Modelo de aprendizaje automático aplicado a la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO

Albeiro Buendía Diago, Karol Stefani Mejia, Oscar Morán Villarreal

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Gloria Inés Alvarez
Director del Trabajo de Grado

Gustavo Adolfo Arteaga
Codirector del Trabajo de Grado

Gerardo Mauricio Sarria
Jurado del Trabajo de Grado

Hernán Dario Vargas
Jurado del Trabajo de Grado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 27 de febrero del 2024

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 27 de febrero del 2024

Autor: Albeiro Buendía Diago, Karol Stefani Mejia, Oscar Morán Villarreal

Título del Trabajo de Grado: "Modelo de aprendizaje automático aplicado a la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO"

Director: Gloria Inés Alvarez Vargas

Codirector: Gustavo Adolfo Arteaga Botero

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado
Gloria Inés Alvarez Vargas

Firma del Codirector del Trabajo de Grado
Gustavo Adolfo Arteaga Botero

Santiago de Cali, 27 de febrero del 2024

Ingeniero
Juan Carlos Martinez Arias
Directora Posgrados de ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado “Modelo de aprendizaje automático aplicado a la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO”, el cual fue realizado por el (los) estudiante (s) Albeiro Buendia Diago, Karol Stefani Mejia, Oscar Moran Villareal con código(s) pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Gloría Inés Alvarez.

El director del Proyecto Aplicado autoriza para evaluar este proyecto, siempre que revise cuidadosamente el documento y avala listo para presentarse y sustentarse oficialmente.

Atentamente,

Albeiro Buendía Diago
C.C. 1113619811 de Palmira

Karol Stefani Mejia Rios
C.C. 1116283254 de Tuluá

Oscar Morán
C.C.1144044921 de Cali

Gloria Inés Álvarez Vargas
C.C. 30306105 de Manizales

Gustavo Adolfo Arteaga Botero
C.C. 75086575 de Manizales

FICHA RESUMEN
PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: Modelo de aprendizaje automático aplicado a la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO

1. **ÁREA DE TRABAJO:** Investigación
2. **TIPO DE PROYECTO (Aplicado, Innovación, Investigación):** Aplicado
3. **ESTUDIANTE(S):** Albeiro Buendía Diago, Karol Stefani Mejia, Oscar Morán Villarreal
4. **CORREO ELECTRÓNICO:** albeirobu@javerianacali.edu.co, karolstefani19@javerianacali.edu.co, ormoran@javerianacali.edu.co
5. **DIRECCIÓN Y TELEFONO:** Carrera 39 # 42-360 b/ Industrial Palmira Colombia 3157435255, Calle 10A #31-78 b/ Colseguros Cali Colombia 3214902673, Carrera 102 # 13 –45 b/ Ciudad Jardín Cali Colombia 3023492243
6. **DIRECTOR:** Gloria Inés Álvarez
7. **VINCULACIÓN DEL DIRECTOR:** Planta
8. **CORREO ELECTRÓNICO DEL DIRECTOR:** galvarez@javerianacali.edu.co
9. **CO-DIRECTOR (Si aplica):** Gustavo Adolfo Arteaga Botero
10. **GRUPO O EMPRESA QUE LO AVALA (Si aplica):** Metro Cali S.A Acuerdo de Reestructuración
11. **OTROS GRUPOS O EMPRESAS:** No aplica
12. **PALABRAS CLAVE (al menos 5):** Modelo de Aprendizaje Autónomo, Seguridad en transporte público, Máquinas de Vectores de Soporte, Random Forest, Perceptrón Multicapa.
13. **FECHA DE INICIO:** 02 de enero de 2023
14. **FECHA DE FINALIZACIÓN:** 12 de enero de 2024

RESUMEN: Este proyecto se enfocó en abordar las deficiencias de seguridad en el sistema de transporte masivo SITM MIO de Santiago de Cali, que experimenta incidentes crecientes de inseguridad. La gestión reactiva y la falta de control han afectado la confianza de los aproximadamente 280 mil usuarios diarios. Se identificó la necesidad de utilizar herramientas tecnológicas avanzadas para mejorar la asignación de recursos de seguridad de manera proactiva. Se desarrolló e implementó un sistema basado en técnicas estadísticas y computacionales, utilizando modelos de aprendizaje automático como Random Forest Regression, Support Vector Regression y Multilayer Perceptron Regression. La herramienta analítica predictiva resultante integra datos históricos y modelos de aprendizaje autónomo, destacando la eficacia del modelo de Random Forest Regression. Este avance marca un hito en la gestión de recursos de seguridad del transporte masivo, demostrando el impacto positivo de la ciencia de datos en la mejora de servicios públicos esenciales y la seguridad ciudadana.

Pontificia Universidad
JAVERIANA
Cali

Modelo de aprendizaje automático aplicado a la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO

Albeiro Buendía Diago
Karol Stefani Mejía
Oscar Morán Villarreal

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

Director(a)
Gloría Inés Álvarez

Codirector(a)
Gustavo Adolfo Arteaga Botero

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, ENERO 15 DE 2024

Tabla de contenido

Lista de Ilustraciones

Lista de Tablas

Introducción

En el contexto urbano de Santiago de Cali, el sistema de transporte masivo SITM MIO se ha consolidado como un eje vital para la movilidad de aproximadamente 280 mil usuarios diarios. Pero esta gran confluencia de personas, que incluye trabajadores, estudiantes y otros ciudadanos, ha expuesto falencias significativas en cuanto a seguridad. La gestión reactiva y la limitada capacidad de control han llevado a un aumento de incidentes de inseguridad, desde pequeños delitos hasta actos más graves, afectando la integridad y la confianza de los usuarios. En este escenario, se identificó una necesidad crítica de emplear y desarrollar herramientas tecnológicas avanzadas para mejorar la distribución de recursos de seguridad y abordar de manera proactiva las problemáticas de seguridad en el SITM MIO.

Este proyecto se centró en el desarrollo e implementación de un sistema basado en técnicas estadísticas y computacionales para optimizar la asignación de recursos de seguridad en las estaciones y terminales del SITM MIO. La iniciativa se inspiró en los resultados positivos observados en empresas del sector público y privado tras la incorporación de herramientas analíticas en sus procesos convencionales, lo que ha llevado a una toma de decisiones más efectiva y basada en datos. El desarrollo y la implementación de esta herramienta en un servicio público tan esencial como el transporte masivo eran una prioridad para mejorar la seguridad de los usuarios y una oportunidad para generar un impacto social significativo en la ciudad.

Para abordar esta problemática, se seleccionaron y aplicaron modelos de aprendizaje automático, incluyendo la Regresión de Bosques Aleatorios (Random Forest Regression), Regresión de Vectores de Soporte (Support Vector Regression) y Regresión de Perceptrón Multicapa (Multilayer Perceptron Regression), cada uno con características únicas para manejar datos complejos y de alta dimensionalidad. A través de una cuidadosa recolección y procesamiento de datos, y con el apoyo de un sistema de visualización en un dashboard interactivo, se logró desarrollar una aplicación web interesante. Esta aplicación no solo cumple con los requisitos técnicos establecidos, sino que ofrece una solución integral para la gestión eficiente de la seguridad en el SITM MIO.

El proyecto culminó con una herramienta analítica predictiva que integra los datos históricos con modelos de aprendizaje autónomo, proveyendo así una base sólida para decisiones estratégicas en la asignación de recursos de seguridad. Los resultados obtenidos destacan la eficacia del modelo de Random Forest Regression, que demostró un rendimiento sobresaliente en términos de precisión y confiabilidad en las predicciones. Esta implementación marca un avance significativo en la forma en que el transporte masivo de Santiago de Cali gestiona y optimiza sus recursos de seguridad, estableciendo un precedente para la aplicación de la ciencia de datos en la mejora de servicios públicos esenciales y la seguridad ciudadana.

1. Definición del problema

1.1. Planteamiento del problema

El Sistema Integrado de Transporte Masivo (SITM) de Santiago de Cali, conocido como MIO, juega un papel crucial en la movilidad urbana, facilitando la movilidad diaria de más de 280 mil usuarios. Con una infraestructura que abarca 7 terminales y 54 estaciones, el MIO enfrenta el desafío significativo de garantizar la seguridad y eficiencia en el manejo de un volumen tan alto de pasajeros. Sin embargo, los recursos de seguridad, compuestos por personal de policía y tecnologías de vigilancia, se ven superados por la demanda, resultando en incidentes de inseguridad que van desde agresiones verbales y vandalismo hasta robos y agresiones físicas severas [1].

La asignación actual de recursos de seguridad y otros, es realizada de manera manual, se centra en la prevención de evasores y en la respuesta a incidentes específicos de seguridad que ocurren día a día. Este enfoque reactivo limita la capacidad del sistema para anticiparse a los incidentes y asignar recursos de manera proactiva y estratégica [35][36].

Desafíos de Ciencia de Datos

La propuesta de desarrollar una herramienta analítica que permita una distribución más eficiente de los recursos de seguridad se enfrenta a varios desafíos significativos en el ámbito de la ciencia de datos:

- Disponibilidad de Datos: La recolección sistemática y consistente de datos operativos y de seguridad es un reto inicial, dada la variabilidad en la documentación y registro de incidentes y actividades del sistema.
- Cantidad de Datos: El volumen de datos generados diariamente por el sistema es amplio, lo que requiere capacidades robustas de procesamiento y análisis para extraer información útil.
- Balanceo de Datos: Debido al gran volumen de información puede que exista un desequilibrio significativo en los datos disponibles, especialmente en lo que respecta a incidentes de seguridad. Los eventos graves, aunque menos frecuentes, son de particular interés y requieren técnicas especializadas de manejo de datos para ser adecuadamente representados y analizados.
- Heterogeneidad de los Datos: Los datos recopilados provienen de múltiples fuentes y formatos, incluyendo reportes de incidentes, datos operacionales del sistema de transporte, y registros de cámaras de seguridad, lo que complica su integración y análisis.
- Ruido en los Datos: La calidad de los datos puede verse afectada por errores de registro, duplicaciones, y omisiones, lo que desafía la precisión y fiabilidad de los análisis realizados.

Ante estos desafíos, el objetivo de esta investigación es desarrollar una herramienta basada en técnicas avanzadas de aprendizaje automático que, al analizar información histórica detallada y actualizada de bitácoras de seguridad y operaciones del MIO, permita establecer criterios predictivos de riesgo y necesidad de recursos de seguridad. Esta herramienta buscará optimizar la asignación de recursos, mejorando significativamente la seguridad y eficiencia del sistema de transporte masivo, transformando la gestión reactiva en una estrategia proactiva y basada en datos.

1.2. Formulación del problema

Los Sistemas Integrados de Transporte Masivo (SITM) en operación, tales como Transmilenio, Megabús, Metroplús, Metrolínea, Transmetro, Transcaribe y el MIO, permitieron una transformación de la estructura de prestación de los servicios de transporte público urbano en Colombia y han generado desde su implementación innumerables beneficios económicos, sociales y ambientales, dejando atrás el modelo de transporte determinado por la “guerra del centavo”, flotas obsoletas, problemas de congestión y de seguridad en la operación de los buses, baja calidad del servicio, entre otros aspectos.

Lo anterior generó cambios en la movilidad de los usuarios, pero también trajo consigo los problemas existentes de inseguridad de la ciudad al sistema de transporte masivo. Actualmente, las cifras por robos al interior del sistema masivo representan el 19,25 % de los vandalismos en las estaciones y terminales del SITM MIO (Proposición 174 concejo de Cali), pero el dato más relevante es de los daños y robos a la infraestructura con una participación del 60,94 %, debido a la falta de control y presencia de la institución pública y privada. A partir del estallido social en el marco del paro nacional del 28 de abril de 2021, la ciudad de Santiago de Cali vivió una crisis social, que afectó en un 90% la infraestructura del SITM MIO, generando un repunte en los casos de violencia al interior de la flota, estaciones y terminales. La gestión de Metro Cali S.A, para velar por la seguridad de los usuarios, recae en la disposición de la administración local de cada entidad, es por esto, que el recurso es limitado y se debió con ello plantear estrategias para mitigar problemas como la evasión, cosquilleo, agresiones, etc. [1].

En consecuencia, el aumento continuo de la infraestructura de transporte masivo ha originado otra problemática relacionada con garantizar la seguridad de los ciudadanos. Actualmente, los sistemas integrados de transporte masivo trabajan para disminuir las cifras de casos asociados a novedades de seguridad, pero no han logrado el impacto suficiente por la forma asignada el recurso. La forma de asignación se realiza de forma reactiva, es decir, ante la novedad y hechos recientes de hurtos, violencia, cosquilleos, se asigna la cantidad de acuerdo con la disponibilidad del momento.

Por lo anterior, la ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer datos en sus diferentes formas, ya sea estructurados o no estructurados, siendo una continuación de áreas de análisis de datos como pueden ser la estadística, la minería de datos, el aprendizaje automático (Machine Learning) y la analítica predictiva. Dentro de la ciencia de datos, existen diferentes campos que se apoyan en los datos y en la computación para sustentar la toma de decisiones. Por el alcance y la posibilidad de realizar modelos de aprendizaje autónomo, a través de esta ciencia, se planteó desarrollar el proyecto para el SITM MIO, teniendo en cuenta, la complejidad de la información a extraer y la histórica inseguridad, que establecieron los criterios necesarios para emplear dichas herramientas.

Al abordar el problema de seguridad en el SITM MIO desde la perspectiva de la ciencia de datos, se enfrentan varios desafíos inherentes a la naturaleza de los datos y al proceso de análisis. Algunos de estos desafíos incluyen:

- **Acceso a los datos:** Obtener acceso a datos relevantes y completos puede ser un desafío, especialmente si involucra información sensible sobre incidentes de seguridad y evasión. Además, garantizar la calidad y la integridad de los datos es crucial para realizar un análisis preciso.

- **Cantidad y disponibilidad de datos:** Asegurar una cantidad adecuada de datos históricos es fundamental para entrenar modelos de aprendizaje automático eficaces. Sin embargo, la disponibilidad de datos históricos específicos de incidentes de seguridad en el SITM MIO puede ser limitada o fragmentada.

- **Heterogeneidad de los datos:** Los datos relacionados con la seguridad en el SITM MIO pueden provenir de diversas fuentes y estar en diferentes formatos, como registros de incidentes, datos de cámaras de seguridad, informes policiales, entre otros. Integrar y normalizar estos datos heterogéneos para su análisis puede ser un desafío.

- **Ruido en los datos:** Los datos recopilados pueden contener ruido, es decir, información irrelevante o inexacta que puede afectar la precisión de los modelos analíticos. Es importante preprocesar los datos para eliminar el ruido y mejorar la calidad del análisis.

- **Escalabilidad y rendimiento:** A medida que la cantidad de datos aumenta, es fundamental que los algoritmos y las herramientas utilizadas para analizar estos datos sean escalables y eficientes en términos de rendimiento computacional.

- **Interpretación de resultados:** La interpretación de los resultados de los análisis de datos en el contexto de la seguridad del transporte masivo puede ser compleja y requiere un entendimiento profundo de los patrones y

tendencias identificados. Además, la comunicación efectiva de estos resultados a las partes interesadas es crucial para la toma de decisiones informadas.

Abordar estos desafíos requiere un enfoque multidisciplinario que combine conocimientos en ciencia de datos, seguridad pública, transporte urbano y tecnologías de la información. Es importante desarrollar soluciones innovadoras que aprovechen al máximo los datos disponibles y que contribuyan a mejorar la seguridad y la eficiencia del sistema de transporte masivo.

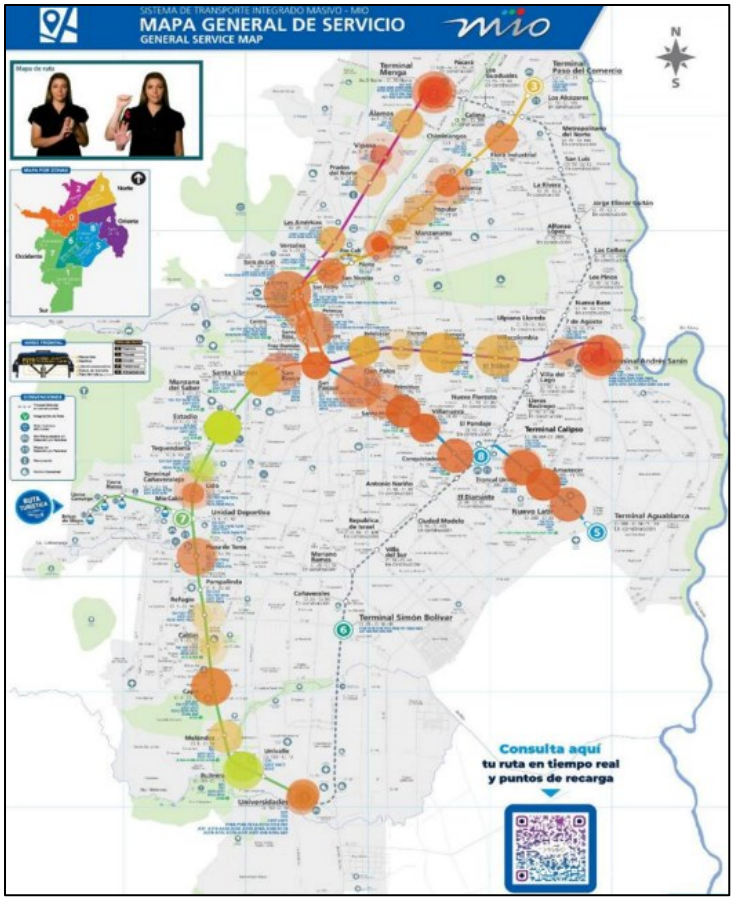

ILUSTRACIÓN 1. MAPA DE EVASIÓN SITM MIO

La Ilustración 1. Mapa de Evasión SITM MIO muestra la evasión con un indicador de color, considerando la relación entre el conteo promedio de evasores y los usos promedios por estación, cuando son aquellas estaciones sin seguridad física y tecnológica, las que más índice de evasión contemplan (rojo: estaciones con más del 10% de evasión). Por lo anterior, el algoritmo de aprendizaje autónomo tendrá en cuenta la información de evasión, para distribuir los recursos institucionales y mejorar la cobertura para mitigar problemas asociados a la evasión y seguridad del sistema.

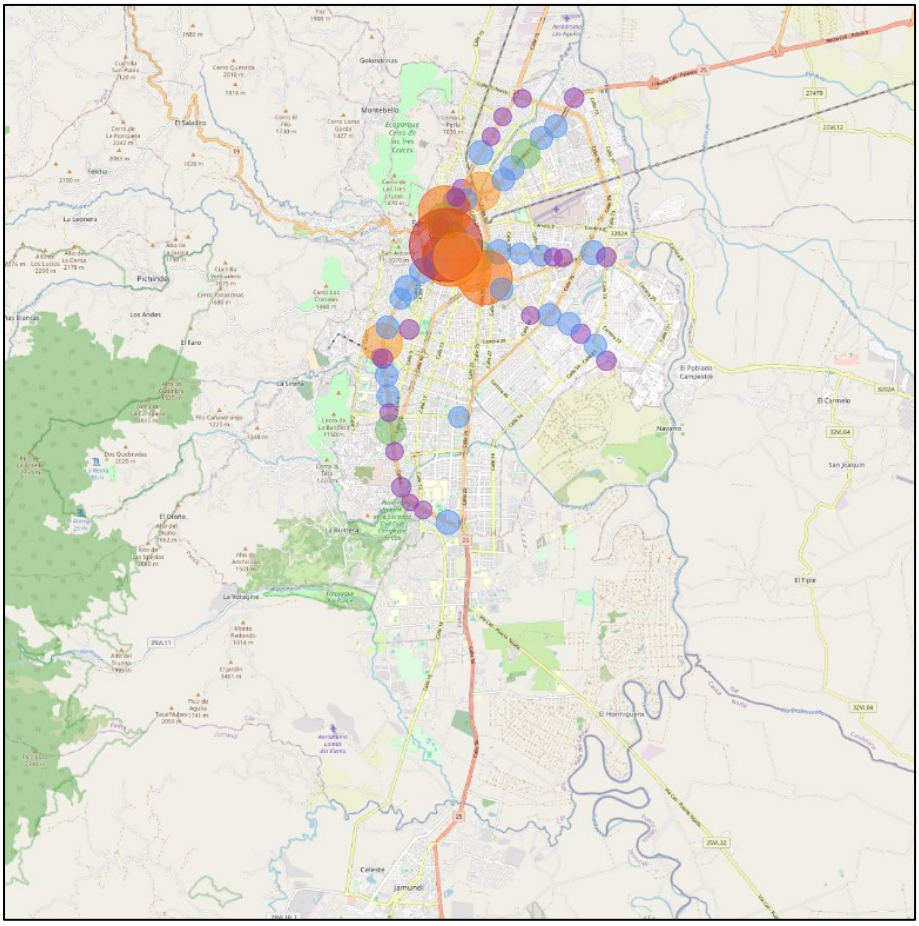

ILUSTRACIÓN 2. MAPA DE VANDALISMOS SITM MIO

La Ilustración 2. Mapa de Vandalismos SITM MIO, diseñado para visualizar los incidentes de vandalismo en el Sistema de Transporte Masivo Integrado de Occidente (SITM MIO), utiliza un sistema de indicadores de color para representar la cantidad de robos y hurtos en la infraestructura del sistema. Cada punto en el mapa corresponde a una ubicación específica dentro del sistema, como estaciones o terminales, donde se han registrado estos incidentes. Los colores varían en intensidad o tonalidad para indicar la frecuencia o severidad de los robos y hurtos, facilitando la identificación rápida de las áreas con mayor incidencia de vandalismo. Por lo anterior, el algoritmo de aprendizaje autónomo tendrá en cuenta la información de vandalismo en estaciones y terminales, para distribuir los recursos institucionales y mejorar la cobertura para mitigar problemas asociados a la evasión y seguridad del sistema.

