempeño en presencia de múltiples variables predictoras. Random Forest es especialmente útil cuando se busca capturar interacciones complejas y reducir la varianza mediante la agregación de múltiples árboles.
- **XGBoost (Extreme Gradient Boosting)**: Este modelo, también basado en árboles de decisión, fue considerado por su alta eficiencia computacional y su capacidad para realizar una optimización precisa mediante técnicas de regularización. XGBoost incorpora mecanismos avanzados para el manejo de ruido, outliers y datos multivariados, lo que lo convierte en una opción potente para tareas de predicción con estructuras complejas.
- **K-Vecinos más Cercanos (KNN)**: Se incluyó este modelo no paramétrico por su simplicidad y enfoque basado en la similitud entre observaciones. KNN permite realizar predicciones a partir de patrones locales en los datos, sin suponer una forma funcional específica entre las variables. Aunque puede ser sensible a la escala de los datos y al número de vecinos seleccionados, su aplicación permite establecer una referencia basada en relaciones directas de proximidad.

Estos modelos fueron seleccionados por su complementariedad metodológica y su capacidad para abordar la predicción de variables continuas en contextos con múltiples atributos explicativos. Su implementación permitió explorar tanto enfoques basados en aprendizaje estadístico como técnicas más sofisticadas de machine learning.

.

6.2 Desarrollo de modelos

6.2.1 Modelo de Bosques Aleatorios (Random Forest)

Para modelar la variable dependiente HORAS_TRABAJADAS, se implementó un modelo de aprendizaje automático supervisado mediante el algoritmo de Random Forest. Este método consiste en construir múltiples árboles de decisión a partir de subconjuntos aleatorios del conjunto de entrenamiento, y promediar sus predicciones para mejorar la precisión y controlar el sobreajuste.

Para optimizar el modelo, se utilizó validación cruzada de 5 pliegues (5-fold cross-validation), lo cual permitió seleccionar el valor óptimo del hiperparámetro mtry (número de variables consideradas en cada división del árbol).

Random Forest construye múltiples árboles de decisión, y en cada división de cada árbol, selecciona aleatoriamente un subconjunto de variables entre todas las disponibles. Este valor es mtry.

- Un mtry bajo hace que los árboles sean más diversos entre sí, pero podría dejar fuera variables relevantes en ciertos splits.
- Un mtry alto reduce la aleatoriedad, lo que puede llevar a árboles más similares (y menor ganancia de agregación), pero con divisiones más informadas.

- **Resultados del Modelo**

Durante el proceso de ajuste, se probaron distintos valores de mtry, y se obtuvieron los siguientes resultados:

mtry	RMSE	R^2	MAE
2	81.70	0.818	32.90
19	41.31	0.908	15.35
37	40.60	0.909	15.15

Tabla 5 Valores mtry del modelo Random Forest

El valor óptimo encontrado fue mtry = 37, el cual fue utilizado para el modelo final. Al evaluar este modelo con el conjunto de prueba, se obtuvieron las siguientes métricas de desempeño:

- Error Cuadrático Medio (RMSE): 34.96
- Coeficiente de Determinación (R^2): 0.91
- Error Absoluto Medio (MAE): 14.45

- **Interpretación y Evaluación**

Los resultados obtenidos indican un excelente desempeño predictivo del modelo. El coeficiente de determinación ($R^2 = 0.91$) sugiere que el modelo es capaz de explicar el 91% de la variabilidad en las horas trabajadas, lo que representa un ajuste muy adecuado.

El RMSE de aproximadamente 35 horas indica que, en promedio, las predicciones se desvían de los valores reales en esa magnitud. Dado que el rango de la variable dependiente es considerable, este error puede considerarse aceptable. Además, el MAE de 14.45 horas refuerza la precisión del modelo, ya que mide directamente la magnitud del error promedio en unidades comprensibles.

En resumen, el modelo de Random Forest se presenta como una herramienta robusta y precisa para la predicción de las horas trabajadas, mostrando un equilibrio entre complejidad y capacidad explicativa.

6.2.2 Modelo de Regresión con XGBoost

Con el fin de predecir la variable dependiente HORAS_TRABAJADAS, se entrenó un segundo modelo utilizando el algoritmo eXtreme Gradient Boosting (XGBoost), una técnica basada en árboles de decisión que emplea el principio de boosting para construir un modelo predictivo robusto a partir de múltiples árboles débiles.

- **Preprocesamiento y configuración**

Previo al entrenamiento del modelo, se realizó la conversión de variables categóricas a factores, asegurando así su adecuada interpretación por parte del algoritmo. Se dividió el conjunto de datos en dos subconjuntos: el 80% para entrenamiento y el 20% para prueba, garantizando que la variable respuesta estuviera adecuadamente representada en ambas particiones.

