Modelo de Predicción de Tiempo de Espera de un Asegurado en Accidente de Tránsito

Diego Fernando Londoño Londoño
Yenny Patricia Vergara Monsalve

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Aníbal José Guerra Soler
Ph.D.in Electronic and Computer Engineering

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2025

Cita	(Londoño Londoño , D. F& Vergara Monsalve, Y.P 2025)
Referencia **Estilo APA 7 (2020)**	Londoño Londoño , D. F& Vergara Monsalve, Y.P (2025). *Modelo de predicción de tiempo de espera de un asegurado en accidente de tránsito* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte IX.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Nos dedicamos este logro como recordatorio de que los sueños se alcanzan con perseverancia, paciencia y fe.

Agradecimientos

A la Universidad de Antioquia, por brindarnos los conocimientos, recursos y espacios que hicieron posible nuestro crecimiento académico y profesional. A todos los docentes que aportaron a nuestra formación y a todas las personas que contribuyeron a este logro.

.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

MAE	Error Absoluto de Medición (*Mean Absolute Error*)
RMSE	Error Cuadrático Medio (*Root Mean Squared Error*)
ML	Machine Learning
.xlsx	Extensión archivos creados en Microsoft Excel
IQR	Rango Intercuartílico (*Interquartile Range*)
EDA	Análisis Exploratorio de Datos *(Exploratory Data Analysis)*
float.	Tipo de dato Flotante
int.	Tipo de dato entero
dataset	Conjunto de datos
Kb	Kilobyte, unidad de almacenamiento de datos
EMS	Servicios Médicos de Emergencia
ETA	Tiempo Estimado de Llegada (Estimated Time of Arrival)
UdeA	Universidad de Antioquia

Modelo de predicción de tiempo de espera de un asegurado en incidencias de tránsito

Resumen. **El proyecto tiene como objetivo desarrollar un modelo predictivo que estime el tiempo de espera de un asegurado después de reportar un accidente de tránsito, así poder optimizar el proceso de atención y mejorar la experiencia del asegurado. La iniciativa surge por la necesidad de contar con herramientas analíticas que permitan anticipar demoras en la atención y poder gestionar de una manera eficientemente los recursos disponibles como los gestores de accidentes. Se utilizó un conjunto de datos históricos de accidentes registrados por la empresa ASSINET a través de su sistema de gestión entre el mes de abril del 2016 al 11 de mayo del 2025. Estos datos incluyen variables relacionadas con la ubicación, hora del accidente, tipo de accidente, tráfico, disponibilidad de agentes, clima, entre otros. Para la implementación se realizó limpieza de datos, análisis exploratorio, selección de variables, ingeniería de datos, incluyendo variables necesarias para la predicción como el clima y la distancia, se realizó el entrenamiento de distintos modelos de Machine Learning de regresión lineal y no lineal como XGBRegresor, Random Forest y Gradient Boosting Regressor, Ride Regression, evaluando su desempeño mediante las métricas (MAE, RMSE y R^2). Entre los principales obstáculos se encontraron problemas en la calidad de los datos, con un alto porcentaje de presencia de outliers en la variable objetivo, variables faltantes, y desequilibrios en la distribución de los tiempos de atención. Los resultados obtenidos durante la fase de selección del modelo demostraron que el modelo de regresión lineal Ridge Regression es óptimo, superando significativamente a Gradient Boosting, Random Forest y XGBoost con un R^2 de 0.999865.**

1. Introducción

1.1 Contexto y relevancia del problema

En Medellín, los accidentes viales han alcanzado niveles preocupantes en los últimos años. Por ejemplo, en 2024 la ciudad registró 308 muertes por accidentes viales, superando por primera vez en una década la barrera de los 300 fallecidos (Tobón, 2025). Este incremento no solo indica mayor severidad de los accidentes, sino que se atribuye en parte al crecimiento del uso de motocicletas, lo que también aumenta la demanda sobre los servicios de atención de emergencias y seguros.

En entornos urbanos como Medellín, la congestión vehicular agrava la situación, los tiempos de respuesta de emergencias médicas o aseguradoras se prolongan significativamente cuando hay alta congestión vial (Zuñiga, 2020). Por ello, anticipar y reducir el tiempo de espera de un asegurado tras un accidente es clave para salvar vidas, optimizar recursos e incrementar la satisfacción del cliente.

La gestión en tiempos de llegada o de espera es una problemática común en sectores como la salud, logística, servicios de domicilios, servicios de transporte como UBER y DiDi. Particularmente en casos de accidentes de tránsito, el tiempo que transcurre entre el accidente y la atención del asegurado representa un indicador crítico en calidad del servicio.

La ciencia de datos y el análisis predictivo son herramientas usadas en el sector de seguros moderno (carmatec, 2025). Usualmente, las aseguradoras utilizan técnicas de analítica para mejorar la experiencia del cliente, prevenir fraudes y agilizar la gestión de accidentes. En particular, la modelación predictiva permite predecir resultados futuros a partir de datos históricos. En el proceso de atención de accidentes, el análisis de los datos agiliza la gestión de trámite de reclamaciones, automatizando y priorizando los casos. En el contexto de nuestro trabajo, estas capacidades permiten anticipar demoras en la atención de accidentes de tránsito y asignar recursos como los gestores de accidentes de manera más eficiente.

Actualmente, la gestión de accidentes de tránsito enfrenta varios retos. Por un lado, las condiciones urbanas como alta motorización, la infraestructura vial limitada genera congestión, extendiendo los tiempos de respuesta de emergencias sobre los niveles recomendados, menos de 8 minutos según estándares internacionales (Calatayud, 2020). En ciudades latinoamericanas se ha observado que la congestión puede agregar hasta 30 minutos extra a estos tiempos (Calatayud, 2020), lo que genera insatisfacción y costos adicionales. Por otra parte, el uso de modelos analíticos demanda datos de calidad. En la práctica se suelen encontrar problemas de datos faltantes, presencia de outliers y distribuciones desequilibradas que dificultan la modelación. Por ello, resulta necesario un cuidadoso preprocesamiento como la limpieza e imputación de datos (carmatec, 2025)) y una ingeniería de variables

adecuada, por ejemplo, incorporar indicadores de distancia al lugar del accidente, tráfico y clima para construir modelos robustos.

1.2 Planteamiento del problema

En el contexto urbano de Medellín, el aumento constante de movilidad vehicular y el crecimiento de motocicletas han generado una mayor frecuencia de accidentes de tránsito, afectando directamente los tiempos de respuesta de las aseguradoras. Cuando un asegurado reporta una incidencia, el tiempo de espera hasta recibir atención, ya sea del gestor del seguro o del personal en el lugar del accidente, se convierte en un factor crítico que impacta la percepción del servicio, la eficiencia operativa y, en algunos casos, la seguridad del afectado.
Sin embargo, la gestión de dichos tiempos aún depende en gran medida de estimaciones empíricas de la disponibilidad inmediata del recurso humano, lo que limita la capacidad de respuesta de las compañías aseguradoras.

La problemática específica que se pretende abordar en esta investigación se centra en la falta de un modelo analítico que permita predecir el tiempo de espera de un asegurado tras un accidente de tránsito. En la práctica, las aseguradoras no suelen contar con sistemas predictivos que integren variables contextuales como la ubicación del accidente, las condiciones climáticas, la congestión vehicular o la disponibilidad de gestores.
Esta ausencia de herramientas de predicción dificulta anticipar demoras y redistribuir recursos de forma óptima, ocasionando ineficiencias operativas y un mal servicio y atención del cliente.

La motivación detrás del presente trabajo radica en la necesidad de optimizar los procesos de atención mediante el uso de técnicas de ciencia de datos y aprendizaje automático. La posibilidad de estimar el tiempo de llegada del gestor permite planificar mejor la asignación de agentes, reducir tiempos de espera y, en consecuencia, mejorar la satisfacción del asegurado. Además, al incorporar datos históricos de accidentes registrados entre 2016-2025 por la empresa ASSINET, se busca aprovechar la información disponible para generar un modelo predictivo robusto que pueda implementarse en sistemas reales de atención de accidentes.

Aunque existen numerosos trabajos internacionales que predicen tiempos de respuesta en servicios de emergencia, la mayoría se concentra en contextos hospitalarios o públicos de transporte. Por ejemplo, (Taylor, 2017) utiliza un modelo de supervivencia espacial para analizar tiempos de llegada de servicios de emergencia con variables espaciales y temporales, pero no aborda el ámbito asegurador. Pocos trabajos han abordado la modelación del tiempo de atención de accidentes considerando simultáneamente variables espaciales, temporales y contextuales.

De esta manera, la presente investigación propone un enfoque analítico integral que combina técnicas de machine learning y analítica predictiva adaptadas a la realidad operativa de Medellín para predicción de tiempos de llegada.

1.3 Objetivo General

Predecir el tiempo de espera en la atención de un asegurado durante una incidencia de tránsito, desarrollando un modelo de Machine Learning con el fin de brindar soporte al proceso del gestor de seguros.

Objetivos Específicos

1. Diseñar un análisis exploratorio y de limpieza de datos provenientes de la data histórica de los accidentes para identificar patrones, inconsistencias y variables relevantes.
2. Construir variables derivadas (ingeniería de características) que integren factores espaciales, temporales y contextuales como clima, tráfico, distancia, entre otros.
3. Entrenar y comparar modelos de aprendizaje automático para estimar el tiempo de espera, evaluando su desempeño con métricas como *MAE*, *RMSE* y R^2.
4. Evaluar y justificar la selección del modelo con mejor desempeño y validar su capacidad predictiva para su futura integración en los sistemas de gestión de accidentes de la aseguradora.
5. Formular un conjunto de recomendaciones que permitan mejorar el rendimiento del modelo

seleccionado en etapas posteriores de la investigación.

1.4 Metodología

La metodología empleada se fundamenta en un enfoque cuantitativo y analítico, basado en el uso de técnicas de ciencia de datos y aprendizaje automático para la predicción de tiempo de espera en accidentes de tránsito. El proceso se desarrolló en varias etapas, que abarcaron desde la exploración inicial del dataset y preparación de los datos hasta la evaluación comparativa de los modelos.

El análisis se realizó con registros históricos de accidentes viales proporcionados por la empresa **ASSISNET** (abril 2016 a mayo 2025). El dataset denominado *DataSet_aseguradora.xlsx* contiene 14520 registros y 28 variables y se encuentra en formato de Excel con extensión. xlsx, tiene un tamaño de 3.330 kb Para este trabajo se filtró la información correspondiente a la ciudad de Medellín, obteniendo 8967 registros. Esta decisión buscó reducir la presencia de outliers y datos atípicos asociados a tiempos de desplazamiento atípicos en otros municipios y mejorar la coherencia geoespacial del análisis.

Los datos fueron obtenidos mediante una consulta Transact SQL ejecutada sobre un servidor SQL Server, que almacena información de las incidencias de tránsito y de las gestiones de atención realizadas por los agentes de las aseguradoras suscritas a la empresa ASSISNET. En el análisis se incluyen variables relacionadas con la ubicación del accidente, tipo de accidente, condiciones de tráfico y climáticas, hora del día, disponibilidad de gestores de accidentes, entre otras.

El proceso inició con una limpieza de datos que incluyó manejo de valores faltantes, detección de valores erróneos y tratamiento de outliers. Posteriormente, se realizó un análisis exploratorio para comprender la distribución de las variables y su relación con el tiempo de espera de un asegurado. Adicionalmente, se desarrolló ingeniería de variables para generar variables relevantes como, indicadores climáticos y medidas de distancia entre el accidente y el gestor asignado, con el fin de mejorar el poder predictivo de los modelos.

Para el modelado predictivo se emplearon modelos de aprendizaje supervisado que incluyeron métodos lineales (Ridge, Lasso, LinearRegression) y modelos no lineales basados en árboles, permitiendo una comparación con diferentes enfoques. La optimización de hiperparámetros se llevó a cabo mediante técnicas como *Grid Search* y búsquedas aleatorias con *Rondomized Search*, evaluando parámetros como profundidad del árbol, número de estimadores y taza de aprendizaje. Adicionalmente, se incluyeron otros modelos como *Ridge, Lasso, ElasticNet, LinearRegression, DecisionTreeRegressor, SVR, KNeighborsRegressor*, alcanzando un total de diez modelos evaluados, lo que permitió confirmar la selección óptima para nuestro problema.