2. Justificación

El proyecto propuesto será útil al gestor del SITM MIO, ya que se entregará una herramienta que ayudará a distribuir mejor los recursos institucionales disponibles, para lograr en las estaciones y terminales del SITM MIO, mejor cobertura y sensación de seguridad, mitigando la inseguridad presentada en el SITM MIO (Sistema Integrado de Transporte Masivo MIO) de Santiago de Cali.

Por otro lado, el proyecto a realizar requiere usar técnicas propias de la ciencia de datos, por lo tanto es viable para realizarlo con el conocimiento aprendido en la maestría y tiene en cuenta tres aspectos: la base de datos con información histórica de vandalismos en estaciones y terminales del SITM MIO desde el año 2010 al 2022, el apoyo de la entidad gestora del SITM MIO para desarrollar el proyecto y la información en caso de necesitarse de cada una de sus dependencias internas que conforman Metro Cali S.A, siendo lo anterior necesario para lograr el desarrollo del modelo de aprendizaje autónomo propuesto.

El modelo ayudará a distribuir recursos según criterios esperados, si el ente gestor sigue las recomendaciones de asignación del algoritmo, mitigar problemas de inseguridad presentados en el principal articulador de la movilidad de Santiago de Cali el SITM MIO, generando un impacto social de gran importancia en la ciudad; pero la medición del impacto lo hará la entidad gestora y si requiere se hará junto con desarrolladores consolidados de dos meses.

3. Objetivos del proyecto

3.1. Objetivo general

Desarrollar un modelo de aprendizaje autónomo, predictivo, que determine la asignación adecuada de los recursos institucionales disponibles en las estaciones y terminales del sistema de transporte masivo del distrito de Santiago de Cali SITM MIO.

3.2. Objetivos específicos

- Analizar los datos suministrados por Metro Cali S.A de novedades de seguridad de las terminales y estaciones.

- Investigar el modelo que más se ajuste para el desarrollo de la herramienta de predicción para los datos de seguridad en las estaciones y terminales suministrados por Metro Cali S.A.

- Desarrollar una herramienta de predicción que se ajuste a la base de datos suministrada de los incidentes históricos de seguridad del SITM MIO.

- Realizar dashboard con mapa dinámico del sistema de transporte masivo de Santiago de Cali para visualizar la asignación de los recursos institucionales sugerida por el modelo.

4. Marco teórico y antecedentes

4.1. Marco teórico

En el ámbito de la ciencia de datos, el presente proyecto de tesis se enfoca en desarrollar un modelo de aprendizaje autónomo y predictivo para optimizar la asignación de recursos en el sistema de transporte masivo del distrito de Santiago de Cali (SITM MIO). La relevancia de integrar métodos de aprendizaje supervisado, minería de datos y técnicas de balanceo de datos es fundamental para abordar eficazmente este desafío.

4.1.1. Minería de datos

La minería de datos, una disciplina en la intersección de la estadística, la inteligencia artificial y la gestión de bases de datos, tiene como objetivo descubrir patrones y relaciones ocultas en grandes conjuntos de datos. Los avances tecnológicos y la creciente disponibilidad de datos han impulsado su evolución en las últimas décadas.

4.1.1.1. Evolución Histórica

La minería de datos ha evolucionado desde métodos estadísticos simples hasta algoritmos complejos de aprendizaje automático. En los primeros días, se enfocaba principalmente en la exploración de datos y la generación de hipótesis. Hoy en día, se enfoca en la predicción y prescripción, aprovechando algoritmos avanzados y potencia computacional.

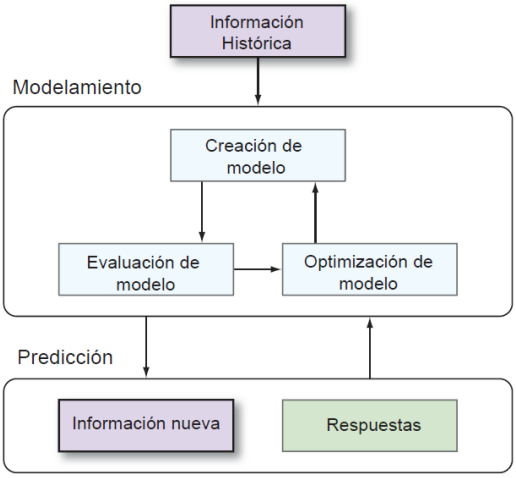

ILUSTRACIÓN 3. WORKFLOW DE LA METODOLOGÍA USUAL USADA EN MACHINE LEARNING

4.1.1.2. Preparación de Datos

Esta etapa es crucial para el éxito de cualquier modelo de minería de datos. Incluye la limpieza de datos, la gestión de valores faltantes, la transformación y normalización de datos, y la codificación de variables categóricas. Un enfoque detallado en esta etapa asegura que los modelos posteriores sean más precisos y eficientes.

4.1.1.3. Creación de Modelo

Implica seleccionar y entrenar algoritmos de aprendizaje automático adecuados. Los algoritmos pueden variar desde regresiones lineales hasta redes neuronales complejas, dependiendo de la naturaleza y complejidad de los datos.

4.1.1.4. Evaluación

Esta fase implica validar la precisión del modelo utilizando métricas como la precisión, la sensibilidad, y el área bajo la curva ROC. La validación cruzada y otras técnicas de evaluación son esenciales para garantizar que el modelo se desempeñe bien en datos no vistos.

- **Precisión:** La precisión es la proporción de verdaderos positivos (TP) sobre la suma de verdaderos positivos y falsos positivos (FP). En otras palabras, mide la proporción de predicciones positivas que fueron correctas.

$$Precisión = \frac{TP}{TP + FP}$$

- **Sensibilidad:** La sensibilidad es la proporción de verdaderos positivos (TP) sobre la suma de verdaderos positivos y falsos negativos (FN). En términos más simples, mide la proporción de positivos reales que fueron identificados correctamente.

$$Sensibilidad = \frac{TP}{TP + FN}$$

- **Área bajo la curva ROC:** El área bajo la curva ROC es una medida de la capacidad de un modelo para distinguir entre clases. Cuanto mayor sea el área bajo la curva (AUC), mejor será el modelo para distinguir entre clases positivas y negativas. La curva ROC es una representación gráfica de la sensibilidad (eje Y) frente a la tasa de falsos positivos (eje X), y el área bajo esta curva proporciona una medida numérica de la calidad del modelo.

$$AUC = \int_0^1 sensibilidad(x) d tasa\ de\ falsos\ positivos\ (x)$$

4.1.1.5. Mejora de Modelo

Incluye técnicas como ajuste de hiperparámetros, selección de características y manejo de desequilibrio en los datos. Esta etapa es iterativa y busca optimizar el rendimiento del modelo.

4.1.1.6. Predicción

La finalidad de la minería de datos. Se utiliza el modelo final para hacer predicciones o tomar decisiones basadas en nuevos datos.

4.1.1.7. Desafíos y Soluciones Contemporáneas

Uno de los mayores desafíos actuales es el manejo de grandes volúmenes de datos. Técnicas como el aprendizaje profundo y el procesamiento distribuido han sido fundamentales para abordar este desafío. Además, la necesidad de modelos interpretables ha llevado al desarrollo de técnicas de explicabilidad de IA, que permiten comprender mejor cómo los modelos toman decisiones [2] [3].

4.1.1.8. Creación de modelo

Para esta etapa se deben hacer las preguntas que pueden ser respondidas por los datos. El trabajo del algoritmo de ML será predecir con éxito el objetivo a través del conjunto de inputs.

4.1.1.9. Evaluación

Aunque no siempre se realiza, suele ser una fase común de la metodología. Una forma de realizar la evaluación es probar la función de hipótesis con datos reales ya mesurado en la data histórica.

4.1.1.10. Mejora de modelo

3 formas comunes son perfeccionar los parámetros de la hipótesis, filtrar el conjunto de los inputs y reprocesamiento y limpieza de los datos.

4.1.1.11. Predicción

Es la última etapa del modelo, y el objetivo final, donde se realiza la predicción o tarea asignada al agente de Machine Learning.

4.1.2. Ventana de tiempo en el análisis temporal

Una ventana de tiempo es una técnica utilizada en el análisis de series temporales que consiste en seleccionar un subconjunto de datos basándose en un período específico transformando la serie temporal en una matriz en la que, cada valor del conjunto de datos está asociado a un valor a predecir. Esta técnica es particularmente útil para transformar series temporales en un formato más adecuado para el modelado predictivo, especialmente para modelos que necesitan considerar dependencias temporales [28].

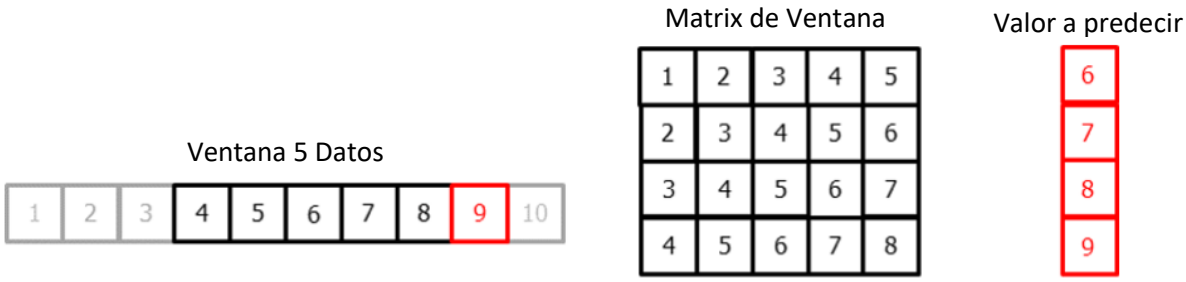

ILUSTRACIÓN 4.SKFORECAST: FORECASTING SERIES TEMPORALES CON PYTHON Y SCIKITLEARN

La ventana de tiempo se incorpora como un componente esencial para reorganizar la estructura de datos y poder realizar el entrenamiento y la evaluación de la efectividad de los modelos de regresión desarrollados en este estudio [4].

La codificación de la serie de tiempo por ventana de tiempo es crucial, ya no permite convertir los datos cronológicos de eventos, en códigos que representan la continuidad de la serie. Esto implica que, en casos donde falten datos de eventos en determinados días y horas, es necesario realizar una reconstrucción de la serie de tiempo para obtener la serie temporal con todos los datos en el tiempo.

4.1.2.1. Propósito del Uso de la Ventana de Tiempo

- **Objetivo Principal:** Utilizar la ventana de tiempo para organizar y transformar la serie temporal de datos, permitiendo una mejor adaptación del modelo de aprendizaje automático a códigos que representan la temporalidad.

- La ventana de tiempo es un instrumento clave que facilita la reorganización de los datos para realizar el entrenamiento, pruebas y usos en un modelo predictivo, dado la naturaleza dinámica de las operaciones y la gestión de recursos [4].

4.1.2.2. Consideraciones Adicionales en el Análisis Temporal

- **Segmentación Temporal:** La ventana de tiempo permitirá la segmentación de los datos en intervalos específicos, facilitando un análisis más detallado de las variaciones temporales y la identificación de patrones relevantes [4].

- **Transformación de la Serie Temporal:** Los datos serán organizados y transformados dentro de la ventana de tiempo antes de ser introducidos en los modelos de aprendizaje automático, asegurando que las relaciones temporales sean adecuadamente consideradas [4].

4.1.2.3. Aplicación Práctica de la Ventana de Tiempo en el Proyecto

- **Evaluación Continua:** La ventana de tiempo se ajustará y adaptará durante diferentes fases del desarrollo del modelo, asegurando que el modelo se mantenga relevante y preciso a medida que se disponga de nuevos datos o se realicen ajustes en la ventana temporal. La retroalimentación continua será esencial para optimizar la capacidad del modelo para abordar las variaciones temporales en el SITM MIO [4].

4.1.3. Machine Learning

Machine learning es un componente importante del creciente campo de la ciencia de datos. Mediante el uso de métodos estadísticos, se entrenan algoritmos para hacer clasificaciones o predicciones, y para descubrir insights clave dentro de los proyectos de minería de datos. Estos insights posteriormente impulsan la toma de decisiones dentro de aplicaciones y empresas, lo que es ideal para influir en las métricas [5] [6].

El aprendizaje automático (machine learning) abarca una variedad de modelos y técnicas algorítmicas diseñadas para procesar datos de manera efectiva y lograr resultados específicos. Dependiendo de la naturaleza de los datos y los objetivos deseados, se recurre a uno de los cuatro modelos de aprendizaje: supervisado, no supervisado, semisupervisado o de refuerzo. Cada uno de estos modelos ofrece enfoques distintos para abordar diferentes problemas en el ámbito del aprendizaje automático [5].

En el contexto de cada modelo, se aplican diversas técnicas algorítmicas que se ajustan a las características particulares de los conjuntos de datos y los resultados buscados. Estas técnicas pueden incluir clasificación, identificación de patrones, proyección de resultados y toma de decisiones informadas [6].

Es esencial destacar que los algoritmos de aprendizaje automático están diseñados para cumplir funciones específicas, ya sea clasificar elementos, descubrir patrones, realizar proyecciones o tomar decisiones fundamentadas. La flexibilidad de estos algoritmos permite su aplicación individual o su combinación estratégica para lograr

la máxima precisión, especialmente cuando se enfrentan a datos complejos y situaciones más impredecibles. En resumen, el aprendizaje automático ofrece un conjunto diverso de herramientas que se adaptan a diversas necesidades y contextos, permitiendo un enfoque personalizado para la resolución de problemas complejos en el análisis de datos [5] [6]. Es esencial destacar que los algoritmos de aprendizaje automático están diseñados para cumplir funciones específicas, ya sea clasificar elementos, descubrir patrones, realizar proyecciones o tomar decisiones fundamentadas. La flexibilidad de estos algoritmos permite su aplicación individual o su combinación estratégica para lograr la máxima precisión, especialmente cuando se enfrentan a datos complejos y situaciones más impredecibles. En resumen, el aprendizaje automático ofrece un conjunto diverso de herramientas que se adaptan a diversas necesidades y contextos, permitiendo un enfoque personalizado para la resolución de problemas complejos en el análisis de datos [5] [6].

4.1.3.1. Métodos de Machine Learning

Los modelos de machine learning se dividen en cuatro categorías principales.

- **Machine learning supervisado:** El aprendizaje supervisado, también conocido como machine learning supervisado, se define por su uso de los conjuntos de datos etiquetados para entrenar los algoritmos para clasificar datos o predecir resultados con precisión [7]. A medida que se introducen datos de entrada en el modelo, este adapta sus pesos hasta que se haya ajustado correctamente. Esto ocurre como parte del proceso de validación cruzada para asegurarse de que el modelo evite el sobreajuste o el subajuste. El aprendizaje supervisado permite a las organizaciones resolver una amplia variedad de problemas del mundo real a escala como, por ejemplo, la clasificación de spam en una carpeta distinta de la bandeja de entrada [8]. Algunos métodos utilizados en el aprendizaje supervisado son las redes neuronales, Naïve Bayes, la regresión lineal, la regresión logística, el bosque aleatorio y la máquina de vectores de soporte (SVM).

- **Machine learning no supervisado:** El aprendizaje no supervisado, también conocido como machine learning no supervisado, utiliza algoritmos de machine learning para analizar y agrupar en clústeres conjuntos de datos sin etiquetar [9]. Estos algoritmos descubren agrupaciones de datos o patrones ocultos sin necesidad de ninguna intervención humana. La capacidad de este método para descubrir similitudes y diferencias en la información lo convierten en ideal para el análisis de datos exploratorios, las estrategias de venta cruzada, la segmentación de clientes y el reconocimiento de imágenes y patrones. También se utiliza para reducir el número de características de un modelo mediante el proceso de reducción de dimensionalidad. El análisis de componentes principales (PCA) y la descomposición en valores singulares (SVD) son dos de los enfoques más habituales para realizar este proceso. Otros algoritmos utilizados en el aprendizaje no supervisado son las

redes neuronales y la agrupación en clúster de medias K.

- **Aprendizaje semisupervisado:** El aprendizaje semisupervisado ofrece un punto intermedio entre el aprendizaje supervisado y no supervisado. Durante el entrenamiento, utiliza un conjunto de datos etiquetados más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos sin etiquetar de mayor tamaño [10]. El aprendizaje semisupervisado puede resolver el problema de no tener suficientes datos etiquetados para un algoritmo de aprendizaje supervisado. También es útil si el coste de etiquetar datos suficientes es demasiado elevado.

- **Machine learning de refuerzo:** Machine learning de refuerzo es un modelo de machine learning que es similar al aprendizaje supervisado, pero el algoritmo no se entrena utilizando datos de muestra. Este modelo aprende a través de prueba y error. Se reforzará una secuencia de resultados satisfactorios para desarrollar la mejor recomendación o política para un problema determinado [10].

4.1.3.2. Funcionamiento del aprendizaje supervisado

El aprendizaje supervisado es el uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados de forma precisa [7]. A medida que los datos se introducen en el modelo, este ajusta sus ponderaciones hasta que dicho modelo se haya ajustado adecuadamente, como parte del proceso de validación cruzada.

El aprendizaje supervisado, también conocido como machine learning supervisado, es una subcategoría del machine learning y la inteligencia artificial [8].

El aprendizaje supervisado permite a las organizaciones resolver una amplia variedad de problemas del mundo real a escala como, por ejemplo, la clasificación de spam en una carpeta distinta de la bandeja de entrada.

El aprendizaje supervisado utiliza un conjunto de datos de entrenamiento para enseñar a los modelos a generar la salida deseada [9].

Este conjunto de datos incluye datos de entrada y resultados correctos, que permiten que el modelo aprenda con el tiempo. El algoritmo mide su precisión a través de la función de pérdida, ajustándose hasta que el error se haya minimizado lo suficiente.

El aprendizaje supervisado puede clasificarse en dos tipos de problemas durante la minería de datos:

- La clasificación

- La regresión

4.1.3.2.1. CLASIFICACIÓN

La clasificación utiliza un algoritmo para asignar con precisión datos de prueba en categorías específicas.

Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo esas entidades deben etiquetarse o definirse.

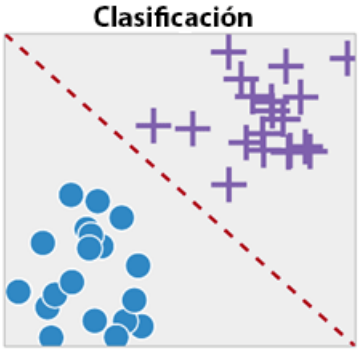

ILUSTRACIÓN 5. EJEMPLO GRÁFICO DE CLASIFICACIÓN

4.1.3.2.2. REGRESIÓN

La regresión se utiliza para comprender la relación entre variables dependientes e independientes.

Se utiliza comúnmente para hacer proyecciones, como los ingresos por ventas de un negocio determinado [7].

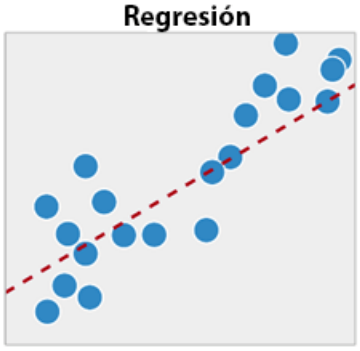

ILUSTRACIÓN 6. EJEMPLO GRÁFICO DE REGRESIÓN

4.1.3.3. Algoritmos de aprendizaje supervisado

4.1.3.3.1. MÁQUINAS DE VECTORES DE SOPORTE

Las Máquinas de Vectores de Soporte (SVM) se presentan como una herramienta analítica robusta y versátil en este contexto. Originadas en los trabajos pioneros de Vapnik [11], las SVM son ampliamente reconocidas por su eficacia en tareas de clasificación y regresión, fundamentales para la predicción de incidentes de seguridad.