El modelo fue ajustado utilizando la función train() del paquete caret, aplicando validación cruzada de 5 pliegues para evitar el sobreajuste y evaluar su capacidad generalizadora. Los hiperparámetros del modelo fueron sintonizados parcialmente, fijando ciertos valores para simplificar el proceso inicial. Los valores finales utilizados fueron:

- nrounds = 50 (número de iteraciones de boosting),
- max_depth = 5 (profundidad máxima de cada árbol),
- eta = 0.3 (tasa de aprendizaje),
- colsample_bytree = 0.6 (porcentaje de columnas seleccionadas por árbol),
- subsample = 0.75 (porcentaje de observaciones usadas en cada iteración),
- gamma = 0 (reducción mínima de pérdida necesaria para realizar una partición),
- min_child_weight = 1 (peso mínimo de observaciones por nodo).

- **Resultados del modelo**

El desempeño del modelo XGBoost sobre el conjunto de prueba se evaluó mediante tres métricas comunes para tareas de regresión:

- RMSE (Root Mean Squared Error): 37.22
- R^2 (Coeficiente de determinación): 0.8988
- MAE (Mean Absolute Error): 16.41

Estos resultados indican que el modelo tiene un buen poder predictivo, explicando aproximadamente el 89.9% de la variabilidad en la variable objetivo. El error medio absoluto de 16.41 horas sugiere una precisión razonable en la predicción de tiempos de trabajo.

6.2.3 K-Vecinos más Cercanos (KNN)

El modelo de K-Vecinos más cercanos (KNN) fue entrenado para predecir las horas trabajadas mensualmente por cliente y tipo de requerimiento, utilizando validación cruzada de 5 pliegues. Durante el proceso de entrenamiento se evaluaron distintos valores de k, seleccionándose el valor óptimo de k = 5 por presentar el menor error cuadrático medio (RMSE).

El rendimiento del modelo con este valor fue el siguiente:

- RMSE: 44.43
- R^2 (coeficiente de determinación): 0.879

- MAE (Error absoluto medio): 16.99

Estos resultados indican que el modelo logra explicar aproximadamente el 87.9% de la variabilidad en las horas trabajadas, con un error absoluto promedio cercano a 17 horas por observación.

6.1 Comparación de Modelos

Una vez implementados los tres enfoques de aprendizaje supervisado, y con el propósito de identificar el modelo más adecuado para predecir la variable *HORAS_TRABAJADAS*, se procedió a su evaluación utilizando tres métricas estándar en tareas de regresión: **RMSE** (Root Mean Squared Error), **R^2** (Coeficiente de Determinación) y **MAE** (Mean Absolute Error). Estas métricas permiten valorar el desempeño de los modelos desde distintas perspectivas: dispersión del error, capacidad explicativa y magnitud promedio del error, respectivamente.

- **Coeficiente de Determinación (R^2):** Indica la proporción de la variabilidad total de la variable objetivo que es explicada por el modelo. Un valor cercano a 1 sugiere un buen ajuste, es decir, que las predicciones se aproximan bien a los valores observados.
- **Error Cuadrático Medio (RMSE):** Evalúa la magnitud promedio del error al cuadrado entre los valores predichos y los reales. Al penalizar de forma más severa los errores grandes, esta métrica es útil para detectar predicciones inestables o poco precisas.
- **Error Absoluto Medio (MAE):** Mide el promedio de las diferencias absolutas entre las predicciones y los valores reales. A diferencia del RMSE, el MAE no penaliza en exceso los errores grandes y se expresa en las mismas unidades que la variable objetivo (en este caso, horas trabajadas), lo que facilita su interpretación. Un MAE más bajo indica una mayor precisión del modelo, al reflejar directamente el error promedio sin importar su dirección.

Modelo	RMSE (prueba)	R^2 (prueba)	MAE (prueba)	Notas relevantes
Random Forest	34.96	0.91	14.45	Mejor desempeño general, alta precisión.
XGBoost	37.22	0.899	16.41	Buen desempeño, ligero subajuste respecto a RF

Modelo	RMSE (prueba)	R^2 (prueba)	MAE (prueba)	Notas relevantes
K-Vecinos (KNN)	44.43	0.879	16.99	Alto RMSE, buen R^2, modelo simple.

Tabla 6 comparativo de métricas de los tres modelos

A partir de la **tabla 6** se concluye que el modelo **Random Forest** demostró el mejor desempeño global entre los modelos evaluados. Obtuvo el menor error cuadrático medio (**RMSE = 34.96**) y el mayor coeficiente de determinación ($R^2 = 0.91$), lo que evidencia una excelente capacidad para explicar la variabilidad de la variable objetivo y una alta precisión en las predicciones. Además, el valor más bajo de MAE (14.45) refuerza su eficacia al reflejar un error promedio reducido en unidades de tiempo. Su arquitectura basada en múltiples árboles y la aleatoriedad en la selección de variables permiten capturar relaciones complejas y reducir el riesgo de sobreajuste, consolidándose como un modelo robusto y confiable.