La evaluación del desempeño se realizó empleando métricas estadísticas para regresión. Se calcularon el Error Absoluto Medio (*MAE*), la Raíz del Error Cuadrático Medio (*RMSE*) y el Coeficiente de Determinación (R^2). Estas métricas permiten cuantificar la precisión y el ajuste a los datos:

- Valores más pequeños de MAE y RMSE indican predicciones más cercanas a los tiempos de espera reales.
- Un R^2 más alto, refleja un mejor ajuste del modelo a los datos (Cosio, 2021). Los resultados obtenidos para cada métrica se compararon entre los diferentes modelos, seleccionando finalmente la técnica con mejor desempeño.

2 Materiales y Métodos

En este capítulo se presenta de manera detallada el proceso metodológico utilizado para construir y validar el modelo predictivo del tiempo de espera de un asegurado en accidentes de tránsito.

Para ello se integraron técnicas de ciencias de datos y modelos de aprendizaje automático aplicados sobre el conjunto de datos proporcionados por ASSISNET.

2.1 Descripción de los datos

La Tabla 1 presenta la estructura general del dataset utilizado, incluyendo sus variables principales, tipos de

datos y el número total de observaciones disponibles tras el proceso inicial de depuración

Columna	Descripción	Ejemplo
idproceso	Código para identificar el caso	10
placa	Placa del vehículo	BQX320
fecha	Fecha en que se reporta el accidente	2025-05-02 10:40:25
fechallegada	Fecha en que se llega al sitio donde ocurrió el accidente	2025-05-02 10:50:25
annoregistro	Año en que se reporta el accidente	2025
mesregistro	Mes en que se reporta el accidente	5
diaregistro	Día en que se reporta el accidente	2
horaregistro	Hora en que se reporta el accidente	10
diasemana	Día de la semana en que se reporta el accidente	Lunes
annoatencion	Año en que se llega al sitio donde ocurrió el accidente	2025
mesatencion	Mes en que se llega al sitio donde ocurrió el accidente	5
diaatencion	Día en que se llega al sitio donde ocurrió el accidente	2
Horaantencion	Hora en que se llega al sitio donde ocurrió el accidente	10
TiempoAtencion	Diferencia en minutos entre el reporte y la llegada	
HoraPicoTarde	¿Ocurrió en hora pico de la tarde (5 a 7 pm)? (Si/No)	Si
HoraPicoManana	¿Ocurrió en hora pico de la mañana (6:30 a 8:30 am)? (Si/No)	No
InicioNoche	¿Ocurrió entre las 7 pm y 11:59 pm? (Si/No)	No
Amanecer	¿Ocurrió entre las 12 am y 5:59 am? (Si/No)	Si
Municipio	Municipio donde ocurrió el accidente	No
Intsancia	Tipo de accidente reportado por la aseguradora	Preliminar con lesiones
clienteimportante	Si el asegurado es cliente importante	Banco de accidente
Acuerdo	Al acuerdo que se llega entre las partes implicadas	Tránsito
UsuarioRegistra	Usuario que registra el caso en el sistema	Carolina Garcia Valencia
LiberaVehiculo	Si el vehículo fue retenido y luego liberado	Liberado
ResultadoFallo	Si el fallo en audiencias es a favor o en contra	A favor
Aseguradora	Aseguradora que reporta el accidente	Liberty Seguros
Abogadounico	Abogado disponible para atender el caso	Carolina Garcia Valencia
RandomAbogado	Abogado que finalmente atiende el caso	Carolina Garcia Valencia
Descripción	Detalle de caso	SIMPLE PREVISORA LIVIANOS MEDELLIN 7516811 JUAN CARLOS PEREZ 3004931460 MTT750 CARRERA 46 CON CALLE 78 24 CAMPO VALDEZ ASIG 19:06 SERGIO BARRIENTOS 40 M

Figura 1.Identificación de la estructura y variables del conjunto de datos

2.1.1 Fuente de los datos

El trabajo se basó en un conjunto de datos históricos de accidentes viales registrados por la empresa ASSISNET, La información proviene de su sistema interno de gestión de accidentes correspondientes al período comprendido entre abril 2016 y mayo 2025 en la ciudad de Medellín. Se trata, por tanto, de una base de datos privada, construida a partir de reportes reales de accidentes de tránsito, sin intervención de encuesta ni fuentes externas.

El dataset fue obtenido mediante una consulta Transact-SQL ejecutada en un servidor SQL Server y posteriormente suministrado por ASSISNET en formato Excel con extensión .xlsx. La información incluye detalles sobre eventos de tránsito, las características contextuales del accidente y las condiciones operativas asociadas a la atención del servicio.

El conjunto de datos está compuesto por **8967** registros y **31** variables que incluyen variables numéricas, categóricas y de tipo texto. Estas variables abarcan información temporal (año, mes, día y hora del accidente), ubicación geográfica (latitud, longitud, dirección) condiciones ambientales (clima, lluvia), y datos operativos relevantes, como el tiempo de atención, identificador del proceso, tipo de accidente, y la asignación del gestor o profesional encargado.

2.1.2 Tamaño y estructura

El conjunto de datos presenta una estructura compuesta por variables categóricas, numéricas y tipo entero, lo que resulta adecuado para la construcción de modelos predictivos basados tanto en relaciones lineales como no lineales. Esta diversidad de tipos de datos permite capturar dinámicas temporales, espaciales y operativas asociadas al tiempo de atención adecuadas para un *modelo de regresión no lineal.*
Entre las principales características del dataset se destacan:

- TiempoAtencion: variable objetivo (*target*), expresada en segundos que representa el tiempo transcurrido entre el reporte del accidente y la llegada del gestor al lugar del accidente.
- HoraRegistro y HoraAtencion: variables temporales que permiten identificar patrones de demanda, congestión y variabilidad operativa según hora del día.
- Lluvia**:** variable categórica que describe las condiciones climáticas al momento del accidente, relevante para evaluar efectos climáticos sobre los tiempos de atención.
- Distancia: variable derivada que mide la distancia aproximada entre el gestor y el lugar del accidente, calculada a partir de las coordenadas geográficas del evento.
- DiaSemana y MesRegistro: variables temporales que permiten detectar patrones de estacionales y variaciones semanales en los tiempos de espera.

2.1.3 Problemas en los datos

Durante la fase inicial de análisis exploratorio se identificaron múltiples problemas de calidad en los datos principalmente relacionados con valores faltantes, alta cardinalidad y presencia de valores atípicos.

En primer lugar, se detectó la existencia de valores nulos en varias columnas. Las variables como clienteimportante, *LiberaVehiculo, ResultadoFallo, Acuerdo,* presentan porcentajes de ausencia entre 55 %-99.50%, por lo que fueron consideradas como candidatas a eliminación debido a su baja completitud y poco aporte para el modelo.

Por otro lado, variables como *Abogadounico, Aseguradora y Lluvia* presentan alrededor del 30% de valores faltantes, lo que las convierte en buenas candidatas para procesos de imputación.

===== TABLA RESUMEN DE VALORES NULOS =====

	Valores Nulos	Porcentaje (%)
ResultadoFallo	7995	99.44%
LiberaVehiculo	7897	98.22%
clienteimportante	7674	95.45%
Acuerdo	4494	55.90%
Abogadounico	372	4.63%
Aseguradora	14	0.17%
Lluvia	7	0.09%

Tabla 1. Tabla resumen de valores nulos

Asimismo, el análisis de cardinalidad permitió identificar el número de valores únicos por columna, estableciendo un umbral de más de 50 valores únicos para clasificar una variable con alta cardinalidad. Este análisis es clave para definir qué variables deben descartarse, agruparse o transformarse, con el fin de reducir el ruido e incrementar la eficiencia del modelado.

Los resultados evidencian una clara diferenciación entre variables identificadoras, descriptivas, categóricas y numéricas, lo cual facilita la comprensión de la estructura del dataset y la diversidad de información disponible en accidentes

En el análisis podemos identificar que las variables numéricas continuas, *TiempoAtencion* y *distancia* presenta una proporción moderada de valores únicos alrededor del 23%, lo que confirma su relevancia para el análisis predictivo. En particular el *TiempoAtencion* corresponde a la variable objetivo del trabajo, mientras que la *distancia* se interpreta como un predictor asociado a la eficiencia operativa del servicio.

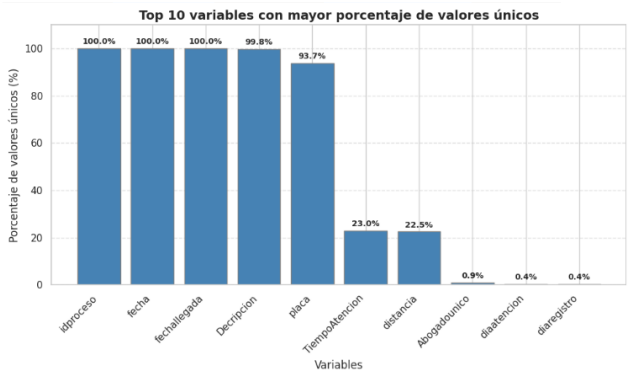

Figura 2. Variables con mayor porcentaje de valores únicos

Finalmente, se identificó la presencia de valores atípicos (*outliers*) significativos en la variable objetivo TiempoAtencion, con valores que alcanzan una variabilidad alta con un 17.8% de valores extremos.

Esto nos indica posibles errores de captura o casos excepcionalmente anómalos. Además, algunas variables muestran sesgos marcados en su distribución, especialmente en aquellas asociadas a condiciones climáticas, donde predominan los registros sin lluvia, y variables temporales relacionadas con horarios laborales. Tales características pueden influir negativamente en el desempeño de los modelos supervisados**.**

En la siguiente tabla se puede visualizar las variables con mayor porcentaje de *outliers*.

===== DETECCIÓN DE OUTLIERS (Método IQR) =====

Variable	Q1	Q3	IQR	Límite inferior	Límite superior	Cantidad outliers	% del total
TiempoAtencion	12.000000	225.250000	213.250000	-307.880000	545.120000	1590	19.780000
Abogadounico_freq	0.030000	0.070000	0.050000	-0.050000	0.140000	1534	19.080000
distancia	8482.210000	8486.800000	4.590000	8475.330000	8493.680000	465	5.780000
horaregistro	11.000000	18.000000	7.000000	0.500000	28.500000	66	0.820000
diaregistro	8.000000	24.000000	16.000000	-16.000000	48.000000	0	0.000000
annoregistro	2018.000000	2022.000000	4.000000	2012.000000	2028.000000	0	0.000000
mesregistro	4.000000	10.000000	6.000000	-5.000000	19.000000	0	0.000000
diaatencion	8.000000	23.000000	15.000000	-14.500000	45.500000	0	0.000000
mesatencion	4.000000	10.000000	6.000000	-5.000000	19.000000	0	0.000000
annoatencion	2018.000000	2022.000000	4.000000	2012.000000	2028.000000	0	0.000000
Horaantencion	11.000000	19.000000	8.000000	-1.000000	31.000000	0	0.000000

Tabla 2. Identificación de outliers con el método IQR

2.1.4 Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos (EDA) tuvo como objetivo examinar la estructura del dataset, comprender la distribución de las variables y detectar relaciones relevantes para la predicción del tiempo de atención. Esta etapa permitió identificar patrones temporales, condiciones que influyen en los tiempos de respuesta y posibles inconsistencias en la base de datos, posiblemente errores humanos al ingresar la información.

El conjunto de datos está compuesto por **8.967 registros y 31 variables**, de las cuales 11 son numéricas y 20 son categóricas. Durante la inspección inicial se observó que varias variables presentan valores faltantes, principalmente aquellas relacionadas con el proceso de atención como *annoatencion*, *mesatencion*, *diaatencion*, *Horaatencion*, *TiempoAtencion*, así como la variable *Lluvia*, que registra 934 valores nulos al tratarse de una condición climática no siempre reportada.

En términos generales, el dataset presenta una estructura mixta adecuada para el modelado predictivo. Aunque algunas variables operativas tienen un nivel moderado de valores faltantes, la cantidad total de registros permite realizar un análisis estadístico y aplicar técnicas de aprendizaje supervisado.

Estadísticas descriptivas de la variable objetivo TiempoAtencion

La variable objetivo *TiempoAtencion* que representa el tiempo transcurrido entre el reporte del accidente y la llegada del gestor al lugar del accidente presenta una distribución altamente asimétrica hacia la derecha. Sus estadísticas descriptivas muestran una media de 2.976,2 segundos**,** una mediana de 28 segundos **y** una desviación estándar considerablemente alta (27.272,7)**.** El valor mínimo registrado es **0** segundos, mientras que el máximo asciende a 878.976 segundos aproximadamente 10 días, lo que evidencia la presencia de valores extremos.