- **Eficiencia en Grandes Dimensiones de Datos:** Las SVM son particularmente adecuadas para manejar grandes volúmenes de datos, una característica esencial para analizar los registros de seguridad del SITM MIO. La capacidad de las SVM para utilizar funciones kernel para transformar los datos a un espacio de mayor dimensión facilita la separación lineal, incluso en casos donde las relaciones entre los datos no son inmediatamente evidentes [12]. Esta propiedad es crucial para identificar patrones complejos y sutiles en los datos de seguridad.

- **Flexibilidad y Adaptabilidad:** La adaptabilidad de las SVM a diferentes tipos de datos y relaciones es una ventaja significativa. A través de la elección adecuada de la función kernel, las SVM pueden modelar una variedad de relaciones no lineales entre las variables, lo que es esencial para comprender y predecir los incidentes de seguridad en un entorno tan dinámico como el del transporte público [13].

- **Precisión y Control del Margen de Error:** Una característica distintiva de las SVM es su enfoque en maximizar el margen, lo que las hace particularmente robustas frente a datos de prueba futuros. Esta propiedad es vital para asegurar que la asignación de recursos de seguridad basada en el modelo sea confiable y efectiva [14]. En el contexto del SITM MIO, esto significa que el modelo puede proporcionar recomendaciones consistentes y fiables para la asignación de recursos, incluso en situaciones de incertidumbre y cambio.

- **Aplicaciones Prácticas en Seguridad de Transporte:** La implementación de SVM en sistemas de transporte masivo como el SITM MIO puede revolucionar la forma en que se asignan los recursos de seguridad. Al predecir con mayor precisión los incidentes de seguridad, las autoridades pueden distribuir los recursos de manera más efectiva, asegurando una respuesta rápida y adecuada a las situaciones emergentes. Esto no solo mejora la seguridad general del sistema, sino que también optimiza el uso de recursos limitados.

4.1.3.3.2. PERCEPTRÓN MULTICAPA

El desarrollo de un modelo de aprendizaje automático para optimizar la asignación de recursos en el sistema de transporte masivo de Santiago de Cali (SITM MIO) puede beneficiarse significativamente del uso del Perceptrón Multicapa (MLP). El

MLP, una forma de redes neuronales artificiales ofrece una estructura flexible y potente para modelar relaciones complejas en datos de gran volumen y variedad, como los asociados con la seguridad en el transporte masivo.

- **Capacidad de Modelado de Datos Complejos:** El MLP es conocido por su habilidad para modelar relaciones no lineales complejas, lo que lo hace ideal para analizar y predecir patrones en los datos de seguridad del SITM MIO [15]. A través de su arquitectura de múltiples capas y neuronas interconectadas, el MLP puede aprender patrones intrincados en los datos, lo que es crucial para identificar factores de riesgo y predecir incidentes de seguridad.

- **Flexibilidad y Generalización:** Una de las fortalezas del MLP es su capacidad para generalizar a partir de los datos de entrenamiento y realizar predicciones precisas sobre datos no vistos anteriormente [16]. Esta característica es esencial para adaptarse a las cambiantes condiciones y patrones de seguridad en un sistema de transporte dinámico como el SITM MIO.

- **Robustez en el Manejo de Datos de Alta Dimensión:** Las redes neuronales, como el MLP, son particularmente eficaces en el manejo de datos de alta dimensión, lo que permite integrar y analizar múltiples factores que influyen en la seguridad del transporte masivo [17]. Esto incluye variables como la hora del día, ubicación, tipo de incidente, y otros datos relevantes.

- **Aplicaciones en Sistemas de Transporte Masivo:** La implementación de MLP en el SITM MIO puede ofrecer una herramienta avanzada para la asignación de recursos de seguridad. Al predecir con precisión los incidentes de seguridad, el modelo puede guiar la distribución de recursos de manera más efectiva, asegurando una respuesta oportuna y adecuada a las situaciones de riesgo [18]. Esto no solo mejora la seguridad general del sistema, sino que también contribuye a un uso más eficiente de los recursos disponibles.

4.1.3.3.3. RANDOM FOREST

El modelo de Random Forest (RF) es una técnica de aprendizaje automático que se ha establecido como una de las herramientas más eficaces y versátiles para el análisis predictivo, especialmente en el contexto de grandes conjuntos de datos con una compleja estructura de atributos, como es el caso de los datos de seguridad en el sistema de transporte masivo de Santiago de Cali (SITM MIO) [19].

- **Fundamentos y Estructura del Random Forest:** El Random Forest es un algoritmo de aprendizaje de conjunto que opera mediante la construcción de múltiples árboles de decisión en el momento del entrenamiento y generando la salida (por ejemplo, la clasificación o la media de las predicciones de los árboles individuales) [19]. La fórmula general para un Random Forest que realiza una tarea de clasificación puede expresarse como:

$$Y = moda\{h(x, \Theta_1), h(x, \Theta_2), \dots, h(x, \Theta_k),\}$$

donde Y es la salida predicha, h representa un árbol individual dentro del bosque, x es el vector de entrada, y $\Theta_1, \Theta_2, \dots, \Theta_k$ son los parámetros aleatorios independientes para cada árbol.

- **Ventajas en el Contexto de Seguridad del Transporte Masivo:** Una de las principales ventajas de Random Forest es su capacidad para manejar grandes conjuntos de datos con una alta dimensión de atributos, lo que lo hace ideal para analizar los complejos patrones de datos de seguridad del SITM MIO. Además, su naturaleza de conjunto reduce el riesgo de sobreajuste, un problema común en los modelos de aprendizaje automático, especialmente en árboles de decisión individuales.

- **Manejo de Diversos Tipos de Datos:** Random Forest es eficaz tanto para variables categóricas como numéricas, lo que permite una amplia gama de aplicaciones, desde la predicción de tipos de incidentes hasta la identificación de patrones temporales y espaciales en los datos de seguridad [20].

- **Importancia en la Asignación de Recursos de Seguridad:** La aplicación de Random Forest en el SITM MIO puede proporcionar predicciones precisas y confiables sobre incidentes de seguridad, lo que a su vez puede guiar una asignación más eficiente y efectiva de recursos. Al anticipar áreas y tiempos de mayor riesgo, el modelo puede ayudar a optimizar la presencia de personal de seguridad y la implementación de medidas preventivas [21].

4.1.3.4. Coeficiente de determinación R^2

El coeficiente de determinación R^2 se establece como una métrica esencial para evaluar la eficacia de los modelos de regresión desarrollados. A continuación, se presenta el criterio adoptado para la aplicación y análisis de R^2 en este estudio específico [22].

4.1.3.4.1. PROPÓSITO DEL USO DE R^2

Objetivo Principal: Utilizar R^2 para determinar qué tan bien el modelo de aprendizaje automático puede explicar la variabilidad en la asignación de recursos y predecir con precisión las necesidades de seguridad y mantenimiento en el SITM MIO.

Importancia en el Contexto del Proyecto: Dada la complejidad y la importancia crítica de la asignación eficiente de recursos, un alto valor de R^2 indica que el modelo puede capturar con precisión las relaciones subyacentes entre las variables y las necesidades operativas.

4.1.3.4.2. INTERPRETACIÓN DE LOS VALORES DE R²

Rango de Valores: Se considera que R² varía entre 0 (sin explicación) y 1 (explicación completa).

Umbrales de Aceptabilidad: Para este proyecto, se establecerán umbrales específicos para R² que determinarán la aceptabilidad del modelo. Por ejemplo, un R² superior a 0.7 puede considerarse como indicativo de un buen ajuste del modelo a los datos.

4.1.3.4.3. CONSIDERACIONES ADICIONALES EN EL ANÁLISIS DE R²

Evitar el Sobreajuste: Se prestará especial atención para asegurar que un R² alto no sea resultado de sobreajuste, especialmente si el modelo muestra un rendimiento significativamente diferente en el conjunto de datos de prueba.

Complementar con Otras Métricas: R² se utilizará en conjunto con otras métricas de rendimiento, como el error cuadrático medio (MSE) y el análisis de residuos, para obtener una evaluación más completa del modelo.

4.1.3.4.4. APLICACIÓN PRÁCTICA DE R² EN EL PROYECTO

Evaluación Continua: El R² se calculará y monitorizará a lo largo de diferentes fases del desarrollo del modelo para asegurar que el modelo se mantenga relevante y preciso a medida que se disponga de nuevos datos o se realicen ajustes en el modelo.

$$R^2 = 1 - \frac{Suma\ de\ Cuadrados\ de\ los\ Residuos\ (SSR)}{Suma\ Total\ de\ Cuadrados\ (SST)}$$

Donde:

- **Suma de Cuadrados de los Residuos (SSR)** se calcula como: $\sum(y_i - \hat{y}_i)^2$ siendo y_i los valores reales y $\hat{y}_i$ los valores predichos por el modelo.
- **Suma Total de Cuadrados (SST)** se calcula como: $\sum(y_i - \bar{y})^2$, donde $\bar{y}$ es el promedio de los valores reales.

En términos más sencillos, el R² mide la proporción de la variabilidad total de la variable dependiente que es explicada por el modelo de regresión.

Para ilustrar mejor esto, se tiene que:

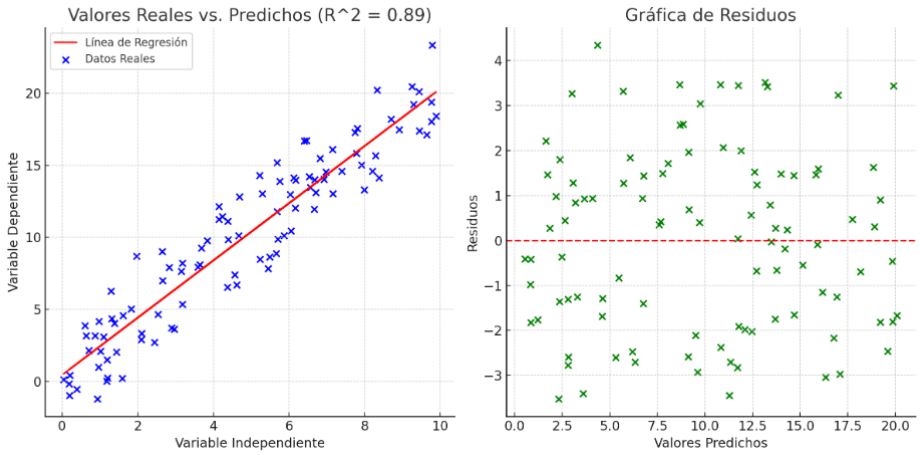

ILUSTRACIÓN 7.DISPERSIÓN DE LOS VALORES REALES VS PREDICHOS

- **La gráfica de dispersión de los valores reales vs. predichos** muestra la línea de regresión. Esta gráfica puede ayudar a visualizar qué tan cerca están los valores predichos de los valores reales. El R^2 calculado para este modelo es 0.91, indicando que el modelo explica aproximadamente el 91% de la variabilidad en los datos. Un R^2 más cercano a 1 indica un mejor ajuste del modelo.

- **Gráfica de los residuos**, que es útil para evaluar la distribución y la constancia de los errores del modelo a lo largo de diferentes valores predichos. Esta gráfica es útil para evaluar si los residuos parecen ser aleatorios y uniformemente distribuidos alrededor de la línea horizontal en $y = 0$, lo cual es deseable en un buen modelo de regresión.

4.1.3.5. Error absoluto medio (MAE)

El Error Absoluto Medio (MAE) se utiliza como una métrica esencial para evaluar la precisión de los modelos de regresión desarrollados en el contexto del presente estudio. A continuación, se presenta el criterio adoptado para la aplicación y análisis del MAE en este proyecto específico [23].

4.1.3.5.1. PROPÓSITO DEL USO DE MAE

Objetivo Principal: Utilizar MAE para evaluar qué tan bien el modelo de aprendizaje automático puede predecir con precisión las necesidades de seguridad y mantenimiento en el SITM MIO, centrándose en la magnitud promedio de los errores de predicción.

Importancia en el Contexto del Proyecto: Dada la complejidad y la importancia crítica de la asignación eficiente de recursos, un bajo valor de MAE indica que el modelo puede realizar predicciones precisas y realistas, contribuyendo a una gestión efectiva de recursos.

4.1.3.5.2. INTERPRETACIÓN DE LOS VALORES DE MAE

Magnitud de Errores: MAE mide la magnitud promedio de los errores absolutos entre las predicciones y los valores reales. Un MAE más bajo indica una mayor precisión en las predicciones del modelo.

Umbrales de Aceptabilidad: Se establecerán umbrales específicos para el MAE que determinarán la aceptabilidad del modelo. Por ejemplo, un MAE inferior a cierto valor (por ejemplo, 5 unidades) puede considerarse como indicativo de un buen ajuste del modelo a los datos.

4.1.3.5.3. CONSIDERACIONES ADICIONALES EN EL ANÁLISIS DE MAE

Evitar el Sobreajuste: Se prestará especial atención para asegurar que un bajo MAE no sea resultado de sobreajuste, especialmente al evaluar el rendimiento del modelo en el conjunto de datos de prueba.

Complementar con Otras Métricas: El MAE se utilizará en conjunto con otras métricas de rendimiento, como el R^2 y el análisis de residuos, para obtener una evaluación más completa del modelo.

4.1.3.5.4. APLICACIÓN PRÁCTICA DE MAE EN EL PROYECTO

Evaluación Continua: El MAE se calculará y monitorizará a lo largo de diferentes fases del desarrollo del modelo para asegurar que el modelo se mantenga relevante y preciso a medida que se disponga de nuevos datos o se realicen ajustes en el modelo. La retroalimentación continua y la optimización del modelo serán fundamentales para garantizar su eficacia a lo largo del tiempo.

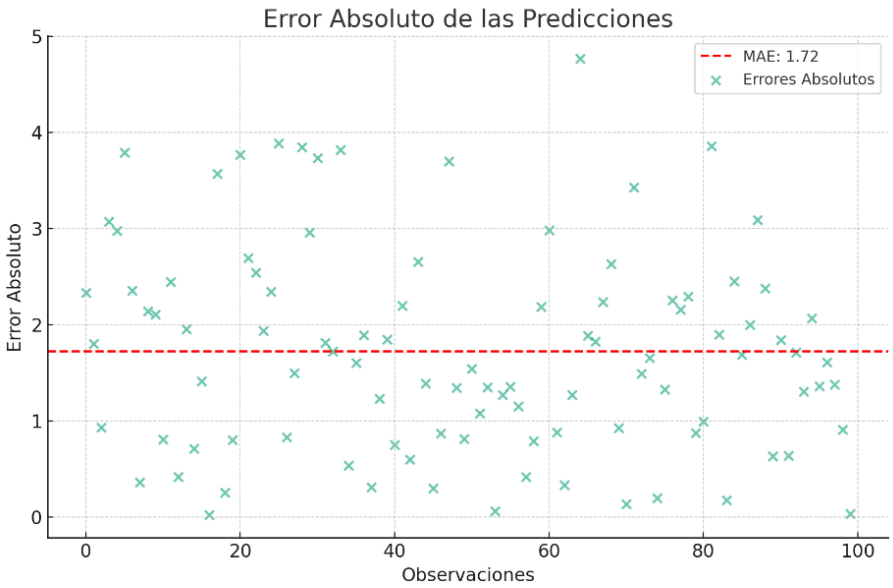

ILUSTRACIÓN 8. ERROR ABSOLUTO MEDIO (MAE)

En la Ilustración 8. Error Absoluto Medio (MAE), se representa visualmente el Error Absoluto Medio (MAE) utilizando un conjunto de datos de ejemplo. Cada punto en el gráfico muestra el error absoluto (la diferencia absoluta entre el valor real y la predicción) para una observación específica. La línea roja discontinua indica el valor medio de estos errores, es decir, el MAE.

Este gráfico ayuda a entender cómo el MAE proporciona una medida de la magnitud promedio de los errores en un modelo de regresión. Un MAE más bajo, indicado por la línea roja más cerca del eje horizontal, señalaría una mayor precisión en las predicciones del modelo.

La fórmula para calcular el MAE es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Donde n es el número total de observaciones, y_i es el valor real y $\hat{y}_i$ es la predicción correspondiente.

4.1.3.6. Balanceo de Datos en problemas de Regresión.

En el ámbito de la modelización estadística y el aprendizaje automático, uno de los desafíos más significativos es el manejo de conjuntos de datos desbalanceados, especialmente en tareas de regresión. Los modelos de regresión, que buscan predecir un valor continúo basándose en una o varias variables independientes,

pueden verse afectados negativamente por desequilibrios en los datos. Estos desequilibrios pueden llevar a sesgos en las predicciones, donde el modelo tiende a realizar mejores predicciones para los valores más comunes, pero falla en predecir con precisión los valores menos frecuentes o 'raros'. Para abordar esta problemática, se utilizan estrategias como SMOTER, SMOGN, el Submuestreo Aleatorio (Random Undersampling), el Sobremuestreo Aleatorio (Random Oversampling) y la Estrategia de Combinación Basada en Relevancia Ponderada (Weighted Relevance-Based Combination Strategy). Estas técnicas buscan equilibrar los datos, permitiendo así que los modelos de regresión aprendan de manera más efectiva y proporcionen predicciones más precisas y menos sesgadas. [24]

4.1.3.6.1. SMOTER (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE FOR REGRESSION)

Descripción: SMOTER es una adaptación del conocido algoritmo SMOTE para regresión, abordando la problemática específica de los datos de regresión desbalanceados [25].

Funcionamiento: Define regiones mayoritarias (frecuentes) y minoritarias (raras) basándose en la densidad original de etiquetas, aplicando submuestreo aleatorio en la región mayoritaria y sobre muestreo en la minoritaria [25].

Técnica de Sobre muestreo: Genera ejemplos sintéticos y emplea una estrategia de interpolación combinando entradas y objetivos de diferentes ejemplos. Utiliza dos casos raros para la interpolación, creando variables objetivo como un promedio ponderado de los casos raros utilizados [25].

4.1.3.6.2. SMOGN (SMOTE AND GAUSSIAN NOISE)

Descripción: SMOGN se basa en SMOTER, pero añade ruido gaussiano durante la fase de sobre muestreo. Se propone como una nueva estrategia de preprocesamiento para lidiar con la regresión desbalanceada [26] [27].

Funcionamiento: Combina las estrategias de SMOTER y ruido gaussiano para la generación de ejemplos sintéticos. Utiliza SMOTER para generar ejemplos sintéticos cuando los ejemplos seleccionados están cercanos entre sí, y aplica ruido gaussiano cuando están más distantes [26] [27].

Personalización y Efectos: SMOGN ofrece aproximadamente tres etapas de personalización para el proceso de remuestreo, que incluyen configuraciones automáticas, automáticas dentro de límites y manuales. La aplicación de SMOGN,

junto con transformaciones logarítmicas, ha demostrado mejorar la cobertura del rango de valores objetivo. En un caso práctico, SMOGN redujo el número total de observaciones en aproximadamente un 20% al aumentar el muestreo de observaciones raras con ruido y disminuir el muestreo de las frecuentes [26] [27].

Ventajas e Impacto: SMOGN se integra con propuestas existentes para resolver problemas detectados en ambas, mostrando ventajas en comparación con otros enfoques. Tiene un impacto diferencial en los algoritmos de aprendizaje utilizados, mostrando más ventajas para aprendices como Random Forest y Multivariate Adaptive Regression Splines [25].

Submuestreo para regresión, (Undersampling Regression): Esta estrategia reduce el tamaño del dominio 'normal' (generalmente la clase mayoritaria en un conjunto de datos) mientras mantiene inalterado el dominio 'raro' (clase minoritaria). El grado de submuestreo se puede controlar de varias maneras [28]:

Balance: Iguala el tamaño de los dominios normal y raro.

Extremo: Invierte la proporción de tamaños entre los dominios normal y raro.

Promedio: Un enfoque intermedio entre balance y extremo. Se suele usar la búsqueda en malla para determinar el grado óptimo de submuestreo, para maximizar métricas como el puntaje F1.

Sobremuestreo para regresión, (Oversampling Regression): Este enfoque mantiene sin cambios el dominio normal y aumenta el tamaño del dominio raro mediante duplicación. El grado de sobremuestreo también se puede configurar como balance, extremo o promedio. El grado óptimo se suele encontrar en un proceso de búsqueda en malla [28].

Estrategia de Combinación Basada en Relevancia Ponderada (WERCS, Weighted Relevance-Based Combination Strategy): A diferencia de UR y OR, WERCS no categoriza los valores objetivo en dominios normal y raro. En su lugar, utiliza valores de relevancia como pesos para decidir qué muestras submuestreo o sobre muestrear. Las instancias con altos valores de relevancia tienen más probabilidades de ser elegidas para sobre muestreo (y menos para submuestreo). Los usuarios pueden especificar el porcentaje de sobre muestreo y submuestreo. A diferencia del sobre muestreo estándar que duplica las muestras, WERCS puede agregar ruido gaussiano a las muestras para crear variabilidad [28].