XGBoost, si bien presentó un rendimiento levemente inferior, se comportó de manera competitiva. Con un R^2 de 0.899 y un MAE de 16.41, mostró una buena capacidad predictiva. No obstante, su RMSE de 37.22 sugiere un ligero subajuste, posiblemente atribuible a una configuración conservadora de hiperparámetros durante el entrenamiento. A pesar de ello, su estructura basada en boosting secuencial le permite detectar interacciones complejas entre variables, haciéndolo una alternativa sólida y eficiente.

K-Vecinos más Cercanos (KNN) obtuvo el desempeño más limitado en comparación con los modelos anteriores. Aunque logró un R^2 de 0.879, su RMSE (44.43) y MAE (16.99) fueron los más altos del análisis, indicando una mayor dispersión en las predicciones. Si bien este modelo destaca por su simplicidad e interpretabilidad, su rendimiento se ve afectado por la sensibilidad al valor de *k* y la ausencia de una estructura de aprendizaje explícito, lo cual puede limitar su eficacia en contextos con múltiples variables y relaciones no lineales.

En conjunto, los resultados ponen en evidencia que los modelos de tipo ensemble, como Random Forest y XGBoost, ofrecen ventajas significativas en escenarios complejos con múltiples variables predictoras. Su capacidad para manejar no linealidades y minimizar errores los posiciona por encima de métodos más básicos como KNN. En función de su superior desempeño, **Random Forest fue seleccionado como el modelo principal para su implementación final**, dada su robustez, precisión y estabilidad en las predicciones.

6.2 Selección del modelo

Con base en la evaluación comparativa de desempeño, el modelo seleccionado para la predicción de la variable *HORAS_TRABAJADAS* fue **Random Forest**. Esta decisión se sustentó en su superior rendimiento en las tres métricas de evaluación consideradas: presentó el menor Error Cuadrático Medio (RMSE = 34.96), el mayor Coeficiente de Determinación (R^2 = 0.91) y el menor Error Absoluto Medio (MAE = 14.45).

Estos resultados reflejan no solo una alta precisión, sino también una excelente capacidad del modelo para explicar la variabilidad de la variable objetivo, superando tanto a XGBoost como a KNN en términos de exactitud y consistencia.

Además de su rendimiento cuantitativo, Random Forest ofrece ventajas cualitativas relevantes: es robusto ante valores atípicos, poco sensible a la multicolinealidad entre variables, y especialmente efectivo en contextos con múltiples características predictoras y relaciones no lineales. La naturaleza del algoritmo, que combina múltiples árboles de decisión mediante el ensamblaje (bagging), proporciona estabilidad y reduce el riesgo de sobreajuste, características fundamentales para un entorno de predicción operativa como el del presente estudio.

6.3 Aplicación del modelo

Una vez seleccionado el modelo de **Random Forest**, se procedió a su aplicación sobre el conjunto de datos, con el objetivo de generar predicciones de horas trabajadas por solicitud en función de diversas características: tipo de requerimiento, cliente, mes, cantidad de usuarios activos, entre otras variables relevantes.

El modelo fue entrenado utilizando el conjunto de entrenamiento y posteriormente validado con el conjunto de prueba, lo que permitió evaluar su capacidad predictiva en datos no vistos. A partir de este modelo final, se generaron estimaciones para solicitudes nuevas y también se analizaron los errores residuales con el propósito de identificar patrones o casos atípicos en las predicciones.