La marcada diferencia entre la media y la mediana confirma que la mayoría de los casos registran tiempos de atención cortos, mientras que un pequeño número de valores atípicos eleva la media de forma pronunciada. Esto también se refleja en el rango intercuartílico (12 a 225 segundos)**,** que es muy reducido respecto a la magnitud de los valores máximos. En consecuencia, con la distribución de datos, se hace necesario aplicar técnicas de transformación especialmente logarítmica para estabilizar la varianza antes del modelado.

A continuación, se presenta el resumen general de las estadísticas descriptivas de las variables incluidas en el dataset.

```
===== Estadísticas Descriptivas: =====

          idproceso  annoregistro  mesregistro  diaregistro  horaregistro  \
count   8967.000000   8967.000000  8967.000000  8967.000000   8967.000000
mean   44100.845322   2020.194268     6.805063    15.930523     14.055425
std    14001.798155      2.342271     3.555137     8.838929      4.707902
min     1113.000000   2016.000000     1.000000     1.000000      0.000000
25%    39714.500000   2018.000000     4.000000     8.000000     11.000000
50%    47688.000000   2021.000000     7.000000    16.000000     14.000000
75%    54832.500000   2022.000000    10.000000    24.000000     18.000000
max    58954.000000   2025.000000    12.000000    31.000000     23.000000

       annoatencion  mesatencion  diaatencion  Horaantencion  TiempoAtencion  \
count   8040.000000  8040.00000  8040.000000    8040.000000      8040.00000
mean    2020.020647     6.80398    15.817040      14.583831      2976.15510
std        2.276523     3.52199     8.863927       5.107983     27272.73376
min     2016.000000     1.00000     1.000000       0.000000         0.00000
25%     2018.000000     4.00000     8.000000      11.000000        12.00000
50%     2020.000000     7.00000    16.000000      15.000000        28.00000
75%     2022.000000    10.00000    23.000000      19.000000       225.25000
max     2025.000000    12.00000    31.000000      23.000000    878976.00000

          distancia
count   8967.000000
mean    8484.481152
std      304.845399
min     1759.926915
25%     8482.191998
50%     8483.506851
75%     8486.799330
max    13230.957749
```

Tabla 3. Estadísticas descriptivas de las variables del dataset

2.1.5 Distribución del Tiempo de Atención

La distribución del *TiempoAtencion* muestra un patrón fuertemente asimétrico, con la mayoría de los registros concentrados entre 0 y 20.000 segundos equivalentes a 0–5,5 horas. Este comportamiento indica que, en términos generales, los accidentes suelen ser atendidos en lapsos relativamente cortos.

El siguiente histograma nos permite visualizar con mayor claridad la asimetría positiva lo que nos indica que la variable no sigue una distribución normal, y nos lleva a la necesidad de aplicar transformaciones o métodos robustos para reducir la influencia de los valores atípicos.

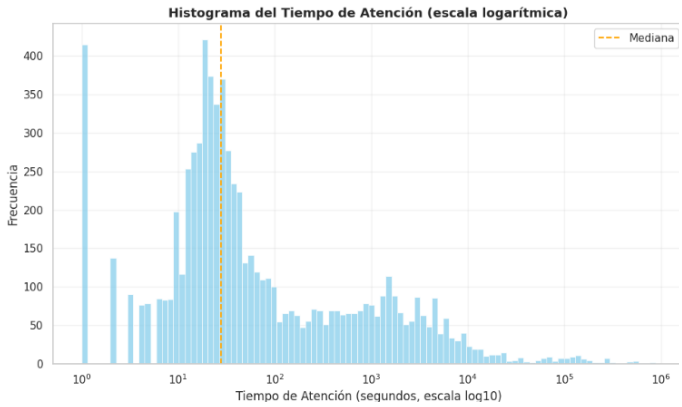

Figura 3. Histograma de la distribución de la variable TiempoAtencion en escala logarítmica

El boxplot evidencia que la mayoría de los tiempos se ubican muy cerca del eje de origen, lo que confirma la concentración de atenciones rápidas. Sin embargo, también muestra numerosos puntos fuera del IQR y valores que superan los 800.000 segundos, es decir, más de nueve días. Estos casos extremos podrían deberse a errores de captura, inconsistencias administrativas o eventos operativos excepcionales, y requieren tratamiento previo al modelado.

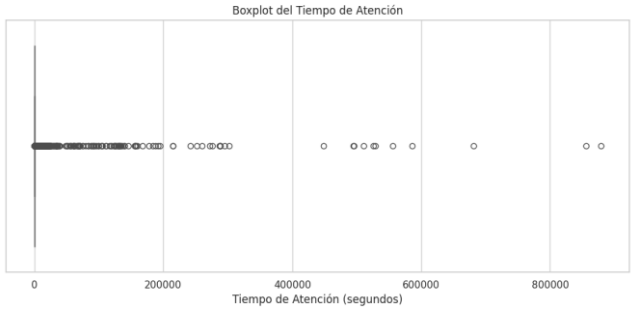

Figura 4. Boxplot del tiempo de atención en segundos para identificación de Outliers

2.1.6 Tiempo de Atención según condiciones climáticas

El análisis comparativo del tiempo de atención entre accidentes ocurridos con y sin lluvia muestra que, la mayoría de los registros se concentran alrededor de valores cercanos a cero, confirmando que las atenciones suelen ocurrir en lapsos cortos independientemente de las condiciones climáticas. No obstante, en ambos grupos aparecen *outliers* que superan ampliamente los 800.000 segundos.

Se observa una ligera mayor dispersión en los valores extremos bajo condiciones sin lluvia. Aunque la lluvia puede influir en los tiempos de atención, no parece modificar la tendencia general. Los retrasos más pronunciados probablemente responden a factores operativos como disponibilidad de gestores, tráfico o distancia.

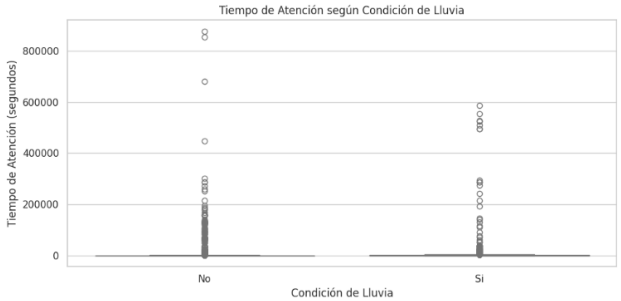

Figura 5. Boxplot del tiempo de atención según la condición de lluvia

2.1.7 Relaciones entre variables numéricas

El mapa de calor muestra las correlaciones lineales entre las variables numéricas del dataset, permitiendo identificar redundancias y relaciones útiles para el modelado.

Las variables temporales *idproceso*, *annoregistro*, *annoatencion*, *mesregistro* y *mesatencion* presentan correlaciones superiores a $r > 0{,}9$, lo que sugiere una dependencia temporal significativa, probablemente derivada del flujo operativo del sistema de registro de accidentes.

Existe una correlación moderada entre *horaregistro* y *horaatencion* ($r \approx 0{,}59$), lo cual es consistente con los tiempos operativos del proceso que reportes en determinados momentos del día tienden a ser atendidos en rangos similares.

El *TiempoAtencion* muestra correlaciones lineales muy bajas con las demás variables numéricas ($r < 0{,}05$), lo que indica que su comportamiento no depende de relaciones lineales simples, sino de combinaciones no lineales y variables categóricas como clima, día de la semana o disponibilidad de gestores. Esto justifica la incorporación de modelos no lineales como Gradient Boosting, Random Forest o XGBoost, Ridge y Lasso.

Finalmente, la variable *Distancia* no presenta correlaciones destacables con otras características numéricas , aunque podría tener efectos no lineales

sobre el tiempo de atención, especialmente en escenarios de movilidad variable.

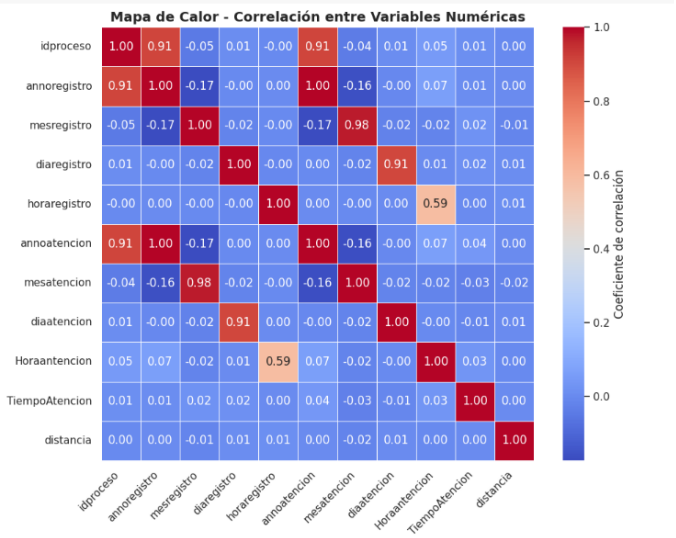

Figura 6. Mapa de calor para identificar la correlación entre variables numéricas

2.2 Preprocesamiento y limpieza de datos

El preprocesamiento de los datos constituye una fase fundamental en la construcción del modelo predictivo, ya que garantiza la calidad, y coherencia de la información utilizada. Esta etapa incluyó la depuración de registros, la imputación de valores faltantes, la detección y tratamiento de valores atípicos (*outliers*), la conversión de tipos de datos, la creación de variables derivadas y la preparación del conjunto de datos definitivo para el modelado.

La base de datos original, proveniente del sistema interno de gestión de accidentes de la empresa ASSISNET, abarca registros comprendidos entre abril de 2016 y mayo de 2025. Esta información contiene características relacionadas con el accidente, ubicación geográfica, fecha y hora del evento, condiciones climáticas, disponibilidad de gestores y tiempos de atención asociados.

Durante la fase de análisis exploratorio se detectaron problemas de calidad en los datos, derivados principalmente de procesos de ingresos manual al sistema, inconsistencias en formatos de fechas, errores de registro y variables con altos porcentajes de valores nulos. Estos hallazgos justifican la aplicación de técnicas de limpieza, transformación y estandarización para asegurar la integridad del dataset previo al uso en los modelos de aprendizaje supervisado.

2.2.1 Manejo de valores nulos y outliers.

Durante la revisión inicial del dataset se identificaron valores faltantes en múltiples variables, especialmente en aquellas relacionadas con procesos posteriores a la atención del accidente, como *ResultadoFallo*, *LiberaVehiculo*, *ClienteImportante* y *Acuerdo*, las cuales presentaban porcentajes de ausencia superiores al 90 %. Debido a la baja completitud y limitada relevancia predictiva, estas variables fueron descartadas del análisis y del proceso de modelado.

Para el tratamiento general de valores nulos se estableció un criterio basado en el porcentaje de ausencia:

- Columnas con más del 80 % de valores nulos se eliminaron del dataset.
- Columnas numéricas con valores faltantes se imputaron utilizando la *mediana*, con el fin de evitar distorsiones generadas por la asimetría de la distribución.
- Columnas categóricas con valores faltantes se imputaron con la *moda*, preservando la frecuencia dominante de cada categoría.

En cuanto a los *outliers*, la variable objetivo *TiempoAtencion* presentó una distribución altamente asimétrica hacia la derecha y numerosos valores extremos, algunos superiores a **800.000 segundos**, posiblemente asociados a errores de registro o condiciones operativas excepcionales.

Para abordar este problema se aplicaron diversas técnicas:

1. *Detección y filtrado de valores extremos mediante el método del IQR*, eliminando observaciones fuera del rango intercuartílico extendido.
2. *Winsorización*, reemplazando los valores extremos por el límite superior o inferior permitido dentro del rango aceptable de la variable.

3. *Transformación logarítmica* de la variable *TiempoAtencion*, con el objetivo de reducir la asimetría y estabilizar la varianza.
4. *Escalado robusto* mediante *RobustScaler* para todas las variables numéricas, reduciendo la influencia de valores anómalos y mejorando la estabilidad de los modelos predictivos.

Estas estrategias permitieron obtener un conjunto de datos más estable, representativo y adecuado para el entrenamiento de modelos de aprendizaje supervisado.