4.2. Antecedentes

En los sistemas de transporte masivo, aunque no se ha realizado un proyecto exactamente igual, existen iniciativas similares que incorporan tecnologías

avanzadas como el Machine Learning y la Inteligencia Artificial, alineadas con la Industria 4.0. Un ejemplo destacado es el proyecto de la Unión Temporal de Recaudo y Tecnológica en Cali, Valle del Cauca, que ha implementado un Sistema de Reconocimiento Facial en el SITM MIO para reforzar la seguridad. Este contexto se enriquece con proyectos que explican los aspectos fundamentales del Machine Learning y su aplicación en la industria, como el "Estado del Arte de la Inteligencia Artificial en el Sector Ferroviario" y el "Informe al Congreso de la República Sector Transporte 2020 - 2021". Estos documentos resaltan cómo la transformación digital a través de tecnologías emergentes puede optimizar la asignación de recursos y mejorar la eficiencia y seguridad en sistemas de transporte como el SITM MIO.

4.2.1. Utilización del machine Learning en la industria 4.0.

Este proyecto proporciona información detallada y relevante sobre la Cuarta Revolución Industrial (Industria 4.0) [29] y el papel del aprendizaje automático (Machine Learning) en este contexto. Aquí hay un resumen de los puntos clave útiles como antecedentes sobre la asignación de recursos institucionales en el SITM MIO:

- **Industria 4.0 y la Transformación Digital**: La Industria 4.0, también conocida como la Cuarta Revolución Industrial, se caracteriza por la fusión de las tecnologías digitales y físicas, como IoT, Big Data, Cloud Computing, y la inteligencia artificial, para crear fábricas inteligentes y sistemas ciber-físicos. Esta transformación digital es crucial para que las empresas se mantengan competitivas en un mercado globalizado donde se demanda personalización y alta calidad en los productos.

- **Objetivos del Documento:** El documento tiene como objetivo principal explicar los aspectos fundamentales del Machine Learning y su aplicación en el sector industrial bajo el paradigma de la Industria 4.0. Esto incluye comprender las tecnologías habilitadoras de la Industria 4.0, conocer la situación de la digitalización en España, establecer un marco teórico sobre el Machine Learning y mostrar sus aplicaciones en distintas áreas de una empresa industrial.

- **Estructura del Documento:** El documento se divide en tres capítulos principales. El primer capítulo aborda la Industria 4.0 y las tecnologías habilitadoras como el Cloud Computing y el IoT. El segundo capítulo se enfoca en el Machine Learning, sus variantes y aplicaciones. El tercer capítulo discute la importancia del Machine Learning en la industria y sus aplicaciones prácticas en áreas como producción, logística, mantenimiento, y recursos humanos.

- **Aplicaciones y Alcance del Machine Learning**: El Machine Learning se aplica para el reconocimiento de patrones y aprendizaje en diversos campos, incluyendo la detección de malware, identificación de células cancerígenas, sistemas de recomendación, y vehículos autónomos. Estas aplicaciones demuestran la capacidad del Machine Learning para trabajar con grandes

volúmenes de datos y extraer patrones, relaciones, tendencias y predicciones valiosas

La relación de este antecedente con el proyecto de asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO, busca aplicar el aprendizaje automático en un contexto específico de transporte y gestión de infraestructuras, lo que se alinea con los principios de la Industria 4.0 y el uso innovador del Machine Learning. Las ideas y estrategias presentadas en el documento proporcionan una base teórica sólida y ejemplos prácticos de cómo integrar el Machine Learning en sistemas complejos, fundamental para contextualizar sobre cómo optimizar la asignación de recursos y mejorar la eficiencia y seguridad en el SITM MIO.

4.2.2. Estado del arte de la inteligencia artificial en el sector ferroviario poniendo en común fabricante, operadora e infraestructura

El documento "Estado del Arte de la Inteligencia Artificial en el Sector Ferroviario" [30] proporciona información valiosa sobre el uso de aprendizaje automático en la asignación de recursos institucionales para el control y seguridad de infraestructuras. Citamos algunos puntos clave del documento que pueden ser relevantes:

- **Aplicación de la Inteligencia Artificial en Diversos Sectores**: El documento señala el creciente interés y la expansiva aplicación de la inteligencia artificial en varios sectores, incluyendo el transporte. Específicamente, en el sector ferroviario, se destaca la inteligencia artificial como un medio para mejorar la eficiencia y sostenibilidad, lo que resalta la versatilidad y la importancia de esta tecnología en contextos de infraestructura y movilidad.

- **Objetivos del Estudio:** El principal objetivo del trabajo es evaluar el estado del arte de la inteligencia artificial en el sector ferroviario. Se enfoca en identificar áreas que requieren más investigación o desarrollo para su aplicación eficiente en problemas de ingeniería relacionados con el transporte ferroviario. Además, busca analizar cómo se resuelven los problemas en el sector, ya sea individualmente o mediante un enfoque colaborativo entre los actores principales (operadoras, infraestructura, y fabricantes).

4.2.3. Informe al congreso de la república sector transporte 2020 – 2021

La relación de este trabajo el proyecto sobre el uso de aprendizaje automático en la asignación de recursos institucionales para el control y la seguridad de la infraestructura móvil, física y tecnológica del SITM MIO es directa. Al igual que en el sector ferroviario [31], el proyecto busca aprovechar las capacidades del aprendizaje automático para mejorar la eficiencia y seguridad en un sistema de

transporte. Las estrategias y hallazgos del documento proporcionan insights valiosos sobre cómo implementar tecnologías de inteligencia artificial en un contexto de infraestructura y movilidad, ofreciendo posibles vías para optimizar la gestión de recursos y la operatividad en el SITM MIO.

El "Informe de la ministra al Congreso 2020 - 2021" del Ministerio de Transporte de Colombia presenta aspectos relevantes, considerados como antecedentes sobre el modelo de aprendizaje automático aplicado a la asignación de recursos institucionales en el SITM MIO.

- **Enfoque en la Modernización del Transporte y la Logística:** El Plan Nacional de Desarrollo 2018 – 2022 destaca la importancia de modernizar el sector transporte para mejorar la eficiencia, seguridad y sostenibilidad. Esta modernización implica la adopción de tecnologías avanzadas, que busca integrar el aprendizaje automático para optimizar la asignación de recursos en el transporte público, contribuyendo así a una operación más eficiente y segura.
- **Innovación y Tecnología en la Gestión del Transporte:** El uso del módulo INSIDE RNDC por el gobierno demuestra el compromiso con la adopción de tecnología para mejorar la logística y eficiencia en el transporte. Esto resalta la relevancia de implementar soluciones tecnológicas avanzadas, como el aprendizaje automático, para la recolección y análisis de datos que pueden mejorar la gestión del transporte público en aspectos como tiempos, seguridad y eficiencia operativa.
- **Priorización de la Seguridad en el Transporte:** La reducción de siniestros viales y la mejora de la seguridad en la infraestructura de transporte son prioridades claras en el informe. Al aplicar el aprendizaje automático para la asignación de recursos, puede contribuir significativamente a estos objetivos, mejorando la seguridad y eficiencia en la infraestructura móvil, física y tecnológica del SITM MIO, alineándose así con las políticas nacionales de seguridad en el transporte.

En resumen, el informe del Ministerio de Transporte proporciona un marco contextual importante, demuestra cómo la adopción de tecnologías avanzadas, en especial el aprendizaje automático, puede jugar un papel crucial en la modernización y mejora de la eficiencia y seguridad en el sector del transporte. Esto valida la dirección del proyecto y el potencial para contribuir a los objetivos nacionales de desarrollo y modernización en el sector de transporte público.

4.2.4. Big data e internet de las cosas para los sistemas inteligentes del transporte

El documento "Big Data e Internet de las Cosas para los sistemas inteligentes del transporte" [32] ofrece antecedentes valiosos sobre el uso de aprendizaje automático en la asignación de recursos institucionales para el control y seguridad de la infraestructura móvil, física y tecnológica del SITM MIO:

- **Aplicación de Big Data en Transporte:** El documento destaca la importancia de Big Data en la transformación digital y la Industria 4.0, resaltando cómo estas tecnologías han permitido superar barreras previas en el procesamiento de datos, detección de patrones ocultos y procesamiento de imágenes en tiempo real. Estos avances son relevantes para tu proyecto, ya que la capacidad de procesar y analizar grandes cantidades de datos es fundamental para optimizar la asignación de recursos en sistemas de transporte como el SITM MIO.

- **Revisión Bibliográfica y Tendencias en Tecnologías de Transporte:** El estudio proporciona una revisión exhaustiva de las tecnologías orientadas al transporte, destacando las tendencias actuales y áreas de oportunidad, especialmente en México. Esta información es útil para identificar cómo se pueden aplicar tecnologías, adaptándolas a las necesidades específicas del SITM MIO.

- **Transformación Digital en Organizaciones de Transporte:** La introducción de nuevas tecnologías y la adopción de nuevos modelos de operación y negocio son cruciales en todos los sectores, incluido el transporte.

- **Aplicaciones Prácticas y Metodologías:** El documento menciona aplicaciones prácticas de Big Data, como la predicción de tráfico, detección de congestión, y optimización de tiempos de respuesta en emergencias. Estas aplicaciones demuestran cómo el aprendizaje automático puede usarse para resolver problemas reales en el transporte, relacionado directamente con el proyecto.

- **Oportunidades de Aplicación y Desarrollo en Ciudades Inteligente**s: El documento identifica varias áreas de oportunidad para aplicar tecnologías avanzadas en el contexto de ciudades inteligentes. Sugiere formas de utilizar el aprendizaje automático y la ciencia de datos para mejorar la infraestructura de transporte y la gestión de tráfico en un entorno urbano.

- **Aplicación de Algoritmos Avanzados en el Transporte**: Se observa un incremento en el uso de algoritmos de aprendizaje automático para mejorar la eficiencia y seguridad en el transporte. Esto incluye algoritmos como K-Means, redes neuronales artificiales (RNAs), y heurísticos. Además, se destaca el uso de la nube y la clusterización con Hadoop, así como la integración de dispositivos inteligentes y algoritmos avanzados de IA para la gestión de datos en transporte.

En resumen, este documento proporciona un marco de referencia sobre cómo el Big Data, y el aprendizaje automático están siendo aplicados en el sector del transporte. Este marco es altamente relevante, ofreciendo insights sobre cómo

estas tecnologías pueden ser utilizadas para mejorar la asignación de recursos y la seguridad en sistemas de transporte como el SITM MIO.

5. Metodología de proyecto

5.1. Objetivo específico 1

En este proyecto de maestría, se logró desarrollar el objetivo específico centrado en la preparación y el manejo óptimo de los datos para el análisis de vandalismo en el sistema de transporte masivo del distrito de Santiago de Cali SITM MIO. Esta fase del proyecto fue fundamental, ya que estableció la base sobre la cual se realizaron los análisis de los conjuntos de datos, a continuación, se detallan las etapas clave y las metodologías empleadas en este proceso.

5.1.1. Recolección de Datos

La recolección de datos es un proceso fundamental en cualquier proyecto de investigación o análisis. Para el sistema de transporte SITM MIO de Metro Cali, la fase inicial de recolección de datos implicó las siguientes acciones:

1. **Reuniones con Metro Cali:** Se realizaron reuniones con representantes de Metro Cali para comprender a fondo las necesidades de información y los desafíos específicos del sistema de transporte. Estas interacciones fueron cruciales para identificar las fuentes de datos más relevantes y para entender el contexto operativo y de seguridad en el que se desarrolla el SITM MIO.

2. **Identificación de fuentes de información:** Se determinaron las fuentes externas adecuadas para proporcionar la información necesaria, teniendo en cuenta la naturaleza confidencial de los datos.

3. **Solicitudes de acceso:** Se gestionaron solicitudes internas a nivel administrativo a la entidad Metro Cali para obtener las autorizaciones requeridas para acceder a la información.

5.1.1.1. Fuentes de Información y Estructura de Datos.

La gestión y acceso a las fuentes de datos se llevaron a cabo a través de permisos y autorizaciones dados por Metro Cali. La entidad gestora permitió el acceso a cada una de las fuentes y consolidar la información necesaria según las necesidades del proyecto. Las fuentes de datos consultadas fueron las siguientes:

1. **Unión Temporal de Recaudo y Tecnología (Utryt):** Esta entidad privada gestiona los conjuntos de datos fundamentales para el proyecto, utilizando infraestructura, hardware y software para facilitar el recaudo en el sistema MIO mediante tarjetas inteligentes y realizar operaciones adicionales como

la seguridad digital a través del centro de control. La Utryt ha colaborado con Metro Cali en la recolección y validación de actos inseguros y delitos en las estaciones y terminales del sistema de transporte MIO. Los datos se almacenan en una base de datos SQL en un entorno Oracle, y la Utryt proporciona bases de datos llamada 'Vandalismos en Estaciones y Terminales' y 'Evasión de Pago' en tablas de Excel en formato XLS.

2. **Plataforma web de la Utryt:** Ofrece un acceso seguro a ciertos conjuntos de datos. A través de esta plataforma, se obtuvo la información para la base de datos "Usos en terminales y estaciones". Debido a problemas técnicos, la descarga de datos se segmentó en varios archivos CSV, que posteriormente se consolidarán en un solo conjunto de datos.

5.1.1.2. Integración de datos.

Para integrar y consolidar la información proveniente de las distintas fuentes, se implementaron los siguientes pasos:

1. **Segmentación y Consolidación:** Los problemas técnicos presentados en la plataforma web para la descarga de conjunto de datos "Usos en Estaciones y Terminales" conllevo a que su descargar se hiciera de forma segmentada por mes. Por lo tanto, se estableció un proceso para consolidar estos segmentos en un único conjunto de datos.

2. **Normalización de Formatos:** Debido a que los distintos conjuntos de datos tenían extensiones de archivos diferentes. Se trabajó en la normalización de los distintos formatos de los conjuntos de datos, pasando de XLS y XLXS a un formato con extensión CSV para unificar los datos en un solo archivo que facilitara el análisis y la gestión de la información.

5.1.1.3. Descripción de los conjuntos de datos obtenidos por las fuentes.

Los conjuntos de datos recopilados de diversas fuentes están organizados por horas. Por lo tanto, presentamos una descripción detallada de estos conjuntos de datos, que incluyen incidentes de vandalismo, casos de evasión de tarifas y patrones de uso de las estaciones y terminales, todos ellos registrados y consolidados cronológicamente para facilitar el análisis y la toma de decisiones en el proyecto.

1. **Conjunto de datos "vandalismo en estaciones y terminales":** Se identificó este conjunto de datos contiene registros del vandalismo diario en estaciones y terminales, tales como robos, ataques, agresiones y atracos. Esta información resultó esencial para seleccionar las entradas de datos adecuadas y crear una secuencia lógica y beneficiosa para el proyecto.

Este conjunto de datos contiene información detallada sobre los eventos de seguridad que ocurren en las terminales y estaciones del sistema de transporte MIO. Aunque la base de datos incluye diversos tipos de información, algunos no son relevantes para nuestro análisis, como los registros de 'primeros auxilios a usuarios', que no se relacionan directamente con delitos y, por lo tanto, no contribuyen significativamente al estudio. La base de datos está compuesta por 19 columnas, cada una establecida y relacionada de la siguiente manera:

- **Numero de Caso (numérico):** Numero de la asignación que se genera automáticamente.
- **Agente Registra (texto):** Nombre de la persona que registra el evento en base de datos, tipo texto.
- **Jerarquía (texto):** Se consolidad los eventos como en el formato: **CENTRO CONTROL SIUR:SEGURIDAD:DELITOS:ATRACO**. En este formato nos indica de que tipo fue el evento (**SEGURIDAD**), luego en tipo de inseguridad (**DELITO**) y por último la acción del delito (**ATRACO**). Esta columna es unas de las más relevante para el análisis.
- **Estado (texto):** Si fue atendido el caso o no. Toma el estado de "Cerrado" o "En proceso"
- **Fecha de Creación (fecha):** Es la fecha de cuando se presentó el evento.
- **Hora de Creación (hora):** Hora en la que sucedió el evento.
- **Tipo (texto):** Quien consolida la información, normalmente es R&T.
- **Cola (texto):** Jerarquía de quien es el responsable de dar respuesta al evento.
- **Asunto (texto):** Describe el tipo de evento o vandalismo.
- **Propietario (texto):** Igual del agente que registra.
- **Descripción Caso (texto):** Descripción del suceso de como cuando y donde ocurrió.
- **Utima Nota (texto):** Lo mismo que Descripción de caso.
- **Impacto (texto):** Tipo de prioridad.
- **Solución (texto):** Describe como se solucionó el evento.
- **Fecha de Solución (fecha):** Fecha de cuando se dio solución del evento.
- **Solución Mínima (numérico):** Tiempo mínimo para dar respuesta.
- **Nivel de servicio (texto):** Tipo de servicio prestando.
- **Tiempo nivel Servicio Min (numérico):** Tiempo máximo para dar respuesta.
- **Ubicación (texto):** En que terminal se presentó evento.

2. **Conjunto de datos "Evasión en terminales y estaciones:** Este conjunto de datos es crucial por su vínculo directo con la seguridad en el SITM MIO. Una parte importante de los incidentes de seguridad se asocia con la evasión de tarifas. Entender estos patrones es clave para crear tácticas efectivas que mitiguen la evasión y fortalezcan la seguridad en el sistema de transporte.

Este conjunto de datos posee una estructura similar al conjunto 'Vandalismo en estaciones y terminales'. La principal diferencia radica en la columna Jerarquía, que sigue el formato: **CENTRO CONTROL SIUR:SEGURIDAD:EVASION: PUERTA TELESCOPICA:USUARIOS**. Este formato nos indica la categoría del evento (**SEGURIDAD**), el tipo de incidente (**EVASION**), el lugar específico donde ocurrió (**PUERTA TELESCOPICA**), y el sujeto involucrado (**USUARIO**). Al igual que el conjunto mencionado anteriormente, esta base de datos también consta de 19 columnas.

3. **Conjunto de datos "Uso en Estaciones y Terminales":** Esta base de datos está compuesta por 8 columnas y está estructurada por las siguientes columnas:

   - **Estación (texto):** Se encuentra el nombre de la estación y terminal como de igual forma el nombre de los buses cuando se accede a un punto de transbordo por fuera de una estación o terminal.
   - **Producto (texto):** Tipo de pago que se realizó cuando se accede a la terminal, estación o flota.
     Integración: Es una constante que especifica que nivel de integración se realizó por ejemplo si fue un transbordo, etc.
   - **Fecha (fecha):** Fecha en la que se accedieron los usuarios a las estaciones.
   - **Hora (hora):** La hora en la que se accedió a la estación terminal, estación o bus.
   - **Veh_id (numero):** Código de la estación, terminal o bus.
   - **Cu_farevalue (numero):** Una constante de estimación del pasaje.
   - **Qpax (numero):** Cuantas personas accedieron a la estación, terminal o bus en la fecha y hora especifica.

5.1.1.4. Base datos de apoyo.

1. **Conjunto de datos "Reconstrucción de la serie de tiempo":** Las bases de datos obtenidas de diversas fuentes no reflejan un comportamiento de serie de tiempo, ya que únicamente registran el momento en que ocurre un evento. Por esta razón, fue crucial desarrollar un conjunto de datos que funcionara como referencia de serie de tiempo, permitiendo así reconstruir el comportamiento temporal de cada conjunto de datos. Esta nueva base está compuesta por los campos **Fecha**, **Hora** y **Día Hábil,** y cubre el período desde el 1 de enero de 2022 hasta el 30 de octubre de 2023. El campo '**Día Hábil**' distingue entre días como festivos, domingos y sábados (valor 1), y los demás días (valor 0).

2. **Conjunto de dato "Ubicación estación":** Los conjuntos de datos provenientes de las distintas fuentes presentan inconsistencias en los nombres de las estaciones, con cada base de datos utilizando nomenclaturas diferentes. Además, no se cuenta con la georreferenciación de cada una de

las estaciones y terminales del sistema MIO. Por lo tanto, se procedió a crear una base de datos unificada de 4 columnas. que contiene:

- **Nombre:** Contiene los diferentes nombres de las estaciones y terminales presente dentro de los conjuntos de datos de "**vandalismo estación y terminales**", "**Evasión en estaciones y terminales**" y para termina "**Usos en estaciones y terminales**"
- **Ubicación:** Contiene el nombre normalizado o a remplazar de las estaciones y terminales para llamarlos a todos de una misma forma. Por ejemplo: En algunos conjuntos de datos tenemos "**7 agosto**" y en otros "**Siete de Agosto**" por lo tanto a estos nombres los convertimos como "**7 de Agosto**".
- **Latitud:** Contiene la coordenada de latitud de la estación y terminal.
- **Longitud:** Contiene la coordenada de Longitud de cada estación y terminal.