- **Predicciones vs valores reales**

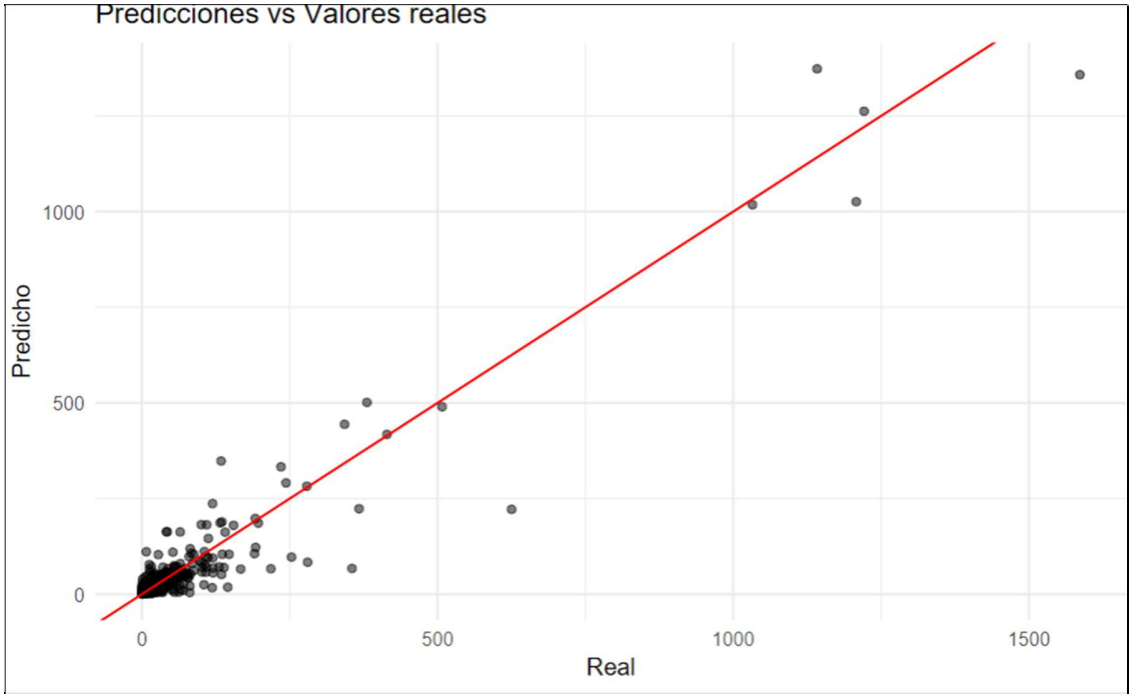

Ilustración 13 Gráfica de dispersión: valores reales vs. valores predichos para la evaluación del modelo

Para complementar la evaluación cuantitativa del modelo, se construyó una **ilustración 13** correspondiente a la dispersión que compara los valores predichos frente a los valores reales de la variable *HORAS_TRABAJADAS*. En la ilustración 10 se observa la distribución de los puntos junto a una línea roja que representa la línea de identidad (y = x), es decir, el ideal en el que las predicciones coinciden exactamente con los valores reales.

En general, se aprecia que la mayoría de las observaciones se agrupan cerca de esta línea, lo que sugiere un buen nivel de ajuste por parte del modelo. No obstante, se identifican algunos valores atípicos con desviaciones significativas, especialmente en el rango de valores altos, lo que podría reflejar la presencia de casos extremos o una mayor variabilidad difícil de capturar con precisión. Esta visualización respalda los resultados obtenidos mediante métricas cuantitativas, al evidenciar que el modelo tiene un desempeño sólido en la predicción de la mayoría de los casos, aunque presenta ciertas limitaciones en escenarios menos frecuentes o más extremos.

- **Análisis Error Residual**

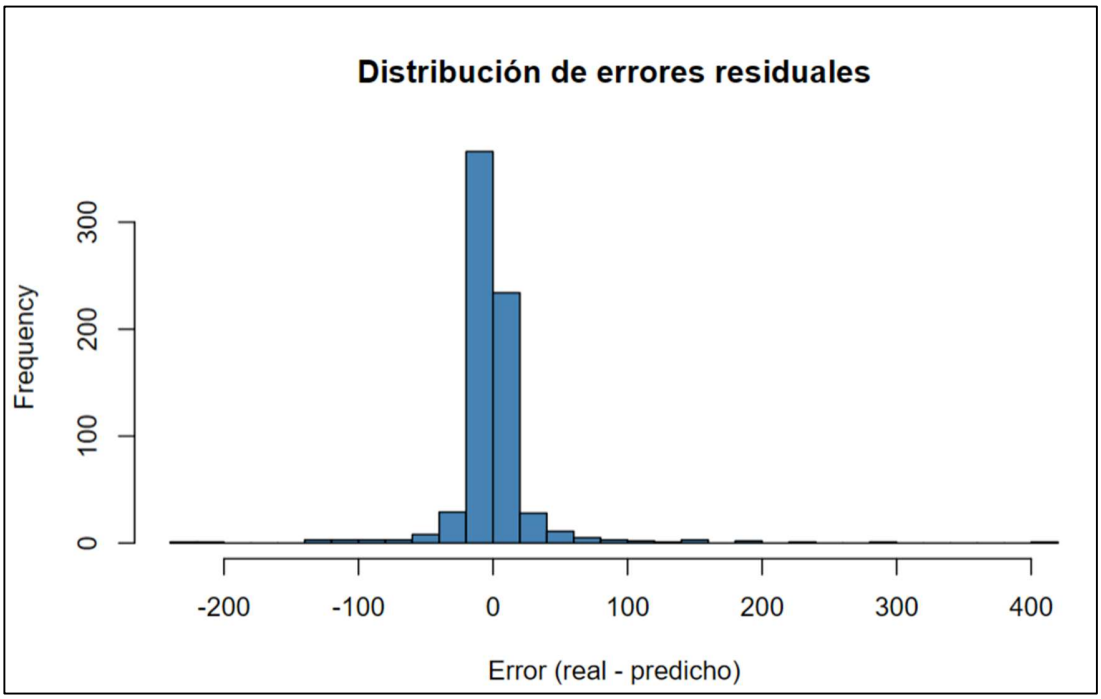

Ilustración 14 Distribución de error residual

La **Ilustración 14** presenta la distribución de los errores residuales obtenidos a partir de la aplicación del modelo de Random Forest sobre el conjunto de prueba. Estos residuos corresponden a la diferencia entre el valor real de las horas trabajadas y el valor predicho por el modelo.