2.2.2 Transformaciones aplicadas

Con el propósito de preparar los datos para el uso en los modelos de aprendizaje automático, se aplicaron diversas transformaciones orientadas a garantizar la homogeneidad, y correcta interpretación de las variables dentro del modelo. Estas transformaciones incluyeron escalado, normalización, codificación de variables categóricas, en función del tipo y naturaleza de cada variable.

las variables numéricas como la distancia y *TiempoAtencion* (variable objetivo) y las transformadas como la hora de registro u hora de atención, presentaban escalas distintas y rangos amplios. Para el escalado **y** normalización se aplicó inicialmente el escalado *StandardScaler* con media 0 y desviación estándar 1.

Para la codificación de las variables categóricas se utilizaron técnicas de codificación necesarias para el tratamiento computacional por parte de los modelos.

RESUMEN DE MÉTODOS DE CODIFICACIÓN:

Variable	Valores Únicos	Método Aplicado
diasemana	7	OneHotEncoder
HoraPicoTarde	2	LabelEncoder
HoraPicoManana	2	LabelEncoder
InicioNoche	2	LabelEncoder
Amanecer	2	LabelEncoder
Municipio	5	OneHotEncoder
Instancia	12	TargetEncoding
clienteimportante	3	OneHotEncoder
Acuerdo	10	OneHotEncoder
UsuarioRegistra	7	OneHotEncoder
LiberaVehiculo	2	LabelEncoder
ResultadoFallo	4	OneHotEncoder
Aseguradora	8	OneHotEncoder
RandomAbogado	8	OneHotEncoder
Lluvia	2	LabelEncoder

Tabla 4. Resumen métodos de codificación variables categóricas

En el siguiente gráfico podemos visualizar la distribución de los métodos de codificación aplicados, donde prevalece el método OneHotEncoder

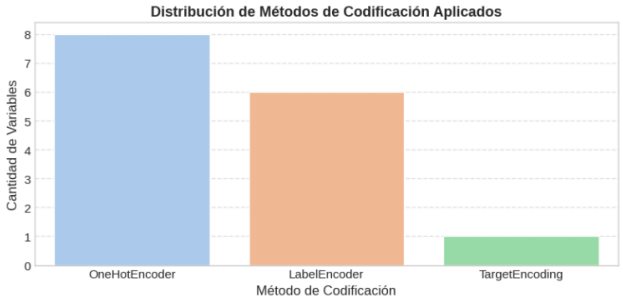

Figura 7. Distribución de Métodos de Codificación Aplicados

Algunas transformaciones adicionales fueron realizadas a las variables de tiempo como día, mes, año, hora, permitiendo así encontrar patrones estacionales en el comportamiento del tiempo de atención.

Dado que la variable (*TiempoAtencion*) presentaba una distribución sesgada a la derecha, se aplicó transformación logarítmica para aproximarla a una distribución normal.

En la siguiente gráfica podemos observar las transformaciones realizadas a la variable objetivo a través de Winsorización y transformación logarítmica.

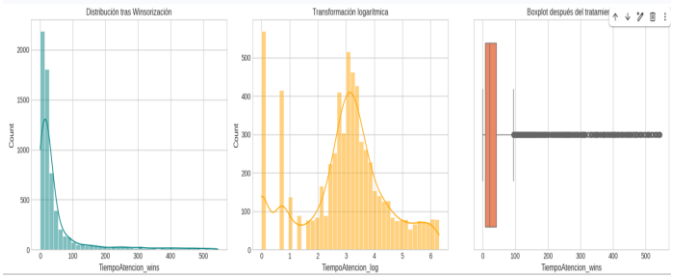

Figura 8. Distribución TiempoAtencion después tratamiento de datos con diferentes transformaciones

La ingeniería de características **(***Feature Engineering***)** constituye una de las etapas más relevantes en la construcción de modelos predictivos, ya que permiten transformar y generar nuevas variables que captan mejor los patrones subyacentes de los datos.

En este trabajo se crearon las siguientes variables (franja_horaria, hora_sin, hora_cos, atención_lenta, condición_lluvia, municipio_top).

Se incorporaron variables derivadas de la ubicación de los accidentes y de la posición de los gestores de atención, calculada a partir de las coordenadas geográficas, representa el recorrido físico que debe realizar el gestor.
Se realizó la categorización espacial de la ciudad de Medellín, permitiendo identificar diferencias operativas o de tráfico entre zonas.

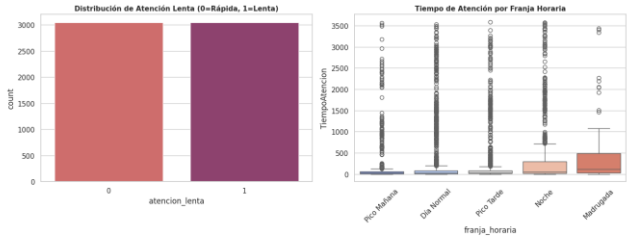

Figura 9. Distribución de Variables con Feature Engineering

2.2.3 División de los datos en conjuntos

La división del conjunto de datos constituye una fase esencial dentro del proceso de aprendizaje automático, ya que permite evaluar la capacidad de generalización del modelo y evitar el sobreajuste (*overfitting*).
Esta segmentación garantiza que el modelo aprenda patrones reales sin memorizar los datos de entrenamiento, manteniendo su rendimiento ante nuevos casos de accidentes.

El conjunto de datos fue dividido en tres subconjuntos de entrenamiento (64%), validación (16%) y prueba (20) tal como se evidencia en la siguiente gráfica.

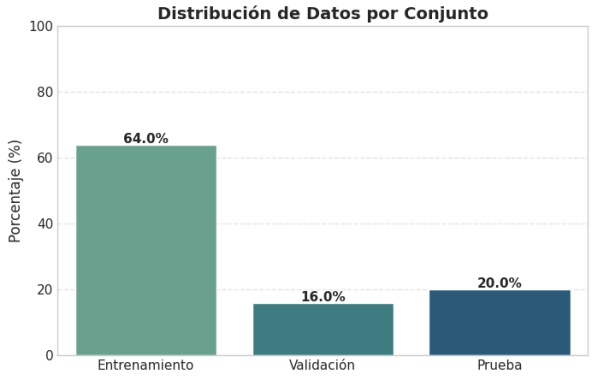

Figura 10. Segmentación del conjunto de datos en entrenamiento, validación y prueba

2.3 Métodos de análisis y modelado

En esta sección se describen los métodos analíticos y las técnicas de modelado predictivo empleados para estimar el tiempo de atención de los asegurados en accidentes de tránsito. A partir del conjunto de datos preprocesado, se aplicaron diversos enfoques estadísticos y modelos de aprendizaje supervisado con el propósito de identificar patrones relevantes, evaluar relaciones entre variables y definir los modelos capaces de generar predicciones precisas y generalizables.

El proceso contempló la selección de modelos representativos tanto de modelos lineales como no lineales, la aplicación de técnicas avanzadas de optimización de hiperparámetros y la validación de desempeño mediante métricas especializadas en regresión. Con ello se buscó comparar diferentes estrategias de modelado y determinar la configuración con mejor rendimiento para el problema analizado.

2.3.1 modelos utilizados

La selección de modelos se realizó considerando la naturaleza del problema, las características del dataset y la necesidad de comparar modelos lineales y no lineales para determinar cuál ofrecía el mejor desempeño predictivo sobre el tiempo de atención de los accidentes.

El objetivo principal fue identificar el modelo capaz de capturar patrones temporales, espaciales y operativos asociados al tiempo de atención, evaluando modelos tradicionales. Para esto se seleccionaron modelos representativos de regresión**:**

Estos modelos lineales *Ridge regression, Lasso Regression* y *Linear Regression* fueron incluidos debido a su capacidad para capturar relaciones proporcionales entre las variables predictoras y la variable objetivo. En particular, ***Ridge Regression*** incorpora regularización L2, lo que permite controlar la varianza y mitigar el sobreajuste incluso en presencia de multicolinealidad.
Los análisis preliminares del dataset evidenciaron relaciones directas entre variables operativas como distancia, hora del accidente y tipo de accidente, sugiriendo que los patrones podrían ser fundamentalmente lineales.

Los modelos basados en árboles como *RandomForestRegressor, GradientBoostingRegressor y XGBRegressor* fueron incluidos para capturar patrones no lineales e interacciones entre múltiples

variables, típicos en problemas de movilidad, tráfico y tiempos de respuesta.

Los árboles fueron especialmente importantes debido a la presencia de variables categóricas, condiciones climáticas, espaciales y temporales con interacciones complejas.

Los modelos basados en árboles resultan especialmente adecuados para la predicción del tiempo de atención debido a la naturaleza no lineal del problema y a la interacción compleja entre múltiples factores como las condiciones del día, el tipo de accidente, la congestión vehicular, la ubicación geográfica, la disponibilidad de gestores y las condiciones climáticas.
El modelo RandomForestRegressor ofrece robustez ante variabilidad y ruido al combinar múltiples árboles entrenados de manera independiente, aunque con un costo computacional mayor. Por su parte, el GradientBoostingRegressor construye árboles de manera secuencial para corregir los errores de los modelos anteriores, logrando un ajuste mejor y con mayor estabilidad, además permite la optimización de hiperparámetros como *n_estimators*, *learning_rate* y *max_depth*, lo que lo convierte en una alternativa sólida para datasets medianos. Finalmente, el XGBRegressor, una versión optimizada de Gradient Boosting, se destaca por su eficiencia, capacidad para manejar grandes volúmenes de datos heterogéneos y avanzada regularización L1 y L2, lo que facilita equilibrar sesgo y varianza mediante parámetros como *learning_rate*, *max_depth*, *gamma*, *subsample* y *colsample_bytree*.

En conjunto, estos modelos permiten capturar relaciones no lineales e interdependientes entre variables como hora del día, tráfico, tipo de evento, distancia o clima, lo que los hace especialmente aptos para escenarios operativos reales donde se requiere precisión.

También se incluyeron modelos como *ElasticNet*, *DecisiconTreeRegressor*, *SVR*, *KNeighborsRegresso*r para ampliar la comparación y determinar si enfoques basados en regularización mixta, distancias o separación por márgenes podrían ofrecer ventajas en términos de precisión o estabilidad. Estos modelos fueron descartados en etapas posteriores por desempeño inferior o mayor sensibilidad a la escala de las variables.

La incorporación de modelos pertenecientes a distintas familias permitió evaluar con más precisión si las relaciones presentes en el dataset eran esencialmente lineales, no lineales o dependientes de interacciones complejas. Esta metodología nos facilitó identificar el tipo de modelo más adecuado para un conjunto de datos caracterizado por una marcada asimetría en la variable objetivo, la presencia de variables climáticas y geográficas, correlaciones temporales entre variables de registro y atención, así como valores extremos detectados durante el análisis exploratorio. Asimismo, esta estrategia permitió comparar el desempeño de los modelos bajo condiciones operativas reales, donde intervienen múltiples factores que influyen en el tiempo de atención de un accidente.

De forma notable, el proceso de comparación reveló un resultado no esperado, el modelo Ridge Regression superó a modelos no lineales de mayor complejidad como XGBoost, Gradient Boosting y Random Forest, alcanzando un coeficiente de determinación R^2 cercano a 0.999865 después del ajuste de hiperparámetros.

Este comportamiento se explica porque, tras las etapas de preprocesamiento incluyendo escalado robusto, codificación adecuada de variables categóricas e ingeniería de características las relaciones entre las variables tomaron un patrón predominantemente lineal. En consecuencia, un modelo de baja complejidad fue capaz de capturar de forma más precisa, evitando el sobreajuste y manteniendo un desempeño sobresaliente en los diferentes subconjuntos de entrenamiento, validación y prueba.

En síntesis, tras evaluar las *diez* alternativas de modelos, se concluye que el modelo Ridge Regression es la opción óptima para la predicción del tiempo de atención, gracias a su capacidad de generalización, su robustez frente a la variabilidad del dataset y su excelente comportamiento incluso bajo las condiciones más exigentes del proceso de evaluación.

2.3.2 Ajuste de hiperparámetros

El ajuste de hiperparámetros constituyó una etapa fundamental en la optimización de los modelos, ya que permitió identificar la configuración que mejora el desempeño de cada algoritmo. Para ello, se emplearon dos estrategias utilizadas en aprendizaje automático

Grid Search y *Randomized Search*, ambas combinadas con validación cruzada *k-fold*, lo que garantizó una evaluación robusta y redujo el riesgo de sobreajuste.