5.1.1.5. Etiquetas de salida.

La consolidación de las etiquetas relacionadas con "**cuántos policías se requieren"** se realizó a partir de un texto no estructurado proporcionado por la Policía Nacional (PONAL) a Metro Cali. Este documento contiene información sobre la fecha y la estación donde se prestó el servicio policial. Adicionalmente, se utilizó otra fuente de información para las etiquetas, realizando el análisis de la columna "**asunto**" en el conjunto de datos de "**Vandalismo en estaciones y terminales**". Con la colaboración de la UTRYT, se identificó que esta columna incluía los requerimientos de presencia policial y el esquema presente de la cantidad de policía se corroboro directamente con PONAL.

5.1.2. Limpieza de Datos

La metodología de limpieza de datos para los conjuntos de Evasión, Vandalismo y Usos por Estación y Terminal en el SITM MIO se desarrolló de la siguiente manera:

- **Identificación y Corrección de Errores:** Se localiza y corrige errores o incoherencias en los datos. Esto incluyó la corrección de formatos de fecha y hora, la unificación de la nomenclatura para las estaciones y terminales, y la normalización de las unidades de medida. Estas correcciones fueron fundamentales para asegurar la coherencia y comparabilidad de los datos a través de todas las estaciones y terminales.

- **Normalización:** Se identificaron y corrigieron inconsistencias en los formatos de los datos. Esto incluyó la estandarización de formatos de fecha y hora, la unificación de la nomenclatura para las estaciones y terminales, y la normalización de las unidades de medida. Estas correcciones fueron fundamentales para asegurar la coherencia y la homogeneización de

terminologías en todos los conjuntos de datos que relaciona las estaciones y terminales.

- **Eliminación de Duplicadas:** Se detectaron y eliminaron los registros duplicados que resultaban de integraciones múltiples. Estos duplicados podrían surgir por errores en la recopilación de datos o en los procesos de integración de datos de diferentes fuentes. La eliminación de estos duplicados fue esencial para proporcionar un recuento preciso y fiable de los usuarios que transitan por cada estación y terminal.

- **Eliminación de columnas:** Se procedió a la eliminación de diversas columnas que carecían de relevancia en cuanto al avance del proyecto y al análisis de datos. Este proceso contribuyó significativamente a reducir las dimensiones de cada conjunto de datos, generando así un manejo más eficiente de la información. Como consecuencia directa de esta depuración, se logró no solo una optimización en la gestión de datos, sino también una notable disminución en la utilización de recursos físicos, como la memoria RAM en los equipos de cómputo. Este enfoque tuvo como resultado en una mejora sustancial en la eficacia y rendimiento general del sistema.

- **Extracción de datos: Dentro de las columnas:** En el proceso de refinamiento de datos, se aplicaron patrones de texto para llevar a cabo la extracción de información pertinente contenida en las columnas. La finalidad principal de este procedimiento fue la selección exclusiva de la información relevante, contribuyendo así a optimizar las dinámicas del análisis. De esta manera, se ha buscado utilizar únicamente la información que aporta valor, permitiendo una mayor eficiencia en la interpretación y aplicación de los datos. Este enfoque estratégico no solo ha facilitado la identificación y utilización de datos clave, sino que también ha mejorado significativamente la calidad y utilidad de la información extraída de cada columna en el contexto del análisis en curso.

- **Reconstrucción:** Se opta por la base de datos de "Reconstrucción de serie temporal" y, mediante una correspondencia adecuada, se establece la relación entre los datos y sus respectivas fechas y horas en cada conjunto de datos. Las horas y días presentes en el conjunto de datos se asocian con el valor cero, indicando así la ausencia de cualquier tipo de delito, uso indebido o evasión en esos momentos específicos

- **Filtro de rango de hora: La** información de los diversos conjuntos de datos se presenta en un formato temporal que abarca las horas del día, iniciando desde las 00:00 hasta las 23:00 después de realizar la “Reconstrucción”. Se tomó la decisión de mantener una unidad de tiempo constante, dado que todos los conjuntos de datos comparten la misma estructura. En consonancia con el horario de inicio de operaciones del sistema de transporte, se definió trabajar con el rango horario desde las 5:00 hasta las 23:00 horas. Esta

elección busca asegurar coherencia y facilitar la comparación entre los datos, manteniendo una perspectiva temporal homogénea para el análisis/

5.1.3. Experimento de conformación del dataset

1. **Creación de DataFrame de Tiempo Completo**: Se crea un DataFrame que incluye todos los días y todas las horas desde el inicio de 2022 hasta el 30 de noviembre del 2023. Esto proporciona una plantilla sobre la cual se pueden mapear o 'matchear' los eventos reales, lo que significa que cualquier evento registrado se alinea con la estructura temporal del DataFrame.

2. **Matching de Eventos con el DataFrame:** Los eventos existentes en las estaciones se alinean con el marco de tiempo mencionado. Por ejemplo, si hay un registro de vandalismo en una estación a las 15:00 el 5 de marzo de 2022, este evento se coloca en la correspondiente ubicación de fecha y hora en el DataFrame para conformar el dataset base.

3. **Estructura del dataset base:** Cuando sea realizado, el 'Matching' de los distintos eventos para conforma el dataset base, obtenemos un nuevo dataset formada por 8 columnas donde las filas esta organizadas de forma temporal y ordenada desde el 1 de enero del 2022 hasta 30 de noviembre 2023, organizados por día desde la hora 00:00 hasta 23:00.

   El resultado final es un dataset base donde cada fila representa una hora específica que contiene una secuencia como, por ejemplo: "1,35,4,0,0,1,0,2" que contiene toda la información de los eventos presentes en una hora especifica. Tomando como ejemplo la fila "1,35,4,0,0,1,0,2", cada dígito tiene el siguiente significado:

   - **Día Hábil o No Hábil (1er digito):** El '1' indica que es un día hábil (no es festivo). Esto podría ser relevante para el análisis, ya que el patrón de uso y los incidentes pueden variar entre días hábiles y no hábiles.

   - **Usos (2do digito):** El '35' indica que en una hora especifica se presentaron 35 personas que usaron la estación en una hora específica.

   - **Evasión (3er digito):** El '4' sugiere que en esa hora cuatro personas accedieron a la terminal o estación sin pagar.

   - **Agresión (4to digito):** El '0' indica que no hubo incidentes de agresión reportados.

   - **Atraco (5to digito):** El '0' indica que no hubo incidentes de atracos reportados.

- **Daños a la Infraestructura (6to digito):** El '1' significa que se presente un reportaron daños a la infraestructura.

- **Robo (7mo digito):** El '0' muestra que no se registraron robos.

- **Cuanto policía se requiere para tratar el delito (8vo digito):** El octavo dígito en nuestros datos representa la cantidad de policías requeridos para manejar un delito en una hora específica y de igual forma es la etiqueta de salida.

El protocolo de seguridad estándar para estos incidentes implica la asignación de dos policías. Sin embargo, dependiendo de la gravedad del delito o del evento en cuestión, este número puede aumentar. Puede que se necesiten 2,4 o más policías para manejar un incidente.

Es importante tener en cuenta que la cantidad de policías asignados puede variar en función de varios factores, incluyendo la naturaleza del delito, el nivel de amenaza percibido y la disponibilidad de recursos en ese momento.

4. **Transformación de datos usando ventana de tiempo:** Para el desarrollo de esta tarea, establecimos dos tipos de metodología del dataset, los cuales se explicarán más adelante con el objetivo de determinar cuál estrategia ofrece los mejores resultados en la estructuración del conjunto de datos.

   Para el desarrollo de los conjuntos de datos, se implementarán dos experimentos que nos permitirán seleccionar el mejor modelo basándonos en las características de R2 y MAE obtenidas durante el entrenamiento de línea base. Una vez seleccionado el modelo, se podrán aplicar otras estrategias para mejorar su rendimiento si es necesario, como la estimación de hiperparámetros. Las dos metodologías que se trabajaran son los siguientes

   El primer experimento se basa en agrupar los dataset de cada una de las terminales de acuerdo con la característica de frecuencia de los vandalismos en las estaciones y terminales, se obtuvieron 17 grupos en total, por lo tanto 17 dataset. Para este enfoque, utilizamos una ventana de tiempo de 19 horas para cada grupo, donde la variable de salida es la cantidad de policías requeridos en la siguiente hora.

   En el marco del segundo experimento, se toma en consideración un único conjunto de datos que engloba todas las estaciones y terminales concatenadas pero codificada con distinta ventana de tiempo. Sin embargo, se introducen variaciones en las longitudes de las ventanas de tiempo, específicamente 6, 12 y 18 horas. Este enfoque implica la manipulación de tres conjuntos de datos distintos durante el desarrollo de este experimento.

Para consolidar los diferentes conjuntos de datos de las estaciones, seguimos un proceso de tres pasos. Primero, codificamos cada estación con la ventana de tiempo establecida. Segundo, eliminamos los duplicados presentes. Luego, concatenamos al final los datos codificados por la ventana de tiempo de otra estación. Este proceso nos permite evitar la superposición de información entre las estaciones y mantener la temporalidad de cada estación.

5.1.7. Experimento de estructura del dataset

Como se mencionó anteriormente, para el desarrollo de los conjuntos de datos, se realizarán dos experimentos que nos permitirán seleccionar el mejor modelo basándonos en las características de R^2 y MAE obtenidas durante el entrenamiento de línea base. Una vez seleccionado el modelo, se podrán aplicar otras estrategias para mejorar su rendimiento si es necesario, como la estimación de hiperparámetros. Los dos experimentos que se trabajaran son los siguientes:

- **Experimento 1: Agrupación de estaciones según las características de la frecuencia de los actos vandálicos:** Mediante el uso de herramientas estadísticas, se crean grupos en función de la frecuencia de los delitos que ocurren en cada estación. En esta metodología, habrá conjuntos de datos con una mayor cantidad de datos debido a que agrupan más estaciones y terminales. Por otro lado, habrá otros conjuntos de datos que no se agrupan con ninguna otra estación o terminal.

- **Experimento 2: Agrupación de todas las estaciones en un solo conjunto de datos:** En este experimento el proceso de agrupación es simple. Se toman el conjunto de datos consolidado y se separan por estaciones para obtener 62 nuevos conjuntos de datos. Una vez separados el conjunto de datos por estación, codificamos cada conjunto de información de forma individual usando la ventana de tiempo. Luego, se concatenan cada conjunto de datos en único dataset. La salida para este experimento es la predicción de la cantidad de policía en la hora siguiente. Este tipo de agrupación es posible gracias a la codificación realizada por la ventana de tiempo, que transforma la serie temporal en un vector. Para este experimento se tiene establecido tres conjuntos de datos con distinta ventana de tiempo que corresponden a 6, 12 y 18 horas.

  Cada línea dentro de un conjunto de datos consolida un vector de características que se vincula con cada estación o terminal, teniendo como base las variables de entrada. Esto asegura independencia en las predicciones. En consecuencia, para llevar a cabo las predicciones mediante el modelo derivado de este experimento, es necesario ingresar los datos de entrada de manera individualizada por estación y conforme al día que se busca predecir.

5.1.7.1. Experimento 1: Agrupación de estaciones según las características de la frecuencia de los actos vandálicos

La estrategia de agrupación propuesta deja de lado la aleatoriedad de los datos de las estaciones y terminales, agrupando los conjuntos de datos en una característica que refleja las relaciones entre ellas. Para el desarrollo de esta metodología, recurrimos a la técnica de Análisis Múltiple por Correspondencia (MCA). Esta es una metodología estadística que facilita la exploración y visualización de las relaciones entre múltiples categorías de datos.

El propósito central de este análisis fue identificar patrones y similitudes entre las distintas estaciones del SITM MIO, basándose en las incidencias de vandalismo y tipo de estación. Este método permitió no solo agrupar estaciones con características similares, sino también entender cómo estos grupos se relacionan entre sí. La técnica de Análisis de Correspondencias Múltiples (MCA) demostró ser eficaz por su habilidad para manejar datos categóricos. Recordemos que, en el conjunto de datos original de vandalismo en estaciones y terminales, los eventos son categóricos, identificados por el nombre del delito a diferencia con el conjunto de datos de uso o evasión, que proporciona información cuantitativa, como cuántas personas utilizaron la estación o terminal en una hora específica, o cuántas accedieron sin pagar.

MCA facilitó la representación de las estaciones y terminales según las variables asociadas en un espacio multidimensional. Esto permitió una interpretación gráfica y visual más clara de las correlaciones entre ellas, basada en la cercanía entre cada uno de los puntos que representan las estaciones y terminales del SITM-MIO. Cuanto más cercanos estén los puntos, mayor será la correlación entre ellos.

En el uso de esta técnica se presentaron hallazgos como valor agregado que proporcionaron insights valiosos para la toma de decisiones y la planificación estratégica dentro del SITM MIO. Por ejemplo, se pudo identificar estaciones que requerían atención prioritaria en términos de medidas de seguridad o mejoras en la gestión de tarifas. Además, este estudio facilitó determinar el camino para futuras investigaciones y aplicaciones de técnicas analíticas similares en otros aspectos del sistema de transporte. De igual forma, se pudo identificar que ciertas estaciones con altas tasas de evasión de tarifas también mostraban altos niveles de incidentes de inseguridad [27].

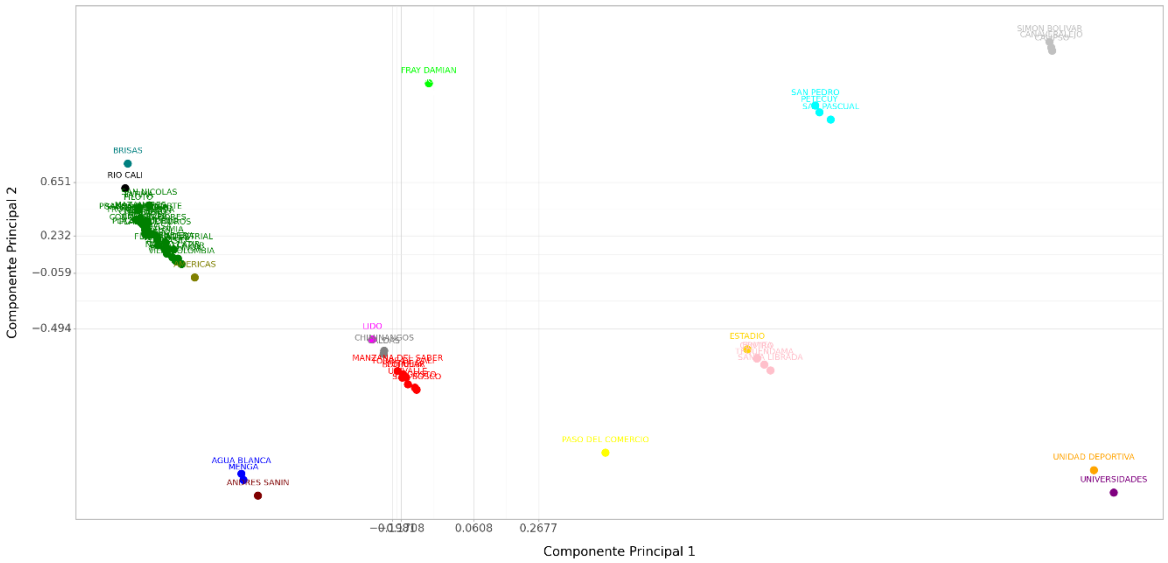

ILUSTRACIÓN 9. AGRUPAMIENTO DE VANDALISMO ESTACIONES DE METRO CALI SITM MIO

5.1.7.1.1. IDENTIFICACIÓN DE LOS GRUPOS

Una vez que sea obtenido la componente 0 y componente 1, como resultado de usar Análisis Múltiple por Correspondencia (MCA), usamos las componentes como coordenada espacial como se observa en la Ilustración 9. Agrupamiento de vandalismo estaciones de Metro Cali SITM MIO.

Para identificar las estaciones se agrupan, calculamos la distancia euclidiana se cada punto (estación) contra todos lo puntos (estaciones) para convertir el componte 0 y 1 en un solo valor que represente la distancia entre los puntos. Entra más pequeña es la distancia euclidiana más cercanía presenta una estación a otra. Como resultado obtenemos una matriz cuadrada donde cada fila consolida las distancias con respecto a una estación especifica.

Una vez obtenida la matriz de relación de distancia, implementamos el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) como solución. Este algoritmo realiza la identificación de clusters basándose en las distancias entre los puntos y la densidad del conjunto de datos. De esta manera, podemos agrupar eficientemente los datos en clúster que reflejen patrones significativos en nuestra información. Como resultado la combinación de la distancia euclidiana y DBSCAN proporcionó una comprensión profunda de la distribución y características de los eventos de vandalismo en el SITM MIO.

Utilizando las técnicas de agrupamiento descritas anteriormente, se clasificó y organizó la información recolectada en la Tabla 1. Asignación de Grupos- Etapa Clusterización que se presenta a continuación.

DATASET	ESTACIONES
Grupo0	7 AGOSTO, BUITRERA, MANZANA DEL SABER, POPULAR, SAN BOSCO, TORRE DE CALI, UNIVALLE
Grupo1	AGUA BLANCA, MENGA
Grupo2	ALAMOS, AMANECER, ATANASIO, BELALCAZAR, CHAPINERO, CIEN PALOS, CONQUISTADORES, FATIMA, FLORA INDUSTRIAL, FLORESTA, MAZANARES, MELENDEZ, NUEVO LATIR, PAMPALINDA, PILOTO, PLAZA CAICEDO, PLAZA DE TOROS, PRADOS DEL NORTE, PRIMITIVO, REFUGIO, SALOMIA, SAN NICOLAS, SANTA MONICA, TREBOL, TRONCAL UNIDA, VERSALLES, VILLA NUEVA, VILLACOLOMBIA, VIPASA
Grupo3	AMERICAS
Grupo4	ANDRES SANIN
Grupo5	BRISAS
Grupo6	CALDAS, CHIMINANGOS
Grupo7	CALIPSO, CANAVERALEJO, SIMON BOLIVAR
Grupo8	CENTRO, ERMITA, SANTA LIBRADA, TEQUENDAMA
Grupo9	ESTADIO
Grupo10	FRAY DAMIAN
Grupo11	LIDO
Grupo12	PASO DEL COMERCIO
Grupo13	PETECUY, SAN PASCUAL, SAN PEDRO
Grupo14	RIO CALI
Grupo15	SUCRE
Grupo16	UNIDAD DEPORTIVA
Grupo17	UNIVERSIDADES

TABLA 1. ASIGNACIÓN DE GRUPOS- ETAPA CLUSTERIZACIÓN

La Tabla 1. Asignación de Grupos- Etapa Clusterización presentada anteriormente es resultado de un análisis de las variables clave para determinar la agrupación adecuada de cada estación. Los criterios considerados incluyen los históricos de vandalismos, que abarcan aspectos como hurtos, evasión de tarifas, daños a la infraestructura, y agresiones. Además, se analizó otro conjunto de datos significativo relacionado con la evasión de cada estación, particularmente, la cantidad de personas que acceden sin pagar; este enfoque metodológico permitió identificar patrones y tendencias específicas asociadas a cada grupo de estaciones.

5.1.7.1.2. ENTENDIMIENTO ADICIONAL IMPORTANTE DEL SITM MIO

Aunque la identificación de características no es el objetivo principal de este trabajo, este proceso permitió comprender el comportamiento de los vandalismos en las estaciones. Esto, a su vez, facilitó la definición de los siguientes conceptos:

- La clusterización facilitó la identificación de patrones ocultos en los datos de las estaciones. Se descubrieron, por ejemplo, agrupaciones de estaciones con altos niveles de vandalismo o con un flujo de pasajeros particularmente elevado.

- La segmentación de las estaciones en grupos homogéneos permitió un enfoque más efectivo en el análisis y la asignación de recursos.

- Se definió un esquema de etiquetado que categoriza las estaciones dentro de cada grupo, y se basa en la cantidad de recursos de policía asignados, correspondiente al número de policías presentes o asignados. Las etiquetas 0, 2, y 4 representan diferentes niveles de asignación de policías y cada grupo tiene una distribución de estas etiquetas basada en los datos del dataset.

5.1.7.1.3. ESTRUCTURA DE LOS DATASET CON RESPECTO A LAS ETIQUETAS

Una vez realizado el agrupamiento de los datos por dataset, se procede a codificar con un nombre asignado como Grupo0 hasta el Grupo17, cada grupo contiene 133 variables de entrada que es el resultado usa una ventana de tiempo de 19 horas.

Etiqueta 0: Indica que no se asignan policías a esa estación o conjunto de estaciones específicas, ya que, según datos históricos o la evaluación de riesgos, no es necesario tener presencia policial en dichos grupos, por una baja incidencia de delitos o actos de vandalismo.

Etiqueta 2: Significa que se asignan dos policías a la estación o grupo de estaciones, debido a una frecuencia intermedia de incidentes o a la importancia estratégica de tener una presencia disuasoria en esa ubicación.

Etiqueta 4: Sugiere una asignación de cuatro policías, lo que implica un nivel más alto de necesidad de seguridad. Las estaciones con esta etiqueta pueden ser puntos de alto riesgo o haber experimentado una alta frecuencia de actos de vandalismo o delitos graves, lo que justifica una presencia policial más robusta.