Se observa una alta concentración de errores alrededor del valor cero, lo cual sugiere que el modelo presenta un buen ajuste general y logra estimar correctamente las horas requeridas en la mayoría de los casos. La forma de la distribución es aproximadamente simétrica, indicando una ausencia de sesgo sistemático. Es decir, el modelo no tiende a sobrestimar ni a subestimar de manera consistente los valores reales.

No obstante, también se identifica una ligera asimetría positiva, con algunos valores extremos hacia la derecha de la distribución. Esto indica que, en ciertos casos, el modelo subestima las horas requeridas, lo cual podría estar asociado a solicitudes de alta complejidad o con características particulares no completamente capturadas por las variables disponibles. A pesar de ello, la frecuencia de estos casos es baja, y no afecta significativamente el desempeño global del modelo. Este análisis confirma que los errores se distribuyen de manera controlada y que el modelo

mantiene una precisión aceptable, incluso en presencia de variabilidad. En consecuencia, los resultados refuerzan la validez del modelo como herramienta de apoyo para la toma de decisiones estratégicas en la planeación de recursos y asignación de cargas operativas dentro de ACTSIS Ltda.

- **Análisis del Desempeño del Modelo Predictivo según Tipo de Requerimiento y Cliente**

Para evaluar la precisión del modelo de predicción de horas trabajadas, se realizó un análisis del error promedio en función de dos factores clave: el tipo de requerimiento y el cliente. Este análisis permitió identificar las áreas donde el modelo presenta mejor desempeño y aquellas en las que se debe mejorar.

Se utilizaron dos métricas de error ampliamente reconocidas: el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE). Ambas proporcionan una medida cuantitativa de la diferencia entre las horas reales trabajadas y las horas predichas por el modelo, siendo el MAE una medida lineal y el RMSE que penaliza más los errores grandes.

Error Promedio por Tipo de Requerimiento

TIPO_DE_RQ	N	Horas_Reales_Medias	Horas_Predichas_Medias	MAE	RMSE
CAP	12	172.808333	130.757824	75.324532	132.257907
PRU	12	141.125000	120.313773	52.604849	72.583762
ADD	60	72.576667	72.328094	35.257489	59.677096
ACT	46	142.200000	139.035494	21.110285	55.994561
IMP	43	38.406977	38.954427	19.827849	33.552028
SP2	51	42.150980	52.909614	17.185555	39.351481
ADS	48	34.500000	34.568778	12.361139	16.514250
AFN	53	22.303774	19.785669	11.318188	19.000727
SP3	42	15.950000	13.671130	10.367442	23.236409
INS	32	20.303125	21.003197	9.499388	18.129201
SCR	40	16.935000	14.770426	8.775535	14.669988
GT2	16	17.875000	12.598223	8.561774	12.541670
LEY	31	11.712903	9.550287	8.351531	20.466028
GT3	26	13.084615	9.721356	7.049538	13.861093
BSG	68	18.407353	19.113100	6.550220	9.162790
SP1	39	8.984615	12.407607	5.482354	11.003520
ATC	27	8.162963	8.292588	4.729763	7.149692
ADL	15	6.786667	8.445547	4.435511	6.876185
GTA	17	8.694118	10.281438	4.434506	5.847385
SIS	1	1.100000	5.423777	4.323777	4.323777
MIG	2	25.450000	23.828127	4.215970	4.517176
ADE	23	5.565217	6.457807	3.312968	4.789794
DBA	5	7.020000	6.901343	2.421368	3.176519

Ilustración 15 error promedio por tipo de requerimiento

Al agrupar los datos por tipo de requerimiento, se calcularon las siguientes estadísticas:

- N: número de observaciones por tipo de requerimiento.
- Horas Reales Medias: promedio de horas efectivamente trabajadas.
- Horas Predichas Medias: promedio de horas estimadas por el modelo.
- MAE y RMSE: métricas de error para cada tipo.

En la **Ilustración 15** se evidencia que ciertos tipos de requerimiento presentan un MAE considerablemente mayor, lo cual indica que el modelo tiene dificultades para predecir con precisión en esos casos específicos. Esta diferencia en el desempeño podría atribuirse a una mayor variabilidad en las horas trabajadas, a la insuficiencia de datos representativos en ciertas categorías o a la presencia de características particulares que el modelo no logra capturar adecuadamente.

Error Promedio por Cliente

CLIENTE	N	MAE	RMSE
ACTSIS	48	51.613948	90.62591
CESAC	17	23.758932	39.23378
ESPY	46	17.794339	36.48169
CELSIA	83	16.469773	41.12589
ESSA_EPM	104	14.167016	30.14400
SERENA	52	12.416285	35.64241
CENTRALES	89	10.864183	18.26033
EEP_E	64	8.669326	16.37367
EDEQ_EPM	84	8.185514	12.66136
CHEC_EPM	81	7.325742	12.53641
SOPESA	41	6.215052	16.44911

Ilustración 16 error promedio por tipo de requerimiento

El análisis por cliente permitió evaluar el desempeño del modelo desde la perspectiva de cada organización beneficiaria de los servicios brindados por ACTSIS. Los resultados revelan diferencias sustanciales en la precisión de las predicciones según el cliente, como se muestra en la **Ilustración 16.**

En particular, el cliente ACTSIS presentó el mayor nivel de error, con un MAE de 51.61 horas y un RMSE de 90.63 horas, lo que indica una alta desviación entre las horas reales trabajadas y las estimadas por el modelo. Otros clientes con errores elevados incluyen CESAC (MAE: 23.76) y ESPY

(MAE: 17.79), lo cual sugiere una mayor variabilidad en la complejidad o características operativas de los requerimientos que gestionan.

