**Logotipo

Descripción generada automáticamente**

**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  
David Manuel Villanueva Valdés

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2025

|  |  |
| --- | --- |
| **Cita** | (Londoño Londoño & Vergara Monsalve, 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. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteIX.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**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

[Resumen 8](#_Toc198842997)

[Abstract 9](#_Toc198842998)

[1. Descripción del problema 10](#_Toc198842999)

[1.1. Problema de negocio 10](#_Toc198843000)

[1.2. Aproximación desde la analítica de datos 11](#_Toc198843001)

[1.3. Origen de los datos 11](#_Toc198843002)

[1.4. Métricas de desempeño 12](#_Toc198843003)

[2. Objetivos 14](#_Toc198843004)

[2.1. Objetivo general 14](#_Toc198843005)

[2.2. Objetivos específicos 14](#_Toc198843006)

[3. Datos 15](#_Toc198843007)

[3.1. Datos originales 15](#_Toc198843008)

[3.2. Datsets 18](#_Toc198843009)

[3.3. Analítica descriptiva 22](#_Toc198843010)

[Referencias 34](#_Toc198843011)

**Lista de tablas**

[**Tabla 1.** Principales Variables del DataSet 16](#_Toc198841751)

[**Tabla 2.** Caracterización de las columnas 17](#_Toc198841752)

**Lista de figuras**

[**Figura 1.** BOXPLOT con Outlier 18](#_Toc198841592)

[**Figura 2**. Rango inter-cuartil (IQR) 19](#_Toc198841593)

[**Figura 3.** Reemplazar atípicos por la media 19](#_Toc198841594)

[**Figura 4.** Distribución de los tiempos después del tratamiento de atípicos 20](#_Toc198841595)

[**Figura 5**. Imputación de valores faltantes 21](#_Toc198841596)

[**Figura 6.** Listado de instancias 21](#_Toc198841597)

[**Figura 7**. Listado de municipios 22](#_Toc198841598)

[**Figura 8.**Listado de aseguradoras 22](#_Toc198841599)

[**Figura 9.** Listado de acuerdos 22](#_Toc198841600)

[**Figura 10.** Porcentaje de datos Null 23](#_Toc198841601)

[**Figura 11.** Valores nulos por columna 24](#_Toc198841602)

[**Figura 12.** Estadísticas Descriptivas 26](#_Toc198841603)

[**Figura 13.** Estadísticas Descriptivas de la variable DiferenciaHoras 26](#_Toc198841604)

[**Figura 14**. Distribución Normal TiempoAtención 27](#_Toc198841605)

[**Figura 15.** Variables categóricas 29](#_Toc198841606)

[**Figura 16.** Frecuencia de instancias variables categóricas 29](#_Toc198841607)

[**Figura 17**.Relación entre Instancia y Acuerdo 31](#_Toc198841608)

[**Figura 18**. Relación entre las variables numéricas 32](#_Toc198841609)

[**Figura 19**. BOXPLOT con Outlier 34](#_Toc198841610)

[**Figura 20**. Atención de incidencias por mes 34](#_Toc198841611)

**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

**NPS** Net Promoter Score (Métrica que mide la satisfacción del cliente)

**UdeA** Universidad de Antioquia

# Resumen

El presente proyecto tiene como objetivo desarrollar un modelo predictivo que estime el tiempo de espera de un asegurado tras reportar un accidente de tránsito, con el fin de optimizar los procesos de atención y mejorar la experiencia del cliente en una compañía aseguradora. La iniciativa surge de la necesidad de contar con herramientas analíticas que permitan anticipar demoras y gestionar de forma eficiente los recursos disponibles en campo, como gestores de siniestros y unidades móviles.

Para ello, se utilizó un conjunto de datos históricos de siniestros, registrados por la aseguradora a través de sus sistemas de gestión durante un período determinado. Estos datos incluyen variables relacionadas con la ubicación, hora del incidente, tipo de accidente, tráfico, disponibilidad de personal y otros factores relevantes.

Se implementó un proceso iterativo que incluyó limpieza, análisis exploratorio, selección de variables y entrenamiento de distintos modelos de machine learning como regresión lineal, árboles de decisión y random forest, evaluando su desempeño mediante métricas como el MAE y el RMSE. Entre los principales obstáculos se encontraron problemas de calidad de datos, presencia de valores atípicos y desequilibrios en la distribución de los tiempos de atención.

Los resultados obtenidos muestran que el modelo tiene un desempeño aceptable para predecir tiempos de espera con un margen de error razonable, y representa una herramienta valiosa para la toma de decisiones operativas en tiempo real.