Por ejemplo, para el **Grupo0**:

- **75348 registros** tienen etiqueta 0, lo que significa que en 99.78% de los casos no se asignarán policías.
- **162 registros** tienen etiqueta 2, indicando que en 0.21% de los casos se asignarán 2 policías.
- **5 registros** tienen etiqueta 4, con un 0.01% de los casos, donde se asignarán 4 policías.

En la Tabla 2. Estructura de datos del experimento 1 se observa estructura de las etiquetas por dataset.

Dataset	Etiquetas			Porcentaje etiquetas		
	0	2	4	0	2	4
Grupo0	75348	162	5	99.78	0.21	0.01
Grupo1	16901	31	0	99.82	0.18	0.00
Grupo2	319499	481	9	99.85	0.15	0.00
Grupo3	12132	17	0	99.86	0.14	0.00
Grupo4	12093	56	0	99.54	0.46	0.00
Grupo5	3133	1	0	99.97	0.03	0.00
Grupo6	22221	56	0	99.75	0.25	0.00
Grupo7	34317	61	1	99.82	0.18	0.00
Grupo8	47530	145	2	99.69	0.30	0.00
Grupo9	12124	23	2	99.79	0.19	0.02
Grupo10	11517	58	1	99.49	0.50	0.01
Grupo11	12115	32	2	99.72	0.26	0.02
Grupo12	9554	12	0	99.87	0.13	0.00
Grupo13	36078	140	6	99.60	0.39	0.02
Grupo14	12143	5	1	99.95	0.04	0.01
Grupo15	12077	71	1	99.41	0.58	0.01
Grupo16	12111	37	1	99.69	0.30	0.01
Grupo17	12098	50	1	99.58	0.41	0.01

TABLA 2. ESTRUCTURA DE DATOS DEL EXPERIMENTO 1

5.1.7.2. Experimento 2: Agrupación de todas las estaciones en un solo conjunto de datos

El proceso de agrupamiento de todas las estaciones y terminales fue una decisión estratégica, tomada tras observar desafíos específicos durante las etapas iniciales del modelado con base a los datos generado en el experimento 1.

En este experimento, unificamos todos los conjuntos de datos en uno solo. Establecimos tres categorías de conjuntos de datos basadas en distintas ventanas temporales: 6 horas, 12 horas y 18 horas. La ventaja de este método radica en su capacidad para integrar todas las estaciones en un único conjunto de datos. Esto se logra mediante la codificación de la serie temporal en un vector de características, lo cual convierte la temporalidad en un formato más manejable. Este procedimiento facilita el análisis de los datos y aumenta la eficiencia al variar los datos en comparación con los conjuntos de datos de la primera metodología y reduce los tiempos de entrenamiento, permitiendo trabajar con mayor agilidad y precisión.

La estructura de las etiquetas y datos para los grupos del experimento 2 se establece de la siguiente forma:

Dataset	Etiquetas			Porcentaje etiquetas		
	0	2	4	0	2	4
Grupo_v6	659392	1415	32	99.781	0.214	0.005
Grupo_v12	670209	1422	32	99.784	0.212	0.005
Grupo_v18	672222	1433	32	99.783	0.213	0.005

TABLA 3. ESTRUCTURA DE DATOS DEL EXPERIMENTO 2

Por ejemplo, para el **Grupo_v18**:

- **672,222 registros** tienen etiqueta 0, lo que significa que en **99.783%** de los casos no se asignarán policías.
- **1,433 registros** tienen etiqueta 2, indicando que en **0.213%** de los casos se asignarán 2 policías.
- **32 registros** tienen etiqueta 4, con un **0.005%** de los casos, donde se asignarán 4 policías.

5.2. Objetivo específico 2

La seguridad en el transporte público, particularmente en estaciones y terminales, es un aspecto crucial para garantizar la integridad de los usuarios y el buen funcionamiento del sistema. En este contexto, el uso de modelos predictivos se presenta como una herramienta estratégica para anticipar incidentes y mejorar proactivamente las medidas de seguridad. Este objetivo se enfoca en una revisión y análisis de características importantes para desarrollar modelos de forma eficaz.

En el contexto de nuestro estudio, hemos seleccionado tres modelos de aprendizaje automático, cada uno con una arquitectura y técnica de aprendizaje distintas. Los modelos que vamos a utilizar para alcanzar nuestro objetivo son los siguientes: Regresión de Bosques Aleatorios (Random Forest Regression), Regresión de Vectores de Soporte (Support Vector Regression) y Regresión de Perceptrón Multicapa (Multilayer Perceptron Regression). Estos modelos fueron seleccionados debido a su capacidad para manejar datos complejos y de alta dimensionalidad, la abundancia de información disponible sobre ellos en la comunidad científica, y la eficiencia que han demostrado en diversas implementaciones.

Para la implementación inicial, adoptamos un enfoque estándar en ciencia de datos para la partición de los datos. Utilizamos el 80% del conjunto para el entrenamiento y el 20% restante para la fase de prueba. El objetivo de esta técnica de división es proporcionar una evaluación justa y equilibrada de la capacidad predictiva de cada modelo, asegurando que el modelo esté bien ajustado, pero no sobre ajustado a los

datos de entrenamiento. De igual forma, para evaluar la precisión de los modelos, usamos las siguientes métricas:

- R^2 o coeficiente de determinación.
- R^2 Ajustado
- MAE o Error Absoluto Medio.

Durante la realización de los entrenamientos, encontramos dificultades al usar la herramienta COLAB de Google. Esto se debió a las constantes advertencias en algunos conjuntos de datos que requerían un uso intensivo de memoria, tanto en la versión gratuita como en la básica de pago. Para entrenar, recurrimos a una laptop con un procesador Core I7-12700H de 12a generación, 16GB de memoria DDR4 a 3200 MHz y una GPU RTX 4070 con 8GB de memoria VRAM.

5.2.1. Entrenamiento de línea base con el experimento de datos 1 y 2

Los experimentos que se describen a continuación se fundamentan en una metodología que prioriza la estructura y arquitectura de los datos, en lugar de enfocarse en la aplicación de técnicas de optimización para mejorar el rendimiento de los modelos. La optimización se llevará a cabo únicamente en aquellos modelos que muestren el mejor rendimiento en la línea base en cada experimento, y solo si se considera necesario. La evaluación del rendimiento se realizará mediante la observación de las métricas R^2 y R^2 Ajustado y MAE en los diferentes modelos. Cuando hablamos de 'línea base', nos referimos a los modelos que se configuran con un conjunto específico de hiperparámetros:

1. **RandomForestClassifier**
   (*n_estimators=100*, *, *criterion='gini'*, *max_depth=None*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features='sqrt'*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *bootstrap=True*, *oob_score=False*, *n_jobs=None*, *random_state=None*, *verbose=0*, *warm_start=False*, *class_weight=None*, *ccp_alpha=0.0*, *max_samples=None*, *monotonic_cst=None*)

2. **SVR**
   (*, *kernel='rbf'*, *degree=3*, *gamma='scale'*, *coef0=0.0*, *tol=0.001*, *C=1.0*, *epsilon=0.1*, *shrinking=True*, *cache_size=200*, *verbose=False*, *max_iter=-1*)

3. **MLPRegressor**
   (*hidden_layer_sizes=(100,)*, *activation='relu'*, *, *solver='adam'*, *alpha=0.0001*, *batch_size='auto'*, *learning_rate='constant'*, *learning_rate_init=0.001*, *power_t=0.5*, *max_iter=200*, *shuffle=True*, *random_state=None*, *tol=0.0001*, *verbose=False*, *warm_start=False*, *momentum=0.9*, *nesterovs_momentum=True*, *early_stopping=False*, *validation_fraction=0.1*, *beta_1=0.9*, *beta_2=0.999*, *epsilon=1e-08*, *n_iter_no_change=10*, *max_fun=15000*)

5.2.1.1. Entrenamiento en línea base del experimento 1

Los resultados del entrenamiento en línea base consolidados en la Tabla 2. Estructura de datos del experimento 1, indicaron que, para ciertos grupos de estaciones, los valores de R^2 eran negativos, lo cual sugiere que los modelos no se ajustaban adecuadamente a los datos, de igual forma, confirmado por la métrica R^2 Ajustado. Sin embargo, los valores de MAE (Error Absoluto Medio) eran relativamente bajos, lo que implica que las predicciones no estaban lejos del valor real en términos absolutos.

Dataset	Random Forest Regression			SVR			MLP Regression		
	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE
Grupo0	-0.158	-0.168	0.012	-1.031	-1.049	0.101	-1.362	-1.383	0.069
Grupo1	-0.137	-0.184	0.011	-0.681	-0.750	0.101	-12.226	-12.767	0.245
Grupo2	0.803	0.802	0.015	-0.073	-0.076	0.107	-0.008	-0.010	0.085
Grupo3	-0.067	-0.129	0.006	-1.035	-1.152	0.100	-206.991	-219.039	0.529
Grupo4	-0.094	-0.158	0.022	-0.319	-0.395	0.107	-147.630	-156.239	1.357
Grupo5	0.000	-0.270	0.007	0.000	-0.270	0.100	0.000	-0.270	0.367
Grupo6	-0.194	-0.231	0.012	-0.874	-0.931	0.102	-39.584	-40.833	0.490
Grupo7	-0.079	-0.101	0.009	-0.951	-0.989	0.103	-455.016	-464.012	1.014
Grupo8	-0.122	-0.138	0.016	-0.663	-0.686	0.105	-65.104	-66.039	0.723
Grupo9	-0.070	-0.132	0.008	-1.020	-1.137	0.099	-557.501	-589.853	1.329
Grupo10	-0.289	-0.367	0.030	-0.275	-0.353	0.086	-0.503	-0.595	0.081
Grupo11	-0.293	-0.368	0.010	-1.341	-1.477	0.100	-31.001	-32.855	0.345
Grupo12	-0.251	-0.344	0.005	-4.313	-4.710	0.097	-909.084	-977.085	1.110
Grupo13	-0.149	-0.170	0.019	-0.541	-0.570	0.106	-11.529	-11.764	0.318
Grupo14	-0.313	-0.389	0.003	-3.640	-3.909	0.077	-3.709	-3.982	0.059
Grupo15	-0.144	-0.210	0.030	-0.224	-0.294	0.098	-0.471	-0.557	0.092
Grupo16	-0.498	-0.585	0.016	-1.078	-1.198	0.102	-431.982	-457.064	1.450
Grupo17	-0.276	-0.350	0.018	-0.636	-0.731	0.104	-347.842	-368.049	1.747

TABLA 4. LABORATORIO Y RESULTADOS LÍNEA BASE

Al analizar cada uno de los resultados del entrenamiento con los dataset consolidada en la Tabla 4. Laboratorio y Resultados línea base, observamos que, en la mayoría de las evaluaciones realizadas a los modelos entrenados, los valores de R^2 y R^2 Ajustado son negativos, lo que indica una correlación inversa entre las variables. De igual forma, alguno de los valores R^2 toma valores que no facilitan su interpretación como los obtenido con el modelo MLP Regression. Sin embargo, este análisis se complementa con el Error Absoluto Medio (MAE). Observamos los siguientes rangos de MAE para cada modelo: Random Forest tiene un MAE entre 0 y 0.030, SVR tiene un MAE entre 0 y 0.16, y MLP tiene un MAE entre 0 y 1.75. Los valores bajos del MAE indican que las predicciones están cerca del valor real, lo

que sugiere que los modelos son precisos a pesar de los valores negativos de R^2 y R^2 Ajustado.

Por otro lado, consideremos los resultados relacionados con el 'Grupo2', el conjunto de datos que contiene la mayor cantidad de información. En el 'Grupo2', los valores de R^2 y R^2 Ajustado para el modelo de Regresión Random Forest muestran un ajuste del 80%, con un MAE de 0.0015. Al revisar la estructura de los datos en la Tabla 4. Laboratorio y Resultados línea base, podemos observar que prácticamente el 99% de los datos, valor 0, están compuestos por frecuencias donde no se requiere policía. Esta misma distribución de datos se presenta en muchos de los conjuntos de datos de los demás grupos, pero no muestra el mismo ajuste. Esto indica que la cantidad de datos es un factor importante para lograr el mejor ajuste posible del modelo.

5.2.1.2. Entrenamiento en línea base con el experimento 2.

Para este experimento se realiza el entrenamiento en línea base de los 3 conjuntos de datos consolidados con ventana de tiempo de 6, 12 y 18 Horas, la salida de este modelo es la predicción de la cantidad de policía de la siguiente hora de acuerdo con la configuración de ventana de tiempo utilizada. Los conjuntos de datos contienen las características individuales de cada estación. Los datos de las pruebas se consolidaron en la siguiente tabla:

Ventana	Random Forest Regression			SVR			MLP Regression		
	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE
6 horas	0.874	0.874	0.032	-0.004	-0.005	0.142	-0.412	-0.413	0.178
12 horas	0.849	0.849	0.033	0.036	0.035	0.138	-0.148	-0.150	0.083
18 horas	0.848	0.848	0.033	0.060	0.058	0.135	-0.165	-0.168	0.127

TABLA 5. PRUEBAS DE USANDO EL DATASET DEL EXPERIMENTO 2

En la Tabla 5. Pruebas de usando el dataset del experimento 2 se puede observar la consolidación de resultados de todos los conjuntos de datos con ventana de tiempo 6, 12 y 18 hora. El resultó mostraron que el modelo de Regresión de Bosques Aleatorios (Random Forest Regression) presento un gran rendimiento con parado con las distintas pruebas realizada en el experimento 1. Este modelo logró un R^2 que explica más del 84% del ajuste de los datos de predicción y un MAE de 0.0032 en los tres conjuntos de datos, un valor relativamente pequeño.

Por otro lado, en cuanto al modelo de Regresión de Vectores de Soporte (SVR), se obtuvo un R^2 de 0 y un MAE que varía entre -0.005 y 0.058. El modelo de Perceptrón Multicapa (MLP) mostró un R^2 en valores negativos y un MAE que se mueve en el rango de 0.083 a 1.178.

El experimento demostró un aumento de rendimiento para el modelo de Regresión de Bosques Aleatorios comparado con el experimento 1 con un R^2 por encima del 0.84. Debido al desbalanceo presente en las frecuencias (observar Tabla 3. Estructura de datos del experimento 2), las predicciones están sesgadas por el valor 0 (no se requiere policía) debido a que es la frecuencia donante. Por lo tanto, es necesario realizar el balanceo de datos con metodologías orientada a balanceo de en problemas de regresión para aumentar la generalidad del modelo.

5.2.1.2.1. BALANCEO DE DATOS

Tras los entrenamientos realizados con los conjuntos de datos de los experimentos 1 y 2, hemos llegado a ciertos descubrimientos clave sobre la importancia del balanceo de datos y la elección del conjunto de datos para aplicar dicha técnica. A continuación, presentamos estos hallazgos y discutimos si es más apropiado aplicar la técnica de balanceo de datos al conjunto de datos del experimento 1 o a el experimento 2:

- El rendimiento superior del modelo Random Forest Regression se evidenció en los conjuntos de datos con mayor concentración de información. Este patrón se observa claramente en la prueba "Grupo2" del Experimento 1 (véase Tabla 2. Estructura de datos del experimento 1 y Tabla 4. Laboratorio y Resultados línea base) y en todas las pruebas del Experimento 2 (véase Tabla 3. Estructura de datos del experimento 2 y Tabla 5. Pruebas de usando el dataset del experimento 2)

- El modelo Random Forest demostró un rendimiento superior en los grupos con mayor concentración. No obstante, se identificó un sesgo en las predicciones, ya que el 99% del conjunto de datos (consulte Tabla 2. Estructura de datos del experimento 1 y Tabla 3. Estructura de datos del experimento 2) indicaba la ausencia de intervención policial. Para mitigar este sesgo, resulta crucial aplicar técnicas de balanceo de datos, garantizando así una representación equitativa de las demás frecuencias que refieren que si se requiere intervención policial.

- La aplicación del balanceo de datos debe llevarse a cabo de manera que existan datos de referencia para realizar un submuestreo adecuado. En el Experimento 2, se observa una mayor concentración de frecuencias (véase Tabla 3. Estructura de datos del experimento 2) donde se requiere presencia policial, lo que contribuye a mejorar la generalidad de las técnicas de equilibrio de datos

Basándonos en lo delineado en los puntos previos, se ha determinado que el experimento 2 constituye con el conjunto de datos más idóneo para implementar las técnicas de balanceo de datos en regresión. Este conjunto de datos integra todas las estaciones en un único conjunto, cada uno con distintas ventanas de tiempo. La técnica de balanceo de datos se llevará a cabo sobre tres conjuntos de datos con ventanas temporales de 6, 12 y 18 horas, respectivamente.

Dentro de este conjunto de datos, se identifican salidas con valores 0, 2 y 4 (ver Tabla 3. Estructura de datos del experimento 2), que corresponden a la cantidad de policías en una estación. El esquema de seguridad del PONAL siempre ha establecido un método de gestión ante novedades mediante un esquema de pares. En este proyecto, que busca sugerir una gestión de este recurso institucional, es esencial detectar cuándo es necesario establecer 1 o 3 policías

5.2.1.2.2. APLICACIÓN DE LA TÉCNICA DE BALANCEO.

Dentro del marco teórico, hemos descrito diversas técnicas para el balanceo de datos en regresión. Destacan SMOTER y SMOGN, técnicas que realizan un submuestreo del conjunto de datos y generan nuevos datos sintéticos para equilibrar las frecuencias. Además, existe otra técnica conocida como Estrategia de Combinación Ponderada basada en Relevancia (WERCS). Esta estrategia ha sido la seleccionada para homogeneizar las muestras en las que no se 'requiere policía' en comparación con las muestras en las que se requieren 2 o 4 policías. Estas técnicas garantizan que los modelos de regresión puedan manejar eficazmente situaciones comunes y menos frecuentes, pero igualmente importantes.

Una de las principales ventajas de utilizar la estrategia WERCS para el balanceo de datos radica en su enfoque en la relevancia. En otras palabras, esta estrategia otorga un mayor peso en el vector de relevancia a las frecuencias minoritarias o raras (aquellas con poca repetición). Esta metodología permite la implementación de técnicas como el sobre muestreo (over-sampling), el submuestreo (under-sampling) y la generación de ruido gaussiano alrededor de los valores de salida mientras se submuestrea la variable de entrada. Estas técnicas permiten un manejo más efectivo de los datos, asegurando que las frecuencias minoritarias se representen adecuadamente y consideren durante el modelado y entrenamiento.

La siguiente tabla consolida la información obtenida del entrenamiento con respecto al balanceo de datos. Se aplicó una estrategia combinada de balanceo a los tres conjuntos de datos. Esta estrategia consistió en aplicar submuestreo (under-sampling) para obtener el 35% del conjunto de datos que se refiere a la frecuencia mayoritaria, es decir, a las situaciones en las que 'no se requiere policía' (valor 0), y para le sobremuestreo (over-sampling) del 75% a las frecuencias en las que sí se requiere la presencia policial. Además, se implementó la estrategia WERCS tanto en su versión original sin ruido gaussiano como en una variante que incorpora ruido gaussiano (WERCS-GN).

Datos		Random Forest Regression			SVR			MLP Regression		
Ventana en horas	Tipo Balance	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE	R^2	R^2 Ajust	MAE
6	Baseline	0.874	0.874	0.032	-0.004	-0.005	0.142	-0.412	-0.413	0.178
	WERCS	0.992	0.992	0.041	0.151	0.150	1.171	0.545	0.545	0.808
	WERCS-GN	0.992	0.992	0.041	0.150	0.150	1.171	0.606	0.606	0.814
12	Baseline	0.849	0.849	0.033	0.036	0.035	0.138	-0.148	-0.150	0.083
	WERCS	0.992	0.992	0.037	0.291	0.290	1.046	0.541	0.540	0.830
	WERCS-GN	0.992	0.992	0.037	0.291	0.290	1.046	0.594	0.593	0.742
18	Baseline	0.848	0.848	0.033	0.060	0.058	0.135	-0.165	-0.168	0.127
	WERCS	**0.994**	**0.994**	**0.036**	0.371	0.370	0.968	0.650	0.649	0.687
	WERCS-GN	0.994	0.994	0.036	0.371	0.370	0.968	0.574	0.573	0.696

TABLA 6. RESULTADOS DEL ENTRENAMIENTO CON LOS CONJUNTOS DE DATOS BALANCEADO USANDO LA ESTRATEGIA WERCS

En la Tabla 6. Resultados del entrenamiento con los conjuntos de datos balanceado usando la estrategia **WERCS**, se presentan los resultados de los entrenamientos de los modelos para cada conjunto de datos balanceado. Es notable la mejora en el rendimiento en cada prueba, en comparación con los resultados obtenidos en el entrenamiento de línea base.