Por el contrario, se observa un desempeño considerablemente más estable en clientes como SOPESA, CHEC_EPM y EDEQ_EPM, cuyos valores de MAE se mantienen por debajo de las 9 horas. Este comportamiento sugiere que el modelo logra una mayor precisión con organizaciones que presentan patrones de requerimiento más consistentes o predecibles.

Estos hallazgos son relevantes para orientar acciones de mejora del modelo predictivo. En particular, invitan a considerar estrategias de ajuste o calibración diferenciadas por cliente, o bien explorar enfoques de modelado personalizados que incorporen variables adicionales que reflejen la heterogeneidad operativa entre organizaciones. Este tipo de análisis no solo fortalece la interpretación de los resultados, sino que también contribuye a aumentar la confiabilidad del sistema en contextos reales de uso.

- **Error relativo según nivel de horas reales**

Con el fin de evaluar el desempeño del modelo de predicción en función del nivel de carga de trabajo, se agruparon los registros según rangos de horas reales trabajadas. La segmentación se realizó en cinco intervalos: 0–20, 21–50, 51–100, 101–200 y más de 200 horas. Para cada grupo, se calcularon dos métricas de error: el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE), como se evidencia en la **ilustración 17**.

Rango	N	MAE	RMSE
0-20	463	6.024371	10.75178
21-50	139	12.429844	20.50625
51-100	55	24.893948	31.79652
101-200	34	56.721587	70.49100
200+	18	134.928869	171.61928

Ilustración 17 error relativo por rangos de horas trabajadas

Los resultados obtenidos evidencian una relación directamente proporcional entre la magnitud de las horas trabajadas y el margen de error del modelo. Específicamente, a medida que los requerimientos implican una mayor dedicación temporal, el error de predicción se incrementa de forma significativa. Este comportamiento puede atribuirse a dos factores principales: la **alta variabilidad inherente a los casos de mayor duración** y su **baja frecuencia relativa** en los datos históricos, lo que reduce la capacidad del modelo para generalizar en dichos escenarios.

El modelo muestra su mejor desempeño en la predicción de requerimientos de baja carga, particularmente en el grupo de 0 a 20 horas, donde se concentra el mayor número de registros. En contraste, el error alcanza su punto máximo en el grupo de más de 200 horas (MAE > 130), lo que refleja una pérdida notable de precisión en tareas de alta demanda.

Este patrón sugiere que el modelo tiende a subestimar o sobreestimar de manera más pronunciada a medida que aumenta la complejidad de los requerimientos, lo que limita su capacidad de generalización en estos casos. Sin embargo, considerando lo evidenciado en el análisis exploratorio sobre la distribución de solicitudes por tipo de requerimiento (**Ilustración 4**), donde la mayoría del volumen histórico gestionado por ACTSIS Ltda. corresponde a casos de ***Soporte***, los cuales se caracterizan por tener una duración breve y concentrarse en los rangos inferiores de horas trabajadas, esta característica del modelo no tendría mayor afectación. Este grupo representa más del 65% de los registros analizados entre 2022 y 2024, lo que implica que el modelo ha sido entrenado mayoritariamente con este tipo de patrones y, como consecuencia, presenta un alto nivel de precisión al predecir este segmento de casos frecuentes y operativamente simples.

En este contexto, si bien se identifican limitaciones en la capacidad del modelo para predecir correctamente solicitudes de mayor duración —como garantías complejas, solicitudes adicionales o actualizaciones continuas—, estas representan una proporción minoritaria del total. Por tanto, su impacto en la precisión general del modelo es limitado, y **no compromete su utilidad práctica para la toma de decisiones cotidianas**. Lejos de ser una debilidad crítica, esta situación refleja un comportamiento esperable en modelos supervisados, cuya efectividad tiende a alinearse con los patrones más representados en los datos de entrenamiento.

Desde una perspectiva de gestión organizacional, estos hallazgos permiten extraer conclusiones relevantes:

- **Planificación táctica:** el modelo puede ser implementado con confianza para anticipar y distribuir la carga de trabajo diaria, especialmente en servicios de soporte, donde se concentra la demanda y donde el modelo ha demostrado mayor precisión.

- **Evaluación de casos especiales:** si bien los requerimientos de alta carga son menos frecuentes, podrían complementarse con estrategias específicas como revisiones manuales, reglas de negocio o modelos secundarios.