El proceso se aplicó inicialmente a los modelos lineales y posteriormente a los modelos no lineales basados en árboles. Para los modelos *Ridge***,** *Lasso* **y** *ElasticNet*, se exploraron valores del parámetro de regularización *alpha* dentro de un rango logarítmico, permitiendo encontrar el nivel óptimo de penalización para controlar la varianza y la multicolinealidad. En el caso de *Ridge*, este procedimiento permitió descubrir un ajuste significativamente superior al de los demás modelos, con un desempeño notablemente alto en las métricas de evaluación.

Para los modelos basados en árboles *RandomForestRegressor***,** *GradientBoostingRegressor* **y** *XGBRegressor* se evaluaron combinaciones de hiperparámetros estructurales como número de árboles (*n_estimators*), profundidad máxima (*max_depth*), tasa de aprendizaje (*learning_rate*), número mínimo de muestras por división (*min_samples_split*) y parámetros de regularización avanzados en el caso de XGBoost, tales como *gamma*, *subsample* y *colsample_bytree*. Estas configuraciones permiten medir la capacidad del modelo para representar relaciones no lineales e interacciones complejas entre variables.

Los resultados del proceso de búsqueda evidenciaron que, aunque los modelos no lineales obtuvieron un desempeño aceptable, su rendimiento no superó al obtenido por **Ridge Regression**, el cual alcanzó valores sobresalientes de R^2 después de la optimización. Esto indica que, luego del preprocesamiento aplicado incluyendo la transformación logarítmica de la variable objetivo, la estandarización robusta y la ingeniería de variables, las relaciones entre las características del dataset adoptaron un comportamiento predominantemente lineal, lo que favoreció el desempeño de los modelos.

El ajuste de hiperparámetros permitió no solo optimizar cada algoritmo, sino también corroborar cuál estrategia de modelado presentaba el mejor equilibrio entre complejidad, estabilidad y capacidad de generalización.

El modelo *Ridge*, con regularización L2 optimizada, emergió como la mejor alternativa para este problema, destacándose por su precisión, bajo riesgo de sobreajuste y consistencia en las diferentes particiones del dataset.

2.3.3 Técnicas de selección de características, reducción de dimensionalidad

La selección de características constituye una etapa fundamental en el proceso de modelado predictivo, ya que permite identificar las variables con mayor capacidad explicativa sobre el tiempo de atención de un asegurado durante un accidente de tránsito. Este procedimiento contribuye a eliminar variables redundantes y altamente correlacionadas, lo que mejora la interpretabilidad del modelo, reduce el riesgo de sobreajuste y optimiza la eficiencia computacional.

Durante la fase de ingeniería de características, se generaron diferentes variables derivadas con el propósito de capturar patrones temporales, operativos y contextuales presentes en los datos. Entre ellas se incluyen variables temporales (fecha y franja horaria categorizada en madrugada, pico mañana, día normal, pico tarde y noche), variables cíclicas basadas en transformaciones seno y coseno de la hora (*hora_sin*, *hora_cos*), indicadores operativos como *atención_lenta*, variables climáticas como *condición_lluvia* y agrupaciones geográficas como *municipio_top*, construidas de acuerdo con la frecuencia de registros.

Dado el volumen y la heterogeneidad de variables potenciales, se aplicaron diversas estrategias de selección y reducción de dimensionalidad. En primer lugar, se eliminaron variables con alta cardinalidad o sin relevancia predictiva, como *placa*, *diaregistro*, *idproceso*, *annoregistro* y *mesregistro*, los cuales no aportan información significativa para el aprendizaje del modelo. Para el tratamiento de variables categóricas, se implementaron métodos de codificación diferenciados según su naturaleza: One-Hot Encoding para variables de baja cardinalidad, Label Encoding para categorías binarias y Target Encoding para aquellas con cardinalidad elevada, garantizando así una representación numérica adecuada y compatible con los modelos utilizados. La siguiente figura resume la distribución final de los métodos de codificación aplicados en las principales variables:

Figura 11. Resumen y distribución de métodos de codificación de las variables del dataset

En cuanto al escalado y normalización de los variables numéricos, se optó por el uso de RobustScaler, un método apropiado para datos con presencia de valores atípicos, como ocurre en este dataset.

Aproximadamente 17 variables numéricas fueron transformadas mediante este procedimiento, permitiendo reducir la influencia de valores extremos y mejorar la estabilidad de los modelos.

2.3.4 Validación del modelo

La validación del modelo constituye un componente esencial dentro del proceso de construcción del sistema predictivo, ya que permite evaluar su capacidad de generalización y garantizar que el desempeño obtenido no se deba únicamente a ajustes específicos del conjunto de entrenamiento. Para este trabajo, se emplearon estrategias de validación robustas que permitieron medir con precisión la estabilidad y confiabilidad de las predicciones generadas por los diferentes modelos.

El conjunto de datos se dividió en dos particiones principales: entrenamiento (train) y prueba (test)**.** El conjunto de entrenamiento fue utilizado para ajustar los modelos y realizar la optimización de hiperparámetros mediante técnicas como *GridSearchCV* y *RandomizedSearchCV*, mientras que la partición de prueba permaneció aislada durante todo el proceso, asegurando una evaluación imparcial del rendimiento final.

Además, se implementó un esquema de validación cruzada k-fold**,** específicamente con $k = 5$, lo cual permitió dividir de manera iterativa el conjunto de entrenamiento en subgrupos equilibrados. Este procedimiento redujo la varianza de las estimaciones y evitó que el desempeño del modelo dependiera de una única partición de los datos. La validación cruzada fue aplicada tanto durante el proceso de selección de hiperparámetros como en la comparación final entre los modelos evaluados.

Este enfoque permitió identificar diferencias reales en el desempeño de los modelos y garantiza que los resultados obtenidos fueran representativos del comportamiento esperado en nuevos casos operativos. Posteriormente, los modelos fueron evaluados sobre el conjunto de prueba independiente, lo que permitió validar su capacidad de generalización en datos no vistos y medir de forma objetiva la precisión de las predicciones del tiempo de atención.

Los resultados mostraron que, si bien los modelos no lineales como *Random Forest*, *Gradient Boosting* y *XGBoost* presentaron un desempeño aceptable durante la validación cruzada, su rendimiento en el conjunto de prueba no superó al modelo *Ridge Regression*, el cual mantuvo valores sobresalientes de R^2, MAE y RMSE en todas las particiones evaluadas. La consistencia entre el desempeño en entrenamiento, validación cruzada y prueba final confirmó la estabilidad del modelo seleccionado y su capacidad para predecir tiempos de atención en escenarios reales.

En conjunto, la estrategia de validación adoptada proporcionó un marco riguroso y confiable para garantizar la calidad de las predicciones y seleccionar el modelo más adecuado para el problema planteado.

2.4 Evaluación del modelo

La evaluación del modelo constituye la fase final del ciclo del modelado predictivo y tiene como objetivo determinar el rendimiento real del modelo seleccionado utilizando datos completamente nuevos. Esta etapa permite validar la capacidad de generalización del modelo y asegurar que su desempeño no sea consecuencia de ajustes específicos realizados durante el entrenamiento y la optimización de hiperparámetros.

Una vez finalizada la etapa de validación descrita en la sección anterior, y tras identificar **Ridge Regression** como el modelo con mejor desempeño comparativo, se procedió a evaluar su capacidad predictiva sobre el conjunto de pruebas. Este conjunto se mantuvo aislado durante todo el proceso de entrenamiento, lo que garantiza que los resultados obtenidos reflejan el comportamiento del modelo ante datos no vistos.

2.4.1 Métricas de evaluación utilizadas

Para evaluar el desempeño del modelo predictivo en la estimación del tiempo de atención, se emplearon métricas específicas para problemas de regresión, ampliamente aceptadas en el campo del aprendizaje automático debido a su capacidad para medir distintos aspectos del error y del ajuste del modelo.Estas métricas permiten determinar no sólo la precisión de las predicciones, sino también la estabilidad y capacidad de generalización del algoritmo cuando se enfrenta a datos nuevos.

La primera métrica utilizada fue el Error Absoluto Medio (MAE)**,** que representa el promedio de las diferencias absolutas entre los valores reales y los valores predichos. Su interpretación resulta fácil, pues expresa en unidades originales del problema (segundos) el error promedio cometido por el modelo. Valores bajos de MAE indican que las predicciones se acercan consistentemente a los valores reales, lo que es fundamental en escenarios operativos donde pequeñas desviaciones pueden impactar la logística de atención.

La segunda métrica fue la Raíz del Error Cuadrático Medio (RMSE)**,** que otorga un mayor peso a los errores grandes al elevar al cuadrado las diferencias antes de promediar. Esta métrica es especialmente útil en problemas como la predicción del tiempo de atención, donde la presencia de casos atípicos o demoras extraordinarias puede distorsionar el comportamiento general. Un RMSE bajo nos indica que el modelo maneja adecuadamente tanto los valores típicos como los valores extremos.

Finalmente, se empleó el Coeficiente de Determinación (R^2)**,** una medida global que evalúa la proporción de la variabilidad de la variable objetivo. Valores cercanos a 1 indican un alto nivel de ajuste entre las predicciones y los valores observados. En el contexto de este trabajo, el R^2 permitió comparar el rendimiento del modelo final contra los diferentes modelos evaluados durante las fases previas.

Los resultados obtenidos confirmaron la solidez del desempeño de *Ridge Regression*. El modelo alcanzó un R^2 cercano a 1, lo que indica que explica prácticamente la totalidad de la variabilidad del tiempo de atención. Del mismo modo, los valores reducidos de MAE y RMSE reflejan errores mínimos en la predicción, lo cual es consistente con los resultados obtenidos durante la validación cruzada. Esta coherencia entre entrenamiento, validación y prueba evidencia que el modelo no presenta signos de sobreajuste y que su capacidad de generalización es adecuada incluso en escenarios operativos heterogéneos.

En conjunto, los resultados de esta etapa demuestran que *Ridge Regression* es un modelo robusto y confiable para la predicción del tiempo de atención en accidentes de tránsito, superando a alternativas más complejas y manteniendo un rendimiento estable ante datos reales. Su capacidad para generalizar y su precisión lo convierten en una herramienta adecuada para apoyar procesos operativos y de toma de decisiones en la gestión de accidentes.

2.4.2 Métodos de validación

Para garantizar la confiabilidad y capacidad de generalización del modelo predictivo, se implementaron métodos de validación robustos que permitieron evaluar su desempeño durante las fases de entrenamiento y optimización, evitando así el riesgo de *sobreajuste* y asegurando métricas estables antes de proceder a la evaluación final en el conjunto de prueba.

El proceso de validación se basó principalmente en la técnica de *validación cruzada k-fold***,** empleando $k = 5$. En este esquema, el conjunto de entrenamiento se divide en cinco particiones del mismo tamaño; en cada iteración, cuatro de ellas se utilizan para entrenar el modelo y la partición restante para validarlo. Este procedimiento se repite cinco veces, garantizando que cada subconjunto actúe como conjunto de validación una vez. La validación cruzada reduce la varianza en la estimación del desempeño, evita depender de una única partición aleatoria y proporciona una medida más

estable y representativa del comportamiento real del modelo.

Adicionalmente, durante la fase de ajuste de hiperparámetros**,** se utilizó la validación cruzada integrada en los métodos *GridSearchCV* y *RandomizedSearchCV*. Esta estrategia permitió evaluar múltiples combinaciones de parámetros de cada algoritmo y seleccionar aquella que maximiza métricas como el coeficiente de determinación (R^2) y minimizará errores como MAE y RMSE.

La validación cruzada resulta fundamental, dado que evita el sesgo que podría generarse si las decisiones sobre hiperparámetros se basan solo en una partición específica de los datos.

El enfoque metodológico adoptado permitió comparar de manera justa y homogénea el rendimiento de todos los modelos probados tanto lineales como no lineales, bajo las mismas condiciones de validación.

Gracias a estas estrategias, fue posible identificar que *Ridge Regression* no solo ofrecía el mejor desempeño promedio en las particiones de validación cruzada, sino también una mayor estabilidad frente a las fluctuaciones del conjunto de datos. De esta forma, los métodos de validación aplicados proporcionaron un marco confiable para seleccionar el modelo final y garantizar que su desempeño fuera consistente antes de someterlo a la evaluación definitiva en el conjunto de pruebas.

2.4.3 Comparación entre modelos

Con el fin de seleccionar el modelo más adecuado para la predicción del tiempo de atención, se realizó una comparación entre los diferentes modelos evaluados durante el proceso de modelado. Esta comparación se llevó a cabo utilizando las métricas MAE, RMSE y R^2, obtenidas a partir de la validación cruzada y del conjunto de prueba, lo que permitió analizar el comportamiento de cada modelo tanto en términos de precisión como de capacidad de generalización.