Para el modelo de Regresión de Vectores de Soporte (SVR), se observó una mejora significativa en el ajuste de las predicciones, alcanzando un R^2 del 37% y un MAE de 0.968. Por otro lado, el modelo de Regresión de Perceptrón Multicapa (MLP) logró un rendimiento bastante alto, con un R^2 de hasta el 65% y un MAE de 0.687. Este rendimiento se observó en el conjunto de datos sin ruido gaussiano y con una ventana de tiempo de 18 horas. Estos resultados destacan la eficacia de las técnicas de balanceo de datos en la mejora del rendimiento de los modelos de regresión.
En pruebas anteriores sin balanceo de datos, observamos que el modelo de Random Forest Regression presento un rendimiento considerablemente bueno, pero con la limitación de no tener en cuenta las frecuencias minoritarias debido a su baja representación dentro del conjunto de datos. Sin embargo, al realizar el balanceo, el rendimiento de este modelo mejoró aún más, ajustándose a cualquier conjunto de datos

Por lo tanto, podemos concluir que el modelo de Random Forest Regression superó a los demás modelos en todos los casos, con un coeficiente de determinación (R^2) que oscila entre el 99.2% y el 99.4%, y un error absoluto medio (MAE) entre 0.036 y 0.041, una desviación realmente muy pequeña. El conjunto de datos que mostró el mayor rendimiento fue aquel con una ventana de tiempo de 18 horas, con o sin

ruido gaussiano. Esto demuestra la eficacia del modelo de Random Forest Regression para realizar la tarea de manera eficiente. De igual forma, el número de nodos resultante después del entrenamiento del modelo se establece en 13967 nodos.

Para concluir, el modelo de Regresión de Bosques Aleatorios (Random Forest Regression) ha demostrado tener el mejor rendimiento para cumplir con los objetivos planteados. En etapas posteriores, se realizará un despliegue sencillo para el desarrollo de la herramienta de predicción. Esta herramienta será de gran utilidad para tomar decisiones basadas en datos y mejorar la eficiencia de las operaciones

5.3. Objetivo específico 3.

Para lograr los objetivos propuestos, se diseñó y desarrolló una aplicación innovadora entorno al modelo de aprendizaje automático especificado en el objetivo 2. Esta herramienta incorpora diversas técnicas y actividades de modelado para procesar y analizar los datos eficientemente. La interfaz de la aplicación está diseñada para ofrecer una visualización clara y detallada de los resultados, permitiendo a los usuarios observar las predicciones generadas por el modelo de Machine Learning. Así, se facilita la interpretación de los datos de las predicciones relacionadas con estaciones y terminales del sistema de transporte SITM-MIO, que se presentan sobre un mapa interactivo de Santiago de Cali, proporcionando una perspectiva organizada y accesible para los usuarios finales. Esta implementación no solo cumple con los requisitos establecidos, sino que también mejora la experiencia del usuario al interactuar con la información procesada.

La herramienta desarrollada es una aplicación web estructurada en dos componentes principales: el back-end y el front-end. El back-end es el núcleo de la aplicación, encargado de interactuar con la base de datos o fuente de datos. Esta sección realiza operaciones críticas como la lectura, transformación y limpieza de los datos, asegurando que la información sea precisa y esté lista para su análisis. Además, gestiona la ejecución del modelo de Machine Learning, modelo random forest regression, procesando los datos y generando predicciones valiosas. Por otro lado, el front-end actúa como la interfaz de usuario, donde se visualizan los resultados de las predicciones. Esta parte de la aplicación es esencial para presentar la información de manera comprensible y accesible, permitiendo a los usuarios interactuar con los datos y obtener insights significativos.

En este objetivo, nos centraremos en profundizar en la descripción y funcionalidades del back-end. Exploraremos cómo se llevan a cabo las tareas de procesamiento de datos y la lógica detrás de la ejecución del modelo. El front-end, que sirve como un dashboard interactivo, será abordado con mayor detalle en el objetivo 4, donde discutiremos su diseño, la experiencia del usuario y cómo facilita la interpretación de los resultados del modelo de Machine Learning. Nuestra meta es proporcionar

una comprensión clara y detallada de cada componente, destacando su importancia y cómo se complementan para ofrecer una solución robusta y eficiente.

5.3.1. Arquitectura básica de la aplicación web

En la Ilustración 10. Arquitectura básica de la aplicación web muestra claramente los dos componentes físicos fundamentales que constituyen la aplicación: el centro de datos y el equipo de cómputo. El equipo de cómputo es una entidad flexible que puede ser una laptop, una máquina virtual o un servidor operando en sistemas operativos Linux o Windows. Este elemento es crucial para la ejecución de la aplicación, ya que se encarga del procesamiento y análisis de los datos.

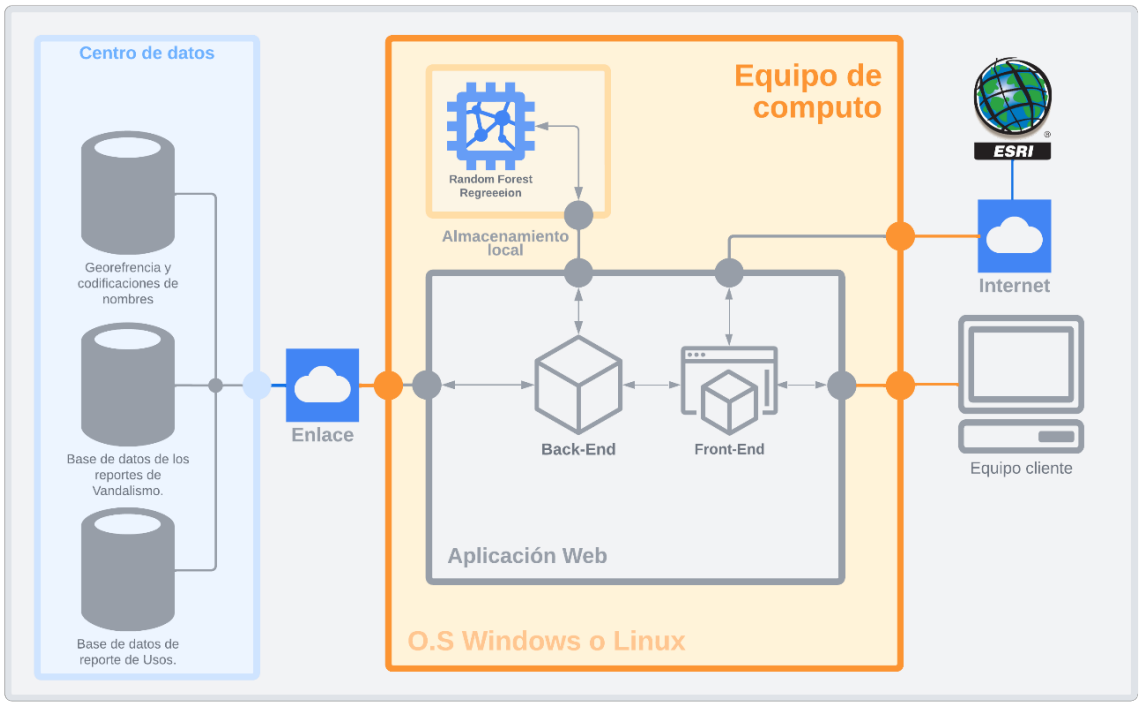

ILUSTRACIÓN 10. ARQUITECTURA BÁSICA DE LA APLICACIÓN WEB

El centro de datos, por su parte, es el repositorio central donde se almacenan todos los datos o insumos necesarios para alimentar el modelo de Machine Learning y realizar las predicciones. En el contexto de nuestra aplicación y como proyecto futuro, la integrar la aplicación web con el centro de datos de la UTR&T.

Para verificar la funcionalidad de la aplicación, se llevó a cabo una simulación del centro de datos, con conjunto de datos en forma CSV con la misma estructura como se encuentra almacenada en los datacenter de la UTR&T. Esta prueba permitió evaluar la capacidad de la aplicación para interactuar con el centro de datos simulado, procesar la información y generar predicciones precisas. Este enfoque garantiza que la aplicación sea capaz de funcionar correctamente en un entorno

controlado antes de su implementación en un escenario real, asegurando así su fiabilidad y eficacia.

En el contexto de nuestra aplicación, otro componente esencial es el servicio de información de mapas (ArcGIS), proporcionado por la compañía ESRI (Instituto de Investigación de Sistemas Ambientales). ArcGIS es una plataforma avanzada que facilita la creación, gestión y visualización de información geográfica, lo que resulta vital para la representación de datos en nuestra aplicación web.

Finalmente, tenemos el equipo cliente, que es el dispositivo a través del cual los usuarios finales accederán y consumirán los servicios ofrecidos. Este equipo cliente puede ser una computadora personal, una tableta o un teléfono inteligente, y es el punto de interacción con el dashboard de la aplicación web. A través de este, los usuarios pueden navegar y ver las predicciones del modelo de Machine Learning.

5.3.2. Back-end de la aplicación web.

El sistema está concebido para simular una conexión auténtica con las bases de datos de la UTRYT. Es crucial destacar que estas bases de datos no han sido sometidas a ningún tipo de preprocesamiento ni limpieza, y tampoco cuentan con datos normalizados. Por consiguiente, es importante implementar metodologías específicas dentro del back-end para realizar la obtención y extracción precisa de la información correspondiente a una fecha específica o día a predecir, facilitando así el análisis detallado de un día en particular.

El desarrollo del back-end de la aplicación se estructura en tres etapas clave que garantizan el procesamiento eficiente y la gestión de los datos:

1. **Etapa de Gestión de Datos:** Esta etapa es fundamental para consolidar y organizar toda la información proveniente de diversas fuentes de datos. Entre estas fuentes, tenemos bases de datos de vandalismo, uso en el sistema de transporte y una base de datos especial que incluye la recodificación de los nombres de las estaciones y terminales y la georreferenciación de cada una. Esta consolidación es crucial para asegurar que la información esté completa y lista para las siguientes fases del proceso.

2. **Etapa Preprocesamiento**: Una vez consolidada la información, esta etapa se encarga de realizar limpieza de datos, eliminación de cualquier inconsistencia o error que pueda afectar la calidad del análisis, reconstrucción de la serie de tiempo y unión de los conjuntos de todos los datos para consolida la información. Además, se realiza la codificación de la serie de tiempo, lo que permite transformar los datos en un formato adecuado para el análisis y lectura para el modelo.

3. **Etapa de Gestión de Modelo de Machine Learning:** La última etapa gestiona la carga de la información de modelo de Machine Learning

previamente generado para que sea funcional al momento de su uso. Aquí, los datos limpios y codificados son organizados y preparados para su análisis. El modelo procesa estos datos y organiza los resultados en formato JSON, el cual es interpretada por el front-end para posteriormente realizar la visualización de resultados en el dashboard. Esta etapa es esencial para observar los resultados, permitiendo a los usuarios finales comprender y explorar las predicciones realizadas por el modelo de manera intuitiva y accesible.

Cada una de estas etapas es vital para el funcionamiento eficaz del back-end, asegurando que los datos sean manejados de manera óptima desde su origen hasta la presentación final en el dashboard de la aplicación web.

5.3.3. Stack de tecnologías back-end.

La aplicación web se creó mediante el uso del lenguaje de programación Python y el microframework Flask, conocido por su simplicidad y flexibilidad en el desarrollo web. Flask ofrece ventajas significativas, como una extensa documentación respaldada por una activa comunidad de desarrolladores, facilitando el aprendizaje y la resolución de problemas. La integración fluida con bibliotecas de ciencia de datos como Pandas, NumPy y Scikit-learn destaca su utilidad en proyectos orientados al análisis de datos. Además, Flask permite la incorporación de plantillas HTML, JavaScript y CSS, mejorando la interacción y presentación de la información entre el back-end y el front-end. En resumen, la elección de Flask ha posibilitado un desarrollo modular y eficiente, permitiendo la integración de diversas funcionalidades y herramientas para mejorar tanto el procesamiento de datos en el back-end como la interactividad en el front-end.

5.3.4. Descripción y diseño genera del front-end.

La interfaz gráfica (véase Ilustración 11. Mockup de la interfaz web) tiene un diseño sencillo y minimalista donde los panes están ubicado sobre el mapa, se compone por 3 paneles principales:

1. **Panel de selección del día y puesta en marcha el análisis:** En este panel se encuentra ubicado en la parte super izquierda y está compuesto por un selector de fecha llamado “**Fecha a predecir**” y un botón de nombre “**Analizar**” que pone en marcha el análisis en la fecha establecida.

2. **Mapa de representación de estaciones:** En un mapa que muestra la ubicación georreferenciada de cada una de las estaciones y terminales que integran el sistema MIO están representados por una marca. Las marcas tienen 2 tipo de representación:

- Cirulo de color verde que indica que no se requiere atención de policía. Con base a la predicción, en todas las horas del día se obtuvo un valor de 0, no se requiere policía.

- Cuadrado de color rojo indica que, dentro de la predicción de las horas del día, existe al menos una hora con requerimiento de policía.

Cada marca posibilita la visualización de un popup al hacer clic sobre ellas.

3. **Popup de visualización de predicción:** Cuando se ha realizado la predicción de cada uno de la estaciones y terminales, se realiza una actualización de las marcas de acuerdo con la predicción realizada sobre el mapa. El popup de visualización se despliega en la parte superior derecha cuando realizamos “click” sobre cada marca mostrando el nombre de la estación, el día de la predicción y las predicciones por hora.

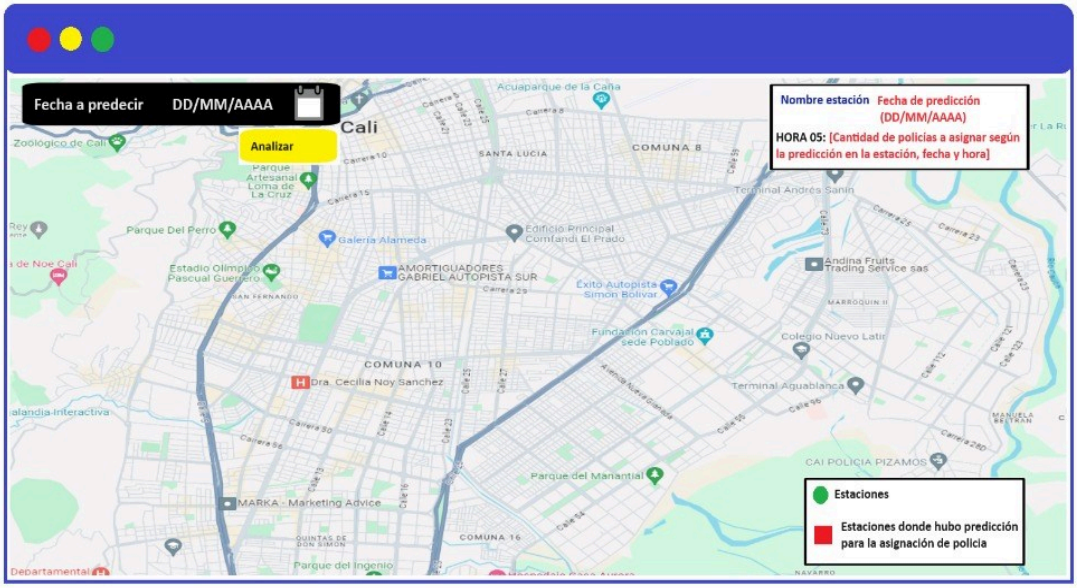

ILUSTRACIÓN 11. MOCKUP DE LA INTERFAZ WEB

5.4. Objetivo específico 4.

Para alcanzar el objetivo propuesto, en la Ilustración 11. Mockup de la interfaz web, se observa el diseñó final de la interfaz web (dashboard) que facilita la visualización de las predicciones realizadas por el modelo de Machine Learning. En la interfaz web se seleccionar una fecha del día a predecir y muestra la cantidad de policía por hora en un popup que se despliega en la parte superior derecha cuando se da “click” sobre cada una de las marchas mostradas en el mapa. La interfaz web incorpora una integración con ArcGIS (servicio de mapas), que permite la visualización de las predicciones en un mapa dinámico como usar distintas herramientas proporcionadas por ArcGIS.

La estructura de la interfaz gráfica está pensada para ofrecer una experiencia de usuario intuitiva y eficiente, asegurando que la información clave sea accesible con facilidad. Además, la integración con ArcGIS enriquece el análisis al dar una perspectiva geográfica de los datos, para la tomar decisiones de forma estratégica en la asignación de recursos policiales.

5.4.1. Diseños finales de la interfaz web.

El diseño final de la interfaz web, se componen básicamente en 3 partes:

1. Panel de selección de fecha y puesta en marcha de la predicción
2. La vista generar.
3. Popup de visualización de resultados.

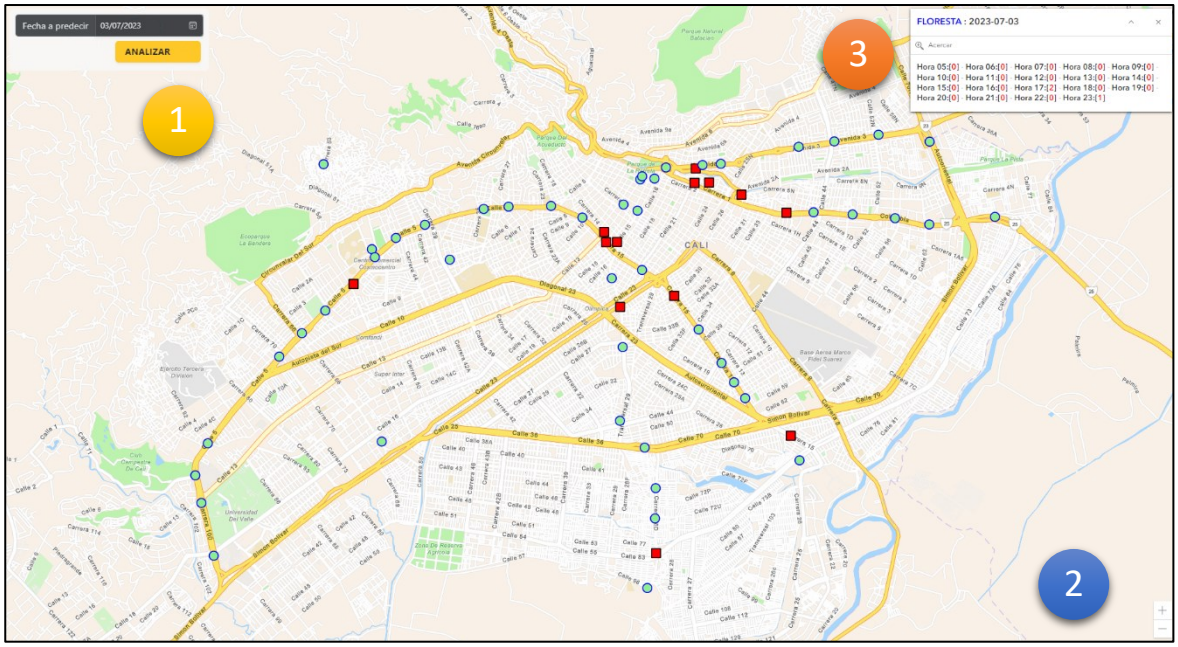

ILUSTRACIÓN 12. INTERFAZ GRÁFICA DE LA APLICACIÓN WEB

Para iniciar el uso de la herramienta, primero accedemos al panel de selección de fecha y puesta en marcha ubicado en el aparte superior izquierda. Dentro de este panel, elegimos una fecha, que no debe exceder el 31 de octubre de 2023, se utiliza una base de datos emulada. Esta selección nos permite simular una predicción como si fuera una situación real. Una vez seleccionada la fecha, procedemos a hacer clic en el botón ‘Analizar’. En ese instante, aparecerá un spinner de color rojo, señalando que la ejecución de las predicciones ha comenzado como se observa en la Ilustración 13. Panel de selección de fecha y puesta en marcha.

ILUSTRACIÓN 13. PANEL DE SELECCIÓN DE FECHA Y PUESTA EN MARCHA

Al concluir el análisis ejecutado por la aplicación web, en el mapa de ArcGIS se despliega unas marcas llamadas 'markerSymbol' sobre cada estación georreferenciada. Los símbolos circulares de color verde y los rectangulares de color rojos representan distintas condiciones del estado de la predicción. El 'marker Symbol' circular de color verde señala que la estación no requiere la intervención policial en ninguna hora del día. Por otro lado, el 'markerSymbol' rectángulo de color rojo indica que, en alguna hora específica, la estación sí necesita la presencia de un número de policías.

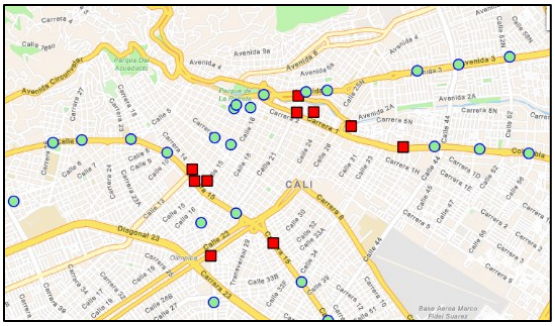

ILUSTRACIÓN 14. VISUALIZACIÓN DE 'MARKER SYMBOL' EN EL MAPA

Para visualizar los resultados de cada predicción, se debe seguir el siguiente procedimiento:

1. Hacer "click" sobre alguna marcar en el mapa (ILUSTRACIÓN 14. VISUALIZACIÓN DE 'MARKER SYMBOL' EN EL MAPA)

2. Al hace "click", se abrirá un popup en la esquina superior derecha (Ilustración 15. POPUP DE VISUALIZACIÓN DE RESULTADO) de la interfaz gráfica. En este popup, se mostrará el nombre de la estación y la fecha analizada.