- **Ajustes futuros del modelo:** es viable considerar estrategias de segmentación o ponderación que fortalezcan el desempeño en escenarios de mayor complejidad, sin afectar la solidez lograda en el núcleo operativo.

En síntesis, la capacidad del modelo para predecir con exactitud los casos de mayor volumen operativo refuerza su aplicabilidad como herramienta estratégica en la planificación de recursos de ACTSIS Ltda., mientras que las oportunidades de mejora en segmentos menos frecuentes abren camino a futuras optimizaciones.

6.4 Limitaciones del modelo

Si bien el modelo de Bosques Aleatorios (Random Forest) demostró un desempeño sobresaliente en la predicción de horas trabajadas, es importante reconocer las limitaciones inherentes a su implementación y aplicación en el contexto operativo de ACTSIS Ltda.

- **Sensibilidad a la distribución de los datos:** El modelo fue entrenado principalmente con casos de soporte de corta duración, los cuales representan más del 65% del total de registros históricos. Esta alta concentración sesga al modelo hacia patrones comunes de baja complejidad, reduciendo su capacidad de generalización frente a casos menos frecuentes, como solicitudes de actualización continua y de ley, o requerimientos complejos que demandan un número elevado de horas.

- **Desempeño reducido en casos extremos:** Los análisis por tipo de requerimiento y por cliente evidenciaron que el error de predicción se incrementa significativamente en escenarios con altos volúmenes de horas trabajadas. Por ejemplo, en los casos que superan las 200 horas mensuales, el MAE supera las 130 horas. Este comportamiento limita la confiabilidad del modelo para gestionar tareas atípicas o de alta complejidad.

- **Ausencia de variables adicionales relevantes:** El modelo fue desarrollado con un conjunto acotado de variables (cliente, tipo de requerimiento, mes y número de solicitudes trabajadas). Variables exógenas como el número de usuarios activos, cambios regulatorios o eventos externos (nuevos clientes) no fueron incluidas por falta de disponibilidad o dificultad en su medición, lo que limita la capacidad del modelo para capturar efectos contextuales relevantes.

- **Estacionalidad y comportamiento futuro:** El modelo se construyó sobre datos históricos comprendidos entre 2022 y 2024. Si bien se identificaron ciertos patrones estacionales, no se garantiza que estos se mantengan en el futuro, sobre todo ante cambios regulatorios o tecnológicos. Por tanto, el modelo requiere ajustes y recalibraciones periódicas para conservar su vigencia.

- **Dependencia de la calidad de los datos:** Aunque se realizó un proceso riguroso de limpieza y consolidación de los datos, cualquier error en la captura, clasificación o codificación de los requerimientos podría afectar la precisión de las predicciones. Esto es especialmente relevante si se considera que algunas categorías de análisis fueron excluidas por inconsistencia en los registros, como el atributo "Módulo".

En resumen, aunque el modelo de Random Forest proporciona una herramienta robusta y confiable para la planificación operativa en ACTSIS Ltda., su aplicabilidad está condicionada a los patrones observados en los datos históricos. Para maximizar su valor, es pertinente complementarlo con modelos especializados para casos complejos, e implementar procesos continuos de actualización y validación.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

El presente trabajo tuvo como propósito el desarrollo de un modelo predictivo para estimar las horas trabajadas en la atención de solicitudes del Sistema Comercial (SAC) de ACTSIS Ltda., con el fin de mejorar la planificación de la capacidad instalada y optimizar la asignación de recursos.

A partir de un riguroso análisis exploratorio de datos históricos entre enero de 2022 y diciembre de 2024, se identificaron variables clave que inciden en la demanda de servicios, tales como el tipo de requerimiento, el cliente y la cantidad de solicitudes mensuales. Se consolidó una base de datos robusta y limpia, sobre la cual se entrenaron y evaluaron distintos algoritmos de aprendizaje supervisado.

Entre los hallazgos más relevantes se destacan los siguientes:

- Se identificaron patrones de comportamiento en la demanda de horas trabajadas a través del análisis exploratorio de más de 17,000 registros, lo que permitió comprender la distribución, estacionalidad y concentración de solicitudes por cliente, tipo de requerimiento y periodo.
- Se implementaron y evaluaron tres modelos de regresión supervisada: **Random Forest**, **XGBoost** y **K-Vecinos más Cercanos (KNN)**. El desempeño de cada uno fue evaluado mediante métricas estándar como RMSE, R^2 y MAE, lo que permitió una comparación objetiva y fundamentada.
- El modelo de **Random Forest** se posicionó como la mejor alternativa para la predicción de la variable HORAS_TRABAJADAS, gracias a su alta capacidad explicativa ($R^2 = 0.91$) y sus bajos niveles de error. Su solidez frente al sobreajuste y su capacidad para manejar relaciones no lineales lo convierten en una herramienta adecuada para entornos operativos complejos.
- La implementación del modelo seleccionado permite estimaciones confiables del esfuerzo requerido por solicitud, lo que habilita mejoras en la **planificación operativa**, **asignación de recursos**, y eventualmente, en la **definición de acuerdos de nivel de servicio (ANS)** más realistas.