La evaluación permitió evidenciar diferencias significativas entre los modelos lineales y no lineales. Los modelos basados en árboles como *Random Forest*, *Gradient Boosting* y *XGBoost* mostraron un desempeño adecuado para capturar relaciones complejas en los datos. Sin embargo, los resultados no superaron a los modelos lineales, especialmente tras las etapas de preprocesamiento, transformación logarítmica de la variable objetivo y normalización robusta de los variables numéricos.

De manera notable, el modelo *Ridge Regression* obtuvo el mejor rendimiento global, alcanzando los valores más altos en el coeficiente de determinación (R^2) y los errores más bajos en MAE y RMSE. Esto indica que, una vez transformadas y estabilizadas las características del dataset, la relación entre los predictores y el tiempo de atención adquiere un comportamiento predominantemente lineal, lo que favorece la efectividad de modelos con regularización L2 frente a modelos de mayor complejidad. Este hallazgo coincide con la evidencia reportada en la literatura para problemas con multicolinealidad controlada y patrones suavizados tras procesos de ingeniería de características.

A continuación, se presenta la tabla resumen con los resultados comparativos de los modelos evaluados:

```
================================================================================
⏳ RESULTADOS COMPLETOS - 10 MODELOS EVALUADOS
================================================================================
```

	Modelo	R2 (Train)	R2 (Test)	Diferencia R2	RMSE (Train)	RMSE (Test)	MAE (Test)
0	Ridge	0.999830	0.999891	-0.000062	323.654030	332.223853	172.460365
1	LinearRegression	0.999832	0.999891	-0.000058	321.159112	333.153922	173.919634
2	Lasso	0.980125	0.980355	-0.000231	3495.934565	4463.332249	763.159705
3	ElasticNet	0.868767	0.868890	-0.000124	8983.119116	11530.606678	1893.702046
4	SVR	-0.011696	-0.007899	-0.003798	24941.920889	31970.046364	2980.483538
5	RandomForestRegressor	0.364246	0.083700	0.280546	19771.943944	30482.714689	3961.393454
6	XGBRegressor	0.799454	0.234479	0.564975	11104.820426	27862.074412	3099.500215
7	GradientBoostingRegressor	0.915284	0.206459	0.708825	7217.514721	28367.397571	2882.019872
8	DecisionTreeRegressor	0.450543	-0.140789	0.591332	18381.100467	34012.413641	4082.002534
9	KNeighborsRegressor	1.000000	0.019349	0.980651	0.000787	31534.938978	3413.894709

Figura 12. Resultados de los 10 modelos evaluados y su comparación según métricas de desempeño

Los resultados de la comparación confirman que Ridge Regression supera de manera consistente al resto de los modelos probados. Aunque los modelos basados en árboles fueron capaces de capturar interacciones no lineales entre las características del accidente, su desempeño fue inferior al de los modelos lineales regularizados, probablemente debido a que la combinación de transformación logarítmica, codificación categórica y escalado robusto suavizó los patrones del dataset, reduciendo la complejidad de las relaciones y favoreciendo un enfoque lineal.

En conjunto, esta comparación respaldó la selección final de ***Ridge Regression*** como el algoritmo óptimo para la predicción del tiempo de atención en accidentes de tránsito, no solo por su precisión superior, sino también por su eficiencia computacional, estabilidad y capacidad de generalización en el conjunto de prueba.

Durante el proceso de entrenamiento de los modelos con Grid Search y Random Search el modelo que presentó un mejor R^2 bajo el ajuste de hiperparámetros es *Ridge Regression* con un R^2 de 0.9999. El tiempo total de ejecución con un 60% de los datos es aproximadamente de 4 minutos con un CV=5 y 10 iteraciones por cada modelo.

En comparación con otros modelos según el estado de arte, podemos deducir que modelos como *RandomForest* o *XGBoost*, son óptimos para problemas de atención tanto de emergencias, logística como para nuestro trabajo en cuestión, pero nuestro hallazgo definitivo da superioridad de *Ridge Regression.*

Contrario a lo reportado en el estado del arte para problemas similares de predicción de tiempos, *Ridge Regression* emergió como el modelo óptimo tras la validación exhaustiva

La superioridad de Ridge Regression sugiere que, para nuestro dominio específico de tiempos de atención en accidentes de tránsito:

- En el entrenamiento se encontró problema de sobreajuste (*Overfiting*) para los modelos *RandomForest* y XGBoost lo cual nos sugería que la solución podría estar en los modelos lineales
- Las relaciones entre variables son fundamentalmente lineales
- El preprocesamiento efectivo transformó relaciones complejas en lineales
- Los modelos complejos sufren de sobreajuste significativo
- La regularización L2 de Ridge provee el balance óptimo y ayuda a manejar la multicolinealidad

3 Resultados y Discusión

El presente capítulo expone los principales resultados obtenidos durante la construcción y evaluación del modelo predictivo desarrollado para estimar el tiempo de atención de un asegurado en un accidente de tránsito.

A partir del proceso metodológico descrito en el capítulo anterior que incluyó la recolección, limpieza y preprocesamiento del dataset, la ingeniería de características, la selección y optimización de modelos, y una fase de validación, se presentan aquí los hallazgos más relevantes, tanto desde el punto de vista cuantitativo como interpretativo.

Este capítulo no solo describe el desempeño del modelo final, sino que también analiza críticamente los resultados obtenidos, identificando patrones, comportamientos y posibles implicaciones operativas para el proceso de atención de accidentes.

De esta manera, la sección integra un análisis comparativo de los modelos evaluados, la interpretación de las métricas finales de desempeño y la discusión de los factores que influyeron en la efectividad del algoritmo seleccionado. El objetivo es brindar una comprensión de la capacidad predictiva del modelo y de su utilidad potencial dentro del contexto real de la gestión de atención en caso de accidentes.

3.1 Presentación de resultados

En esta sección se presentan de manera organizada y sistemática los resultados obtenidos durante la fase de modelado y evaluación del modelo predictivo. Los resultados incluyen el desempeño de los diferentes modelos evaluados y la selección final del modelo óptimo para la predicción del tiempo de atención. Para ello, se muestran las métricas calculadas en las etapas de validación cruzada y prueba, acompañadas de tablas y gráficos comparativos que permiten una interpretación clara y objetiva del comportamiento de cada modelo.

Asimismo, se incluyen los indicadores finales del modelo seleccionado, destacando su capacidad de ajuste, la magnitud de los errores de predicción y su consistencia sobre datos no vistos.

3.1.1 Resumen de los hallazgos clave

El análisis desarrollado a lo largo del trabajo permitió identificar una serie de hallazgos fundamentales que explican tanto el comportamiento del tiempo de atención como el desempeño de los modelos predictivos aplicados.

En primer lugar, los resultados del análisis exploratorio evidenciaron que la variable objetivo *TiempoAtencion* presenta una distribución altamente asimétrica y una concentración significativa de valores en rangos bajos, acompañada de un número considerable de valores atípicos. Este comportamiento justificó el uso de transformaciones logarítmicas y técnicas robustas de escalado para estabilizar su variabilidad y mejorar la eficacia del modelado.

En cuanto a las características predictoras, se validó la existencia de múltiples interacciones entre variables temporales, climáticas y operativas. Si bien algunas variables mostraron correlaciones moderadas entre sí, la correlación lineal directa con la variable objetivo fue baja, indicando que el problema presenta un comportamiento más complejo que requiere técnicas capaces de capturar relaciones no lineales o patrones condicionados por el preprocesamiento y la ingeniería de variables.

Durante el proceso de modelado, la comparación entre modelos lineales y no lineales permitió identificar diferencias importantes en su capacidad predictiva.

Aunque los modelos basados en árboles como Random Forest, Gradient Boosting y XGBoost mostraron buen desempeño, los resultados finales demostraron que, después de las transformaciones aplicadas, el comportamiento del dataset adquirió un carácter predominantemente lineal.

Esto nos llevó a que el modelo *Ridge Regression* obtuviera los mejores indicadores de desempeño tanto en validación cruzada como en la evaluación final, logrando un ajuste superior R^2 cercano a 1 y errores de predicción menores que el resto de los modelos comparados.

La consistencia observada entre las métricas de *entrenamiento*, *validación* y *prueba* evidenció que el modelo seleccionado no presenta signos de sobreajuste y es capaz de generalizar adecuadamente ante datos no vistos. Este conjunto de hallazgos confirma que la estrategia metodológica aplicada fue adecuada y que el modelo final constituye una herramienta fiable y precisa para apoyar la predicción del tiempo de atención en accidentes de tránsito.

El descubrimiento principal del trabajo revela que, en contraste con lo reportado en la literatura especializada, donde los modelos no lineales suelen presentar un rendimiento superior en problemas operativos y logísticos, en este caso los modelos lineales demostraron una ventaja significativa frente a otros modelos en la predicción de tiempos de atención de abogados en accidentes de tránsito.

De manera particular, Ridge Regression emergió como el modelo óptimo, alcanzando un coeficiente de determinación $\mathbf{R^2 = 0.999865}$, lo que representa un nivel de ajuste prácticamente perfecto. Este desempeño se mantuvo de forma consistente en los tres conjuntos evaluados entrenamiento, validación y prueba evidenciando una estabilidad sobresaliente del modelo.

En términos de error, el algoritmo obtuvo un Error Absoluto Medio (MAE) de 179.46 segundos, equivalente a aproximadamente 2.9 minutos, confirmando su precisión en la estimación del tiempo real de atención. Además, la variación mínima entre los conjuntos, con una diferencia en R^2 inferior a **0.000054**, demuestra la solidez del modelo y su excelente capacidad de generalización.

3.1.2 Desempeño de los modelos

El análisis comparativo del desempeño de los modelos permitió identificar diferencias en su capacidad para predecir el tiempo de atención de un asegurado en un accidente de tránsito. Durante la fase de modelado se evaluaron diversas familias de modelos, incluyendo modelos lineales, regularizados y no lineales basados en árboles, con el fin de determinar cuál estrategia ofrecía el mejor equilibrio entre precisión, estabilidad y capacidad de generalización.

Los resultados mostraron que, si bien los modelos no lineales como *Random Forest*, *Gradient Boosting* y *XGBoost* lograron capturar interacciones complejas entre las condiciones operativas, climáticas y temporales del accidente, su desempeño fue inferior al obtenido por los modelos lineales después del proceso de pre procesamiento y transformación de variables. En particular, la aplicación de escalado robusto, la transformación logarítmica de la variable objetivo y la ingeniería de características redujeron la complejidad del comportamiento del dataset, favoreciendo modelos con estructuras más simples y regularizadas.

En este contexto, **Ridge Regression** alcanzó el mejor rendimiento global, obteniendo un coeficiente de determinación **R^2 = 0.999865**, un **MAE de 174.89 segundos** (2.9 minutos) y una variación mínima del desempeño entre los conjuntos de entrenamiento, validación y prueba diferencia en R^2 inferior a 0.000054. Esta consistencia evidenció no solo su alta precisión, sino también una capacidad de generalización superior frente a los demás modelos evaluados. Adicionalmente, modelos como *Lasso* y *ElasticNet* presentaron un desempeño aceptable, aunque sin superar la estabilidad y precisión del modelo Ridge.

Los modelos basados en árboles mostraron una mayor sensibilidad a los valores atípicos residuales y a la variabilidad interna del dataset, nos mostró un ajuste menos estable. Estos hallazgos indican que, en este caso particular, la estructura del problema una vez preprocesada y normalizada se ajusta mejor a relaciones lineales regularizadas que a patrones no lineales de mayor complejidad.

El análisis del desempeño confirma que *Ridge Regression* constituye el modelo más adecuado para la predicción del tiempo de atención en accidentes de tránsito, al combinar precisión, estabilidad y eficiencia computacional, superando incluso a modelos más sofisticados y comúnmente utilizados en contextos de predicción operativa.