[Sigue este enlace para ver el repositorio del proyecto.](https://github.com/ferdilo04/Datos_Analitica/tree/master/Monografia_2025)

*Palabras clave*: Modelos predictivos, Tiempo de espera, Accidentes de tránsito, Seguros, Aprendizaje automático, Analítica de datos, Gestión de siniestros, Optimización de recursos, Experiencia del asegurado, Predicción operativa

# Abstract

This project aims to develop a predictive model to estimate the waiting time of an insured person after reporting a traffic accident, with the goal of optimizing service response and improving customer experience in an insurance company. The initiative arises from the need for analytical tools that allow the company to anticipate delays and efficiently allocate field resources, such as claims adjusters and mobile units.

To achieve this, a dataset composed of historical traffic incident records collected by the insurer over a defined period was used. The data includes variables such as accident location, time of the incident, type of accident, traffic conditions, and resource availability.

An iterative process was followed, including data cleaning, exploratory analysis, feature selection, and training of various machine learning models, such as linear regression, decision trees, and random forest. These models were evaluated using metrics like MAE and RMSE. Some of the main challenges encountered included data quality issues, the presence of outliers, and an imbalanced distribution of waiting times.

The results show that the model performs reasonably well in predicting waiting times within an acceptable error margin. This predictive tool has the potential to significantly support real-time operational decision-making and improve overall service delivery in traffic accident scenarios.

*Keywords***:**   
Predictive Modeling, Waiting Time, Traffic Accidents, Insurance, Machine Learning, Operational Optimization, Claims Management, Customer Experience, Data Analytics, Resource Allocation

# Descripción del problema

En el contexto de los accidentes de tránsito, uno de los principales factores que impacta en la satisfacción de los asegurados es el tiempo de espera para recibir asistencia por parte de su compañía de seguros. Este tiempo puede variar significativamente según diversos factores como la ubicación del siniestro, la disponibilidad de gestores, la hora del día, el tipo de accidente, entre otros.

Actualmente, muchas aseguradoras enfrentan dificultades para estimar de forma precisa el tiempo de atención que tomará resolver un incidente, lo cual impide una asignación eficiente de recursos y afecta negativamente la experiencia del cliente. Además, la falta de una predicción adecuada puede provocar demoras operativas, sobrecarga de gestores o uso ineficiente de unidades móviles.

Frente a esta problemática, surge la necesidad de desarrollar un modelo predictivo que permita estimar, con base en información histórica y en tiempo real, el tiempo de espera estimado que tendrá un asegurado tras reportar un accidente de tránsito. Esta estimación no solo permitirá optimizar la asignación de recursos humanos y logísticos, sino que también contribuirá a mejorar los niveles de servicio, reducir costos operativos y aumentar la confianza del cliente en la aseguradora.

Por lo tanto, esta investigación busca aplicar técnicas de aprendizaje automático (Machine Learning) para construir un modelo capaz de predecir con precisión dicho tiempo de espera, facilitando la toma de decisiones proactiva por parte de las aseguradoras.

## Problema de negocio

En el sector asegurador, especialmente en el ámbito de los seguros vehiculares, la capacidad de respuesta ante incidentes de tránsito es un factor determinante en la percepción del cliente sobre la calidad del servicio. Las aseguradoras enfrentan el desafío constante de coordinar de manera eficiente la atención en campo, gestionando la disponibilidad de gestores de siniestros, grúas, unidades móviles y otros recursos necesarios para asistir al asegurado.

Actualmente, la falta de herramientas predictivas dificulta la planificación operativa y la asignación óptima de estos recursos. Como resultado, se presentan tiempos de espera prolongados o inconsistentes, generando insatisfacción en los asegurados, pérdida de confianza y un aumento en los reclamos. Además, estas ineficiencias pueden traducirse en costos operativos elevados, uso inadecuado del personal en campo, y riesgos de saturación en momentos de alta demanda (por ejemplo, durante horarios pico o condiciones climáticas adversas).

La empresa AssisNet busca entonces mejorar sus procesos de gestión de siniestros mediante la implementación de un sistema que le permita predecir con anticipación el tiempo estimado de atención de un asegurado, a partir de variables como la localización, hora del incidente, tipo de accidente, historial de eventos similares, entre otros. Esta capacidad predictiva permitirá anticipar demoras, redistribuir recursos en tiempo real, y ofrecer al cliente información más precisa y confiable sobre el tiempo de espera.

## Aproximación desde la analítica de datos

Para abordar el problema del tiempo de espera en la atención de asegurados durante accidentes de tránsito, se plantea una solución basada en técnicas de analítica de datos y modelos predictivos de machine learning (ML). Esta aproximación permitirá transformar los datos históricos de siniestros en una herramienta estratégica para la toma de decisiones operativas en tiempo real.