3. En la parte inferior del popup, se destacarán en color rojo la cantidad de policía que se requiere en dicha hora donde el modelo de Machine Learning predijo que existe la necesidad de intervención policial para la estación seleccionada.

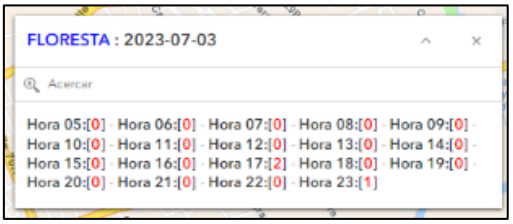

ILUSTRACIÓN 15. POPUP DE VISUALIZACIÓN DE RESULTADO

5.2.2. Stack de tecnologías de la interfaz gráfica.

La interfaz gráfica ha sido desarrollada utilizando JavaScript y HTML, aprovechando las capacidades del microframework Flask para la integración. La programación de las funcionalidades específicas presentadas en el mapa de ArcGIS se llevó a cabo con JavaScript, mientras que el maquetado de los objetos, como paneles y el propio mapa, se realizó en HTML. Esta estructura permite una integración efectiva con el servidor de ArcGIS, lo que facilita el uso de sus propiedades y características avanzadas para mapas.

El uso de JavaScript permite una interactividad dinámica y una respuesta ágil a las acciones del usuario, mientras que HTML proporciona la estructura y el diseño visual de la interfaz. La combinación de estas tecnologías, junto con la conexión al servidor de ArcGIS, resulta en una experiencia de usuario fluida y eficiente, donde se pueden manipular y visualizar datos geoespaciales con gran detalle y precisión".

6. Conclusiones y trabajos futuros

6.1. Conclusiones

- El objetivo general de desarrollar un modelo de aprendizaje autónomo y predictivo para optimizar la asignación de recursos de seguridad en el sistema de transporte masivo SITM MIO de Santiago de Cali se cumplió con éxito. Este modelo integró técnicas avanzadas de ciencia de datos y aprendizaje automático para analizar eficazmente los patrones de seguridad y predecir las necesidades de recursos en estaciones y terminales. La implementación de este sistema será valiosa, mejorando la seguridad y la eficiencia en la asignación de recursos policiales en las estaciones y terminales del SITM MIO, y sirviendo como un modelo replicable para otros sistemas de transporte masivo enfrentando desafíos similares. El éxito de la implementación de este proyecto representará un paso significativo hacia la mejora de la seguridad y la gestión eficiente de los sistemas de transporte masivo**.**

- El Objetivo Específico 1 del proyecto abordó con éxito la tarea crítica de preparar y manejar de manera óptima los datos para el análisis de actos de vandalismo en el sistema de transporte masivo SITM MIO de Santiago de Cali. Este objetivo se centró en la recolección, limpieza, y codificación de datos relevantes, estableciendo así una base sólida para análisis posteriores y la aplicación de modelos de aprendizaje automático.

  La recolección de datos fue un proceso meticuloso, donde se identificaron y obtuvieron conjuntos de datos clave de fuentes externas y internas, incluyendo información sobre actos de vandalismo, evasión de tarifas y uso de estaciones y terminales. La calidad de estos datos fue esencial para garantizar la precisión y relevancia de los modelos predictivos desarrollados.

  El proceso de limpieza y codificación de datos demostró ser fundamental. Se eliminaron inexactitudes y se estandarizaron los formatos, lo que permitió una representación coherente y precisa de la información. La normalización de los datos fue crucial, especialmente en la codificación de variables categóricas y la transformación de series temporales, lo que facilitó su posterior análisis y modelado.

  Este objetivo estableció un estándar alto en la gestión de datos dentro del ámbito de seguridad en el transporte masivo, demostrando que una preparación detallada y cuidadosa de los datos es esencial para cualquier análisis predictivo eficaz. Los esfuerzos realizados en esta etapa del proyecto permitieron comprender más las problemáticas de seguridad en el SITM MIO, y proporcionaron una base confiable para aplicar técnicas de aprendizaje en los objetivos siguientes.

- El Objetivo Específico 2 se centró en definir el modelo más eficaz para la herramienta de predicción. A través de un proceso meticuloso de entrenamiento y evaluación con diversas estructuras de datos, se optimizó un modelo de aprendizaje automático que responde específicamente a las características del conjunto de datos proporcionado. La implementación de metodologías variadas en los entrenamientos de línea base, aplicadas a modelos como Regresión de Bosques Aleatorios, Regresión de Vectores de Soporte y Regresión de Perceptrón Multicapa, fue esencial para comprender su rendimiento y capacidades, estableciendo una metodología clara para maximizar la eficiencia de los modelos.

  El empleo de técnicas avanzadas de balanceo de datos, incluyendo la Estrategia de Combinación Ponderada basada en Relevancia (WERCS), con y sin ruido gaussiano, resultó clave para mejorar el rendimiento de los modelos. Esto fue particularmente efectivo en el manejo de datos desbalanceados en problemas de regresión. Además, el uso de distintas ventanas temporales permitió una representación más equilibrada y homogénea de las frecuencias en varios escenarios, lo que contribuyó significativamente a la mejora del rendimiento predictivo.

  Entre todos los modelos evaluados, la Regresión de Bosques Aleatorios sobresalió por su excepcional rendimiento, reflejado en métricas de R^2 y MAE particularmente favorables, indicando un ajuste excelente a los datos. Este éxito fue decisivo para avanzar en el desarrollo de la herramienta de predicción, culminando en un modelo altamente ajustado al conjunto de datos.

- El Objetivo Específico 3 se cumplió con éxito, logrando el diseño y desarrollo de una herramienta de predicción innovadora basada en el modelo de aprendizaje automático seleccionado. La aplicación web resultante, que consta de un back-end y un front-end, es una solución integral eficiente tanto en el procesamiento y análisis de datos como en la presentación clara y accesible de los resultados para los usuarios finales.

  La arquitectura de la aplicación, que distingue claramente entre las funciones de procesamiento de datos y la interfaz de usuario, ha garantizado una gestión efectiva y una visualización intuitiva de las predicciones generadas por el modelo de Random Forest Regression. El back-end se encargó con éxito de la consolidación, limpieza y codificación de los datos, proporcionando una base sólida para el análisis predictivo. La integración con el servicio de información de mapas ArcGIS ha aportado un valor significativo, facilitando una representación geográfica precisa y detallada de

las predicciones de seguridad en un mapa interactivo de la ciudad de Santiago de Cali.

Por otro lado, el front-end ofrece una interfaz de usuario amigable y eficiente, que permite a los usuarios acceder y comprender fácilmente las predicciones y sus implicaciones para la seguridad en el transporte masivo. Esta interfaz es una herramienta valiosa para la toma de decisiones informadas y estratégicas por parte de los gestores del sistema de transporte, contribuyendo a una asignación de recursos de policía más eficiente y efectiva.

Esta conclusión se resaltaron los logros del Objetivo Específico 3, enfatizando la importancia de la herramienta desarrollada y su impacto en la gestión de la seguridad del transporte masivo.

- El Objetivo Específico 4, enfocado en el desarrollo y perfeccionamiento del front-end de la aplicación web, se cumplió exitosamente, culminando en la creación de un dashboard interactivo y amigable. Esta interfaz visual juega un papel crucial en la presentación de los resultados de predicción generados por el modelo de aprendizaje automático, facilitando la interpretación y el análisis de los datos para los usuarios finales.

  El dashboard desarrollado en este objetivo actúa como un enlace entre la complejidad técnica del back-end y la accesibilidad que requieren los usuarios no especializados en análisis de datos. Mediante una interfaz que destaca por su claridad, intuición y atractivo visual, el dashboard facilita a los usuarios del SITM MIO, incluyendo administradores y personal de seguridad, la visualización y comprensión rápida de áreas con riesgo potencial, patrones de criminalidad y otros indicadores de seguridad pertinentes. La integración efectiva de herramientas avanzadas de visualización de datos, junto con la capacidad de interactuar con el mapa de Santiago de Cali, enriquece significativamente el proceso de toma de decisiones basado en datos.

- El Objetivo Específico 2 se centró en definir el modelo más eficaz para la herramienta de predicción. A través de un proceso meticuloso de entrenamiento y evaluación con diversas estructuras de datos, se optimizó un modelo de aprendizaje automático que responde específicamente a las características del conjunto de datos proporcionado. La implementación de metodologías variadas en los entrenamientos de línea base, aplicadas a modelos como Regresión de Bosques Aleatorios, Regresión de Vectores de Soporte y Regresión de Perceptrón Multicapa, fue esencial para comprender su rendimiento y capacidades, estableciendo una metodología clara para maximizar la eficiencia de los modelos.

El empleo de técnicas avanzadas de balanceo de datos, incluyendo la Estrategia de Combinación Ponderada basada en Relevancia (WERCS), con y sin ruido gaussiano, resultó clave para mejorar el rendimiento de los modelos. Esto fue particularmente efectivo en el manejo de datos desbalanceados en problemas de regresión. Además, el uso de distintas ventanas temporales permitió una representación más equilibrada y homogénea de las frecuencias en varios escenarios, lo que contribuyó significativamente a la mejora del rendimiento predictivo.

Entre todos los modelos evaluados, la Regresión de Bosques Aleatorios sobresalió por su excepcional rendimiento, reflejado en métricas de R2 y MAE particularmente favorables, indicando un ajuste excelente a los datos. Este éxito fue decisivo para avanzar en el desarrollo de la herramienta de predicción, culminando en un modelo altamente ajustado al conjunto de datos.

- El Objetivo Específico 3 se cumplió con éxito, logrando el diseño y desarrollo de una herramienta de predicción innovadora basada en el modelo de aprendizaje automático seleccionado. La aplicación web resultante, que consta de un back-end y un front-end, es una solución integral eficiente tanto en el procesamiento y análisis de datos como en la presentación clara y accesible de los resultados para los usuarios finales.

  La arquitectura de la aplicación, que distingue claramente entre las funciones de procesamiento de datos y la interfaz de usuario, ha garantizado una gestión efectiva y una visualización intuitiva de las predicciones generadas por el modelo de Random Forest Regression. El back-end se encargó con éxito de la consolidación, limpieza y codificación de los datos, proporcionando una base sólida para el análisis predictivo. La integración con el servicio de información de mapas ArcGIS ha aportado un valor significativo, facilitando una representación geográfica precisa y detallada de las predicciones de seguridad en un mapa interactivo de la ciudad de Santiago de Cali.

  Por otro lado, el front-end ofrece una interfaz de usuario amigable y eficiente, que permite a los usuarios acceder y comprender fácilmente las predicciones y sus implicaciones para la seguridad en el transporte masivo. Esta interfaz es una herramienta valiosa para la toma de decisiones informadas y estratégicas por parte de los gestores del sistema de transporte, contribuyendo a una asignación de recursos de policía más eficiente y efectiva.

  Esta conclusión se resaltaron los logros del Objetivo Específico 3, enfatizando la importancia de la herramienta desarrollada y su impacto en la gestión de la seguridad del transporte masivo.

- El Objetivo Específico 4, enfocado en el desarrollo y perfeccionamiento del front-end de la aplicación web, se cumplió exitosamente, culminando en la creación de un dashboard interactivo y amigable. Esta interfaz visual juega un papel crucial en la presentación de los resultados de predicción generados por el modelo de aprendizaje automático, facilitando la interpretación y el análisis de los datos para los usuarios finales.

  El dashboard desarrollado en este objetivo actúa como un enlace entre la complejidad técnica del back-end y la accesibilidad que requieren los usuarios no especializados en análisis de datos. Mediante una interfaz que destaca por su claridad, intuición y atractivo visual, el dashboard facilita a los usuarios del SITM MIO, incluyendo administradores y personal de seguridad, la visualización y comprensión rápida de áreas con riesgo potencial, patrones de criminalidad y otros indicadores de seguridad pertinentes. La integración efectiva de herramientas avanzadas de visualización de datos, junto con la capacidad de interactuar con el mapa de Santiago de Cali, enriquece significativamente el proceso de toma de decisiones basado en dato.

6.2. Trabajos futuros

6.2.1. Predicción de las Entradas

- Explorar la inclusión de variables adicionales para mejorar la precisión de la predicción, como eventos públicos, condiciones climáticas, festividades locales, etc.
- Analizar la variabilidad de incidentes según días específicos de la semana y meses del año para ajustar el modelo a patrones estacionales.
- Considerar la inclusión de información demográfica y características específicas de las estaciones para capturar aspectos socioeconómicos que puedan influir en los incidentes.

6.2.2. Agregar Otras Salidas

- Incluir nuevas categorías de incidentes o eventos, como protestas, emergencias médicas, vandalismo, etc., para brindar una visión más completa de la seguridad en las estaciones.
- Evaluar la posibilidad de predecir el tiempo de respuesta de las autoridades ante incidentes, proporcionando una estimación del tiempo necesario para abordar situaciones específicas.
- Incorporar datos sobre la presencia de personal de seguridad privada en el modelo, considerando el contrato firmado, para prever la necesidad de refuerzos o redistribución de este recurso según la demanda y la ubicación.

6.2.3. Optimización del Modelo y Validación Continua

- Continuar refinando el modelo mediante técnicas de optimización, como la selección de características más relevantes y el ajuste de parámetros, para mejorar su rendimiento a medida que se recopilan más datos.
- Establecer un proceso de validación continua del modelo para garantizar su eficacia a lo largo del tiempo, considerando la evolución de patrones de seguridad y la posible adaptación del modelo a cambios en el entorno.

6.2.4. Interacción en Tiempo Real

- Desarrollar capacidades de interacción en tiempo real con el modelo, permitiendo a los responsables de seguridad tomar decisiones informadas y rápidas ante situaciones emergentes.
- Implementar alertas automáticas basadas en las predicciones del modelo para notificar sobre posibles incidentes inminentes, facilitando una respuesta proactiva.

Bibliografía

[1] Oficina de Análisis de Información y Estudios Estratégicos, Alcaldía Mayor de Bogotá D.C, «Evaluación del plan integral contra la evasión del pasaje en Transmilenio,» 18 Diciembre 2017. [En línea].

[2] IBM, «Metodología Fundamental para la Ciencia de Datos,» Junio 2015. [En línea]. Available: https://www.ibm.com/downloads/cas/6RZMKDN8. [Último acceso: 26 Noviembre 2023].

[3] T. MARCHANT MORALES, «MÉTODOS DE ANÁLISIS PARA BIG DATA Y SU PARTICIPACIÓN EN LA INDUSTRIA: ESTUDIO APLICADO A LA PREVENCIÓN DE FALLOS EN EMPRESAS FERROVIARIAS,» Universidad Técnica Federico Santa María, 2018. [En línea]. Available: https://repositorio.usm.cl/handle/11673/43410. [Último acceso: 26 Noviembre 2023].

[4] G. RÌos, «Series de Tiempo,» Universidad de Chile, 2008. [En línea]. Available: https://www.u-cursos.cl/ingenieria/2010/1/CC52A/1/material_docente/bajar?id_material=296003. [Último acceso: 2023 Noviembre 10].

[5] Z.-H. Zhou, Machine learning, China: Springer Nature Singapore Pte Ltd., ISBN 978-981-15-1966-6. [En línea]. Disponible en: https://link.springer.com/book/10.1007/978-981-15-1967-3., 2021.

[6] J. W. P. P. N. R. P. W. N. &. L. S. Biamonte, Machine Learning Algorithms - A Review, Quantum machine learning. Nature, 2017.

[7] D. W. T. H. a. R. T. G. James, An Introduction to Statistical Learning, New York, NY, USA: Springer, 2013.

[8] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY, USA: 1st ed., Springer Science+Business Media, LLC, New York, NY, 2006, ISBN 978-0387310732. Disponible en: https://link.springer.com/book/9780387310732, 2006.

[9] R. T. a. J. F. T. Hastie, «The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed,» Springer Series in Statistics, Springer, 2009. [En línea]. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7. [Último acceso: 20 11 2023].

[10] E. Alpaydin, Introduction to Machine Learning, Cambridge, MA, USA: The MIT Press, 2020.

[11] V. Vapnik, The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag, New York, NY, 1995, ISBN 978-1-4757-2440-0. Disponible en: https://link.springer.com/book/10.1007/978-1-4757-3264-1., 1995.

[12] J. S.-T. a. N. Cristianini, Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[13] B. S. a. A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.

[14] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[15] F. Rosenblatt, The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, 1958.

[16] S. Haykin, Neural Networks and Learning Machines, New York, NY, USA: Pearson, 2009.

[17] Y. B. a. G. H. Y. LeCun, Deep learning, Nature, vol. 521, no. 7553, pp. 436–444, May 2015. DOI: 10.1038/nature14539. Disponible en: https://www.nature.com/articles/nature14539., 2015.

[18] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks, vol. 61, pp. 85–117, Jan. 2015. DOI: 10.1016/j.neunet.2014.09.003. Disponible en: https://www.sciencedirect.com/science/article/pii/S0893608014002135, 2015.

[19] L. Breiman, Random Forests, 2001.

[20] R. T. a. J. F. T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York, NY, USA: 2nd ed., Springer Series in Statistics, Springer, New York, NY, USA, 2009, ISBN 978-0387848570. Disponible en: https://link.springer.com/book/10.1007/978-0-387-84858-7, 2009.

[21] D. E. a. L. W. P. Geurts, Extremely randomized trees, Machine Learning, vol. 63, no. 1, pp. 3-42, Apr. 2006. DOI: 10.1023/A:1010933404324, 2006.

[22] S. Turney, «Coefficient of Determination (R^2) | Calculation & Interpretation,» [En línea]. Available: https://www.scribbr.com/statistics/coefficient-of-determination/. [Último acceso: 20 Noviembre 2023].

[23] M. C. Johnson, Análisis Estadístico: Error Absoluto Medio y Sus Aplicaciones., Editorial Acme, 2020.

[24] L. T. y. R. P. R. P. Branco, «SMOGN: a Pre-processing Approach for Imbalanced Regression en Proceedings of the Machine Learning Research,» vol. 74, pp. 36-50, 2017. Disponible en: http://proceedings.mlr.press/v74/branco17a/branco17a.pdf, 2017. [En línea].

[25] P. Canuma, «How to Deal With Imbalanced Classification and Regression Data,» Neptune.ai, [En línea]. Available: https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data.

[26] J. Rahn, «Tackling Imbalanced Regression with SMOGN,» [En línea]. Available: https://jorahn.github.io.

[27] J. Rahn, «Tackling Imbalanced Regression with SMOGN,» [En línea]. Available: https://jorahn.github.io.

[28] J. Gado, «Tutorial02_resampling,» [En línea]. Available: https://github.com/jafetgado/resreg/blob/master/tutorial/tutorial02_resampling.ipynb.

[29] A. M. Cuadrado, «Utilización del Machine Learning en la Industria 4.0,» Universidad de Valladolid. Escuela de Ingenierías Industriales, Septiembre 2019. [En línea]. Available: https://uvadoc.uva.es/handle/10324/37908. [Último acceso: Noviembre 2023].

[30] J. S. González-Llanos, «Estado del arte de la inteligencia artificial en el sector ferroviario poniendo en común fabricante operadora e infraestructura,» Junio 2023. [En línea].

[31] Ministerio de Transporte de Colombia, «Informe de la Ministra al Congreso 2020 - 2021,» 2021. [En línea].

[32] A. B. R. J. A. B. C. A. R. Z. D. y. C. U. F. M. J. A. Ascencio Laguna, «Big Data e Internet de las Cosas para los sistemas inteligentes del transporte. Características y áreas de oportunidad,» 2020. [En línea].

[33] Ministerio de Transporte, «Sistema de Información, Evaluación y Seguimiento al Transporte Urbano,» [En línea]. Available: https://sisetu.mintransporte.gov.co.. [Último acceso: 26 Noviembre 2023].

[34] «Pandas - Python Data Analysis Library,» [En línea]. Available: https://pandas.pydata.org/. [Último acceso: 12 2023].

[35] U. d. Alcalá, «Todo lo que necesitas saber sobre Data Science,» [En línea]. Available: https://www.master-data-scientist.com/que-necesitas-saber-sobre-data-science/.

[36] green4T, «La tecnología y los retos del transporte público en 2023,» 17 Marzo 2023. [En línea]. Available: https://www.green4t.com/es/insights/la-tecnologia-y-los-retos-del-transporte-publico-en-2023/.