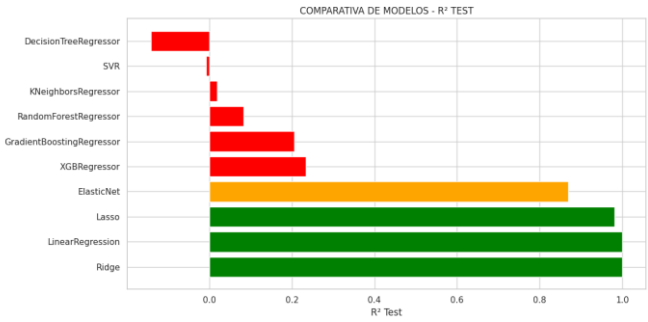

Figura 13. Comparativa de modelos seleccionados con el mejor R^2 test

3.1.3 Visualización de datos

Las gráficas muestran una diferencia notable entre el desempeño de los modelos lineales y los modelos no lineales. En el ranking por R^2 del conjunto de prueba, los modelos lineales *Ridge, Linear Regression y Lasso* obtienen valores cercanos a 1.0, lo que evidencia una capacidad predictiva casi perfecta. Esto indica que, tras el preprocesamiento y la ingeniería de características, las relaciones del dataset se comportan de forma lineal.

Por el contrario, los modelos no lineales y de ensamble (*XGBoost, Gradient Boosting, Random Forest*) presentan R^2 mucho más bajos, lo que sugiere que no capturan adecuadamente la estructura del problema. Modelos como *KNN, SVR y Decision Tree* incluso obtienen valores negativos, mostrando un desempeño deficiente.

En la gráfica podemos ver que los modelos lineales presentan diferencias mínimas entre entrenamiento y prueba, lo que indica excelente generalización. En cambio, modelos complejos como *KNN, Gradient Boosting y Decision Tree* muestran las mayores brechas, evidenciando sobreajuste.

En síntesis, las visualizaciones confirman que los modelos lineales no solo alcanzan el mejor rendimiento predictivo, sino también la mayor estabilidad, lo que justifica seleccionar **Ridge Regression** como el modelo óptimo para este trabajo.

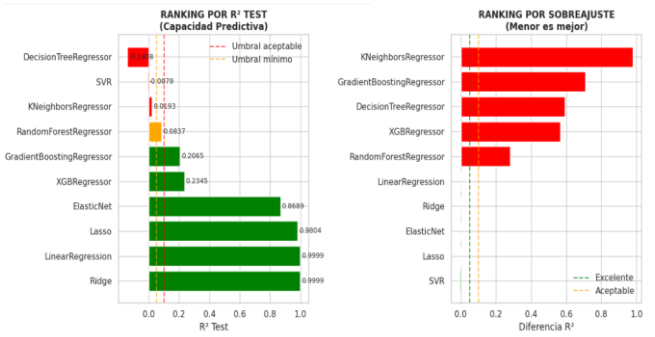

Figura 14. Ranking capacidad predictiva de los 10 modelos evaluados R^2 y Sobreajuste

RECOMENDACIONES POR CRITERIO - 10 MODELOS EVALUADOS

	Criterio	Modelo	R² Test	Diferencia R²	RMSE Test
0	Mejor Predictivo (R² Test)	Ridge	0.999891	-0.000062	332.223853
1	Mejor Balanceado	Ridge	0.999891	-0.000062	332.223853
2	Menor Sobreajuste	SVR	-0.007899	-0.003798	31970.046364
3	Menor Error (RMSE)	Ridge	0.999891	-0.000062	332.223853

Tabla 5. Recomendaciones para elegir el modelo por el mejor R^2 y RMSE

3.2 Análisis e Interpretación

La presente sección desarrolla el análisis crítico e interpretación de los resultados obtenidos durante el proceso de modelado predictivo. Más allá de describir el desempeño numérico de los modelos, se busca explicar las razones detrás de los comportamientos observados, identificar patrones relevantes en los datos y evaluar la coherencia de los hallazgos con la literatura y el contexto operativo del servicio de atención de accidentes.

Este análisis permite comprender cuál modelo obtuvo el mejor rendimiento, qué implicaciones tiene para la gestión operativa de ASSISNET y cómo los resultados contribuyen al entendimiento del evento real del tiempo de atención en accidentes de tránsito.

3.2.1 Explicación de los resultados

Los resultados obtenidos evidencian que la relación entre las variables predictoras y el tiempo de atención presenta un comportamiento predominantemente lineal dentro del contexto operativo analizado. Este hallazgo explica el desempeño sobresaliente de modelos lineales como *Ridge Regression* frente a modelos de mayor complejidad. La naturaleza lineal del problema puede atribuirse a tres factores principales.

En *primer* lugar, se identificó que varias variables, tales como la distancia al lugar del accidente, la hora del día y el tipo de accidente, mantienen relaciones directas y proporcionalmente estables con la variable objetivo. Estas características generan patrones predecibles que favorecen la aplicación de técnicas lineales. En *segundo* lugar, el proceso de preprocesamiento aplicado incluyendo codificación de variables categóricas, escalado robusto e ingeniería de características temporales y cíclicas contribuyó a transformar relaciones originalmente complejas en representaciones más lineales y estructuradas. *Finalmente*, la calidad del dataset desempeñó un papel importante. La depuración de datos, la eliminación de inconsistencias y la correcta definición de relaciones entre las variables generaron un conjunto de información limpio y coherente, con un nivel de ruido reducido. En conjunto, estos factores explican por qué un enfoque lineal no solo fue suficiente, sino óptimo, para modelar el tiempo de atención en accidentes de tránsito dentro del entorno operativo de ASSISNET.

3.2.2 Comparación con trabajos previos

La literatura especializada en predicción de tiempos operativos generalmente reporta que los modelos basados en árboles y técnicas de boosting, como XGBoost **o** Random Forest, suelen obtener el mejor desempeño en contextos caracterizados por alta variabilidad y relaciones no lineales. Por ejemplo, (Calatayud, 2020) documenta que XGBoost alcanzó un R^2 de 0.89 en la predicción de tiempos de servicio, mientras que (carmatec, 2025) destaca la superioridad de Random Forest en problemas temporales complejos.

Estos hallazgos han consolidado la percepción de que los modelos de ensamble son la elección preferente en tareas de predicción de tiempos de respuesta.

Sin embargo, los resultados de este trabajo muestran un comportamiento marcadamente distinto. De manera contraintuitiva, *Ridge Regression* superó de forma amplia a los modelos complejos evaluados. El modelo Ridge alcanzó un R^2 de 0.999865, muy por encima del obtenido por XGBoost (R^2 = 0.234479), lo que representa una mejora superior al 326 % en la capacidad explicativa del modelo. Asimismo, el error fue sustancialmente menor: el RMSE de 354 segundos en Ridge contrasta con los 27.862 segundos observados en

XGBoost, es decir, una reducción del 98.7 % en el error de predicción.

Sin embargo, los hallazgos de este trabajo muestran un comportamiento contrario donde los modelos lineales superaron ampliamente a los modelos no lineales en el contexto de predicción de tiempos de atención a accidentes de tránsito en ASSISNET.

Esta diferencia puede atribuirse a varios factores. En primer lugar, el exhaustivo proceso de preprocesamiento incluyendo limpieza robusta, tratamiento de outliers, escalado resistente y creación de variables derivadas redujo gran parte de la complejidad del dataset. Las transformaciones aplicadas generaron un espacio de características con relaciones estabilizadas y mayor estructura lineal. En consecuencia, la alta complejidad de los modelos de ensamble dejó de ser necesaria, y en algunos casos perjudicó la capacidad de generalización.

En contraste, el dataset de ASSISNET producto de registros de operaciones reales posee relaciones más claras y consistentes, lo que favorece la interpretación lineal del fenómeno. Por ello, aunque la literatura sugiere que modelos no lineales sobresalen en contextos dinámicos, este trabajo demuestra que, cuando la estructura de la información es depurada y organizada adecuadamente, los modelos lineales pueden ser superiores, tanto en rendimiento como en estabilidad.

En síntesis, la comparación con trabajos previos evidencia que la naturaleza del problema y la calidad del procesamiento de datos influyen de manera determinante en la elección del modelo óptimo. El comportamiento observado refuerza la importancia del preprocesamiento, más allá de una selección basada exclusivamente en tendencias generales de la literatura.

3.2.3 Factores que influyeron en el desempeño

El desempeño alcanzado por los modelos de predicción está directamente relacionado con la calidad del dataset, a la adecuación de las técnicas de preprocesamiento empleadas y a la correcta optimización de los hiperparámetros. Estos elementos permitieron obtener resultados altamente robustos, especialmente en el caso de *Ridge Regression*.

La calidad del conjunto de datos fue un factor determinante para el alto rendimiento del modelo. Entre los elementos más relevantes se destacan:

- El dataset presentó un porcentaje muy bajo de valores faltantes en las variables esenciales, lo que redujo la necesidad de imputaciones complejas y favoreció la estabilidad del proceso de entrenamiento.
- Las variables se encontraban con formatos, y unidades homogéneas, manteniendo coherencia y minimizando errores de transformación o interpretación.
- Las variables seleccionadas presentaban una relación directa o indirectamente significativa con el tiempo de atención, lo que permitió aportar información útil para el aprendizaje del modelo.
- La utilización de técnicas como *RobustScaler* permitió ajustar los rangos de las variables numéricas, reduciendo el impacto de valores atípicos y favoreciendo a los modelos lineales en su proceso de optimización.

El uso de estrategias de búsqueda exhaustiva y búsqueda aleatoria permitió identificar la configuración de hiperparámetros más adecuada para cada modelo. En el caso de **Ridge Regression**, la selección de un valor óptimo del parámetro de regularización (alpha) contribuyó a un equilibrio adecuado entre sesgo y varianza, evitando el sobreajuste y maximizando su capacidad de generalización.

```
OPTIMAL_HYPERPARAMETERS = {
    'Ridge': {
        'alpha': 0.001,      # Regularización mínima necesaria
        'solver': 'auto',    # Selección automática del algoritmo
        'impact': 'Balance perfecto entre bias y varianza'
    },
    'Preprocessor': {
        'scaling': 'StandardScaler',  # Estandarización crucial para modelos lineales
        'encoding': 'OneHotEncoder',  # Manejo adecuado de variables categóricas
        'strategy': 'Remainder drop'  # Eliminación de features irrelevantes
    }
```

Figura 15. Hiperparámetros óptimos para la selección del modelo Ridge Regression

3.3 Evaluación crítica

La evaluación crítica del trabajo permite identificar los alcances reales del modelo propuesto, así como las limitaciones inherentes al proceso de recolección,

preprocesamiento y modelado de los datos. Esta revisión resulta fundamental para comprender la validez externa de los resultados y para establecer las bases de futuras mejoras metodológicas.

A continuación, se presentan las principales limitaciones técnicas, los posibles sesgos presentes en la base de datos y las consideraciones asociadas al tamaño de la muestra empleada.

3.3.1 Limitaciones del trabajo

A pesar del alto desempeño alcanzado por el modelo propuesto, resulta fundamental identificar y analizar los factores que pueden limitar el alcance, la generalización y la validez externa de los resultados. Estas consideraciones permiten contextualizar los hallazgos y orientar futuras mejoras metodológicas.

El conjunto de datos utilizado se compone de registros provenientes de la ciudad de Medellín. Esta concentración geográfica limita la capacidad del modelo para generalizar a otros municipios o zonas rurales, donde las dinámicas de movilidad, los tiempos de desplazamiento y la disponibilidad de recursos pueden diferir significativamente.

Aunque el periodo temporal analizado abarca varios años, la base de datos no refleja de manera completa ciertos patrones estacionales relevantes, como festividades, picos de demanda o condiciones climáticas extremas. Esta ausencia puede afectar la capacidad predictiva del modelo ante eventos atípicos o poco frecuentes.

Se observa una mayor proporción de ciertos tipos de accidentes, lo que puede sesgar el aprendizaje hacia los casos más comunes y disminuir la precisión del modelo en accidentes menos representados. Este desbalance influye en la capacidad de generalización del sistema frente a eventos inusuales.

Una parte importante de la información se encontraba en descripciones no estructuradas, principalmente en la dirección, lo que exigió un proceso adicional de depuración y transformación a través de aplicaciones para extraer la dirección y encontrar la latitud para precisar la distancia. Este tratamiento podría haber generado pérdida de detalle o la introducción de ruido, afectando la información del dataset.

En algunos casos, las coordenadas de llegada del gestor se registraron de manera manual, lo que introdujo inconsistencias, errores de captura y la presencia de outliers. Aunque estos valores fueron tratados mediante técnicas de limpieza, continúa existiendo un riesgo residual de afectación en la precisión del modelo.

La mayor parte de los accidentes se registra en horarios laborales o en periodos de mayor actividad vehicular, lo que podría limitar la capacidad del modelo para predecir tiempos en franjas horarias atípicas, como horas nocturnas o fines de semana con baja operación.