El objetivo principal es desarrollar un modelo de predicción que estime con precisión el tiempo de espera que enfrentará un asegurado desde el momento en que reporta un accidente hasta que recibe atención en el lugar del siniestro. Este modelo se entrenará utilizando variables históricas como: Ubicación geográfica del accidente, hora y día del evento, condiciones climáticas, tipo de accidente, nivel de tráfico en la zona, disponibilidad de gestores y unidades móviles, información demográfica del asegurado, e historial de atención en eventos similares

Mediante esta aproximación, la aseguradora podrá anticipar con mayor exactitud posibles demoras, optimizar la distribución de recursos, mejorar la experiencia del cliente al proporcionar tiempos estimados de atención realistas, y reducir los costos operativos asociados con una mala planificación.

## Origen de los datos

Los datos utilizados en este proyecto provienen de registros históricos recopilados por la compañía aseguradora AssisNet que brinda servicios de atención en incidentes de tránsito. Esta información ha sido recolectada de manera sistemática a través de sus plataformas digitales de gestión de siniestros, aplicaciones móviles utilizadas por los asegurados.

Los registros cubren un período de tiempo representativo y comprenden eventos reales de accidentes de tránsito donde el gestor de servicio prestó servicios de asistencia. Los datos fueron obtenidos bajo condiciones operativas normales, abarcando distintos tipos de siniestros, zonas geográficas, franjas horarias y niveles de complejidad en la atención. La recolección se realizó en cumplimiento con políticas internas de calidad de datos y estándares de confidencialidad, sin incluir información que permita identificar de manera directa a los asegurados.

Este conjunto de datos constituye la base sobre la cual se entrenará y validará el modelo predictivo, permitiendo identificar patrones relevantes y variables significativas que influyen en el tiempo de espera del asegurado.

## Métricas de desempeño

En el contexto de un modelo de predicción del tiempo de espera de un asegurado en un accidente de tránsito, es fundamental evaluar tanto las métricas de desempeño del modelo de machine learning como su impacto en el negocio. Desde el punto de vista técnico, al tratarse de un problema de regresión, las métricas más relevantes son el MAE (error absoluto medio), el RMSE (raíz del error cuadrático medio) y el coeficiente de determinación R². Un modelo con un MAE menor o igual a 10 minutos, un RMSE menor o igual a 15 minutos y un R² igual o superior a 0.85 puede considerarse técnicamente sólido, ya que ofrece predicciones suficientemente precisas para la toma de decisiones operativas.

En cuanto a las métricas de negocio, el modelo puede contribuir significativamente a reducir el tiempo promedio de espera de los asegurados, mejorar la satisfacción del cliente, disminuir el número de reclamos por demoras y optimizar los costos operativos mediante una mejor asignación de recursos logísticos.

Se puede establecer que el modelo es viable siempre que cumpla con los umbrales mínimos de MAE ≤ 10 minutos y R² ≥ 0.85, ya que a partir de ese punto se alcanza un ROI suficiente para cubrir los costos de operación y generar valor para la organización.

* Disminuir el tiempo de espera del asegurado entre un 10% a 30% en relación con el tiempo de espera actual aproximado de 52 minutos sin la implementación del modelo.
* La ejecución del modelo se hará en línea por medio de un Api expuesta para el uso exclusivo de la aplicación de asistencias de AsisNet, y se espera un tiempo de respuesta no mayor a 2 segundos en condiciones normales, es decir, que no se presenten inconvenientes de conectividad.
* Según estudios de consultoras como McKinsey y Deloitte, reducciones del 20-30% en tiempos de espera están correlacionadas con mejoras del 5-15% en satisfacción del cliente (NPS) y disminuciones del 3-8% en abandono de clientes.

# Objetivos

## Objetivo general

Predecir el tiempo de espera en la atención de un asegurado durante una incidencia o accidente de tránsito desarrollando un modelo de Machine Learning con el fin de facilitar la asignación adecuada del gestor de seguros.

## Objetivos específicos

1. Recolectar y preprocesar datos históricos relevantes sobre incidentes de tránsito, incluyendo variables como tipo de accidente, ubicación, hora del evento, tiempo de respuesta, criticidad del caso y perfil del asegurado.
2. Identificar y seleccionar las variables más significativas que influyen en el tiempo de espera, mediante técnicas de análisis exploratorio de datos (EDA) y selección de características.
3. Diseñar y entrenar modelos de Machine Learning supervisados, comparando distintos algoritmos (como regresión lineal, árboles de decisión, random forest o XGBoost), para predecir el tiempo de espera en función de las condiciones del accidente.
4. Evaluar el rendimiento de los