Se demostró la viabilidad técnica de aplicar **modelos de machine learning** al análisis de operaciones en servicios tecnológicos, aportando valor tanto en la automatización de estimaciones como en el soporte a la toma de decisiones estratégicas.

En definitiva, este modelo constituye una herramienta valiosa para anticipar la carga de trabajo y facilitar la toma de decisiones estratégicas en la gestión operativa de ACTSIS. Su aplicación permite prevenir la sobrecarga de personal, reducir tiempos de respuesta y evitar incumplimientos en los acuerdos de nivel de servicio, contribuyendo así a una mayor eficiencia y calidad del servicio.

No obstante, es importante reconocer las limitaciones identificadas durante el desarrollo del modelo. En particular, se evidenció un desempeño inferior en casos de requerimientos complejos o atípicos, así como una fuerte dependencia del modelo respecto a los patrones dominantes en los datos históricos (principalmente solicitudes de soporte). Además, la exclusión de variables contextuales —como regulaciones externas, cambios tecnológicos o carga operativa no registrada— limita la capacidad del modelo para adaptarse a condiciones extraordinarias. También se identificó una limitada capacidad de interpretabilidad del modelo, característica inherente a técnicas de tipo "ensemble" como Random Forest.

A pesar de estas restricciones, el modelo demuestra ser una solución robusta para el núcleo operativo de ACTSIS, especialmente en tareas de soporte, que representan el grueso de la demanda.

7.2 TRABAJOS FUTUROS

Si bien los resultados obtenidos son satisfactorios y demuestran el valor del enfoque predictivo aplicado, el proyecto deja abiertas múltiples oportunidades de mejora, expansión y consolidación, tanto desde una perspectiva metodológica como desde su aplicación práctica en contextos organizacionales más amplios.

En este sentido, se proponen a continuación diversas líneas de trabajo futuro que podrían enriquecer y extender los aportes logrados:

- **Ajuste avanzado de hiperparámetros**: Aunque los modelos empleados muestran buen desempeño, es pertinente profundizar en la optimización de sus configuraciones, con el fin de maximizar su precisión y capacidad de generalización.
- **Incorporación de nuevas variables predictoras**: Incorporar nuevas variables exógenas, como número de usuarios activos por cliente, cambios regulatorios, calendarios de implantaciones o indicadores económicos del sector, que podrían aportar contexto adicional a la demanda observada y mejorar la precisión del modelo en escenarios no rutinarios.

- **Segmentación y modelado específico por cliente o clúster**: Desarrollar submodelos especializados para cada tipo de requerimiento o grupo de clientes, dada la heterogeneidad en los patrones de trabajo. Esto permitiría aumentar la precisión en segmentos donde el modelo actual presenta mayor margen de error.
- **Mantenimiento y recalibración periódica del modelo**: institucionalizar un proceso de actualización periódica del modelo, que contemple su reentrenamiento con datos recientes, evaluación de desempeño con métricas actualizadas y, de ser necesario, la sustitución del modelo por versiones más eficientes. Este enfoque garantizaría la vigencia del modelo a lo largo del tiempo y su adaptabilidad a contextos cambiantes.
- **Exploración de enfoques alternativos en modelado predictivo**: Evaluar el desempeño de modelos más complejos como redes neuronales profundas o modelos híbridos (por ejemplo, combinando series temporales y regresores) que puedan capturar relaciones no lineales más profundas, especialmente en casos extremos o de baja frecuencia.

Estas líneas de trabajo futuro representan una ruta viable para seguir fortaleciendo la capacidad analítica y predictiva de la organización. La implementación progresiva de estas mejoras no solo permitirá una asignación más eficiente de los recursos, sino que también consolidará una cultura de toma de decisiones basada en datos, con beneficios tangibles en la planeación operativa, la satisfacción del cliente y la sostenibilidad del servicio.

8 REFERENCIAS BIBLIOGRÁFICAS

[1] S. Chopra y P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, 5th ed. Boston: Pearson, 2013.

[2] M. Kuhn y K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2013.

[3] D. Mishra, P. Kumar y A. Kumar, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *Journal of Business Research*, vol. 131, pp. 308–320, 2021.

[4] X. Yang, S. Wang y Y. Li, "Demand Forecasting Based on Machine Learning Algorithms on Customer Information: An Applied Approach," *Journal of Business Analytics*, vol. 4, no. 3, pp. 210–224, 2021.

[5] A. Mukherjee, "Holt Winter's Method for Time Series Analysis," *Analytics Vidhya*, 26 de abril de 2023. [En línea]. Disponible en: https://www.analyticsvidhya.com/blog/2021/08/holt-winters-method-for-time-series-analysis/. [Accedido: 17-may-2025].

[6] 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.

[7] D. Ulrich, *Human Resource Champions: The Next Agenda for Adding Value and Delivering Results*. Boston: Harvard Business School Press, 1997.