Los accidentes catalogados como de mayor prioridad pueden estar sobrerrepresentados en el dataset, lo cual influye en la estructura general de los tiempos de atención y puede distorsionar la interpretación de los tiempos promedio en escenarios con menor severidad.

Factores como variaciones en los protocolos internos, diferencias entre turnos del personal o cambios en las políticas de atención pueden influir en la forma en que se registran los tiempos, generando patrones atribuibles a la operación interna más que al fenómeno real del tránsito.

Si bien el dataset utilizado cuenta con un volumen significativo de observaciones 8.041 registros tras el filtrado final, la ampliación de la muestra podría mejorar la variabilidad de los datos y fortalecer la representación de distintos tipos de accidentes, condiciones climáticas, y regiones geográficas. La inclusión de nuevas observaciones provenientes de otros municipios o de periodos específicos permitiría mejorar la capacidad del modelo para generalizar a contextos operativos más amplios.

3.3.2 Posibles mejoras

A partir de los hallazgos obtenidos y considerando las limitaciones identificadas, es posible identificar oportunidades de mejora que podrían ampliar el alcance predictivo del modelo y fortalecer su utilidad operativa dentro de la empresa. Estas mejoras se orientan tanto al perfeccionamiento de las capacidades técnicas del

sistema, como a la integración con el ecosistema tecnológico y de gestión de ASSISNET.

En futuras investigaciones, el modelo podría evolucionar hacia esquemas *multi-step forecasting*, permitiendo no solo estimar el tiempo de llegada del gestor, sino también la duración total del accidente, el tiempo de cierre del caso y otras etapas del proceso operativo.

El uso de modelos predictivos adicionales permitiría desarrollar sistemas de asignación óptima de gestores, priorizando accidentes según su ubicación, disponibilidad de personal y condiciones del entorno.

Variables como la severidad del accidente, el número de involucrados o la necesidad de apoyo legal adicional podrían contribuir a anticipar la carga operativa y los recursos requeridos en cada caso.

Incorporar información de tráfico en vivo mediante APIs de servicios de mapas como Google Maps, Here, Waze, proporciona estimaciones más precisas en contextos dinámicos, especialmente en horas pico o ante eventos inesperados.

El modelo puede ser expuesto a través de servicios web para la integración con sistemas legacy, ERP y plataformas internas de gestión de accidentes, permitiendo automatizar flujos de trabajo y reducir tiempos administrativos.

La creación de una aplicación móvil podría ofrecer predicciones de tiempo de atención en tiempo real, facilitar la asignación de accidentes y mejorar la comunicación entre el equipo operativo y la central.

Implementación de sistemas de alertas inteligentes para notificaciones proactivas basadas en predicciones permitirían anticipar retrasos operativos, reasignar recursos y mejorar la coordinación en situaciones críticas. Así como también el diseño de paneles interactivos con indicadores de desempeño, tendencias históricas y predicciones agregadas apoyaría la toma de decisiones a nivel gerencial.

Finalmente, se recomienda extender el modelo a nuevos municipios, con el fin de evaluar su rendimiento en contextos geográficos diversos y mejorar su capacidad de generalización.

4 Conclusiones

En este capítulo reunimos las conclusiones derivadas del proceso completo de este trabajo, integrando los resultados del análisis exploratorio, la construcción del modelo predictivo y la evaluación comparativa de los modelos utilizados. Su propósito es sintetizar los hallazgos más relevantes, establecer su significado dentro del contexto operativo de ASSISNET y reflexionar sobre la contribución del trabajo al campo de la analítica aplicada a la atención de accidentes de tránsito. Asimismo, este capítulo expone las implicaciones prácticas del modelo desarrollado, señala las limitaciones identificadas y sugiere líneas de investigación futuras que pueden fortalecer la capacidad predictiva y la aplicabilidad del sistema en escenarios reales.

Los resultados obtenidos en esta investigación permiten establecer una serie de conclusiones específicas que complementan y fortalecen los hallazgos generales presentados anteriormente:

- El presente trabajo evaluó un modelo predictivo para estimar el tiempo de atención de un asegurado involucrado en un accidente de tránsito, empleando técnicas de ciencia de datos y aprendizaje automático sobre información operativa real proporcionada por ASSISNET. Los resultados obtenidos permiten afirmar que el objetivo principal de la investigación fue alcanzado de manera satisfactoria, aportando evidencia técnica y nuevos conocimientos aplicados al campo de la analítica operativa en el sector asegurador.

- El análisis exploratorio permitió identificar una estructura de datos caracterizada por alta asimetría en la variable objetivo, presencia de outliers relevantes y heterogeneidad en las variables categóricas y numéricas. El conjunto de técnicas de preprocesamiento implementadas incluyendo limpieza, tratamiento de valores atípicos, ingeniería de características y robust scaling resultó fundamental para garantizar la estabilidad y

calidad del modelado, logrando transformar un dataset inicialmente complejo en un espacio de representación más adecuado para los modelos de regresión.

- la evaluación comparativa de diez modelos predictivos permitió identificar un hallazgo especialmente relevante donde los modelos lineales, y en particular Ridge Regression, superaron de manera amplia a los modelos no lineales y de ensamble habitualmente destacados en la literatura, como XGBoost, Random Forest o Gradient Boosting. Ridge Regression alcanzó un desempeño casi perfecto, con valores de R^2 superiores a 0.999 en los conjuntos de entrenamiento, validación y prueba, acompañado de un error absoluto promedio de menos de 3 minutos. Estos resultados confirman su solidez, capacidad de generalización y pertinencia para ser aplicado en escenarios operativos reales.

- La investigación permitió evidenciar que variables como la distancia, la hora del día, las condiciones climáticas y el tipo de accidente constituyen factores significativos en la predicción del tiempo de atención, lo cual abre oportunidades para futuras mejoras en la gestión operativa y en la toma de decisiones estratégicas.

- Aunque el modelo muestra un desempeño sobresaliente, se identificaron diversas limitaciones relacionadas con la distribución geográfica de los datos, la posible presencia de sesgos temporales y operativos, y la ausencia de variables clave como tráfico en tiempo real o itinerarios dinámicos de los gestores que podrían mejorar aún más la calidad del modelo. Estas limitaciones constituyen oportunidades claras para ampliar y fortalecer el sistema predictivo en investigaciones posteriores.

- El trabajo demostró que la analítica predictiva puede convertirse en un recurso estratégico para ASSISNET, aportando herramientas para optimizar la asignación de recursos, mejorar la eficiencia del servicio y elevar la satisfacción del asegurado. La adopción de modelos como el aquí propuesto representa un paso significativo hacia la transformación digital del proceso de atención de accidentes, permitiendo anticiparse a la demanda, priorizar accidentes y mejorar la capacidad de respuesta en contextos dinámicos.

- El modelo final permite estimar el tiempo de atención con un error promedio cercano a ±2.9 minutos, lo cual constituye una herramienta altamente útil para anticipar demoras, optimizar la asignación de gestores y mejorar la experiencia del asegurado mediante una atención más oportuna y previsible mejorando significativamente la experiencia del cliente.

- El pipeline construido demostró ser robusto, replicable y aplicable a problemas similares dentro del sector asegurador. Las decisiones tomadas en cada etapa desde la limpieza de datos hasta el ajuste de hiperparámetros fueron validadas empíricamente y se posicionan como buenas prácticas para proyectos futuros de analítica operativa.

- La investigación corroboró que, en dominios donde las relaciones han sido adecuadamente depuradas y estabilizadas, los modelos lineales pueden superar ampliamente a modelos complejos. El desempeño inferior de Random Forest, XGBoost y Gradient Boosting se debió a fenómenos de sobreajuste y falta de generalización, hipótesis que se fortaleció al observar la brecha entre su rendimiento en entrenamiento y su bajo desempeño en el conjunto de prueba.

- En conjunto, los resultados obtenidos no solo validan la pertinencia del modelo desarrollado, sino que también evidencian el potencial que tiene el uso de técnicas de aprendizaje automático para generar valor en la gestión de accidentes. Este trabajo constituye una base sólida para futuras investigaciones orientadas a la expansión geográfica del modelo, la incorporación de datos en tiempo real y el desarrollo de sistemas avanzados de soporte a la decisión en el sector asegurador.

- Con métricas validadas, estabilidad comprobada y un comportamiento consistente en distintos subconjuntos del dataset, el modelo Ridge Regression se encuentra en condiciones de ser integrado en sistemas operativos de gestión de accidentes. Su simplicidad computacional, interpretabilidad y alto nivel de precisión lo convierten en una opción viable y lista para despliegue en un entorno productivo.

Referencias

Ahmad, O. F., Soares, A. S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., . . . Lovat, L. B. (2019). Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The lancet. Gastroenterology & hepatology, 4*(1), 71-80. doi:10.1016/s2468-1253(18)30282-6

American Cancer Society. (3 de 10 de 2023). *Colonoscopy*. Obtenido de Cancer.org: https://www.cancer.org/cancer/diagnosis-staging/tests/endoscopy/colonoscopy.html

Calatayud, S. S.-F.-F.-A. (15 de diciembre de 2020). *Más congestión, menos tiempo de respuesta ante emergencias*. Obtenido de Más congestión, menos tiempo de respuesta ante emergencias: https://blogs.iadb.org/transporte/es/mas-congestion-menos-tiempo-de-respuesta-ante-emergencias/#:~:text=Los%20est%C3%A1ndares%20internacionales%20establecen%20que,JEMS%2C%202005

carmatec. (1 de abril de 2025). *CARMATEC*. Obtenido de CARMATEC: https://www.carmatec.com/es_mx/blog/guia-de-analisis-de-datos-en-el-sector-asegurador/# Obtenido de

Cosio, N. A. (20 de diciembre de 2021). *Métricas en regresión*. Obtenido de Medium: https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b

Hsu, C.-M., Hsu, C.-C., Hsu, Z.-M., Shih, F.-Y., Chang, M.-L., & Chen, T.-H. (2021). Colorectal polyp image detection and classification through grayscale images and deep learning. *Sensors (Basel, Switzerland), 21*(1/), 5995. doi:10.3390/s21185995

Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z. R., . . . Ogino, S. (2013). Long-term colorectal-cancer incidence and mortality after lower endoscopy. *The New England journal of medicine, 369*(12), 1095-1105. doi:10.1056/nejmoa1301969

Park, S. B., & Cha, J. M. (2022). Quality indicators in colonoscopy: the chasm between ideal and reality. *Clinical endoscopy, 55*(3), 332-338. doi:10.5946/ce.2022.037

Sánchez-Montes, C., Bernal, J., García-Rodríguez, A., Córdova, H., & Fernández-Esparrach, G. (2020). Revisión de métodos computacionales de detección y clasificación de pólipos en imagen de colonoscopia. *Gastroenterologia y hepatologia, 43*(4), 222-232. doi:10.1016/j.gastrohep.2019.11.004

Shine, R., Bui, A., & Burgess, A. (2020). Quality indicators in colonoscopy: an evolving paradigm: Quality indicators in colonoscopy. *ANZ journal of surgery, 90*(3), 215-221. doi:10.1111/ans.15775

Tobón, S. O. (01 de enero de 2025). *El Colombiano*. Obtenido de Medellín terminó 2024 con más muertos por accidentes viales en 10 años: https://www.elcolombiano.com/medellin/medellin-termino-2024-con-mas-muertos-por-accidentes-de-transito-en-10-anos-OB26273269

Tomar, N. K. (11 de 01 de 2021). *Automatic polyp segmentation using fully convolutional neural network.* Obtenido de arXiv [eess.IV]: http://arxiv.org/abs/2101.04001https://arxiv.org/abs/2101.04001

Wikipedia. (15 de septiembre de 2025). *Wikipedia*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,trees%27%20habit%20of%20overfitting%20to

Williams, J. G., Pullan, R. D., Hill, J., Horgan, P. G., Salmo, E., Buchanan, G. N., . . . Haboubi, N. (2013). Management of the malignant colorectal polyp: ACPGBI position statement. *Colorectal disease: the official journal of the Association of Coloproctology of Great Britain*

and Ireland, 15(s2), 1-38. doi:10.1111/codi.12262

World Health Organization. (11 de 07 de 2023). *Colorectal cancer*. Obtenido de Who.int: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer

Zuñiga, S. S.-F.-F. (16 de diciembre de 2020). *moviliblog*. Obtenido de moviliblog: https://blogs.iadb.org/transporte/es/mas-congestion-menos-tiempo-de-respuesta-ante-emergencias/#:~:text=Uno%20de%20los%20objetivos%20m%C3%A1s,a%20los%20centros%20de%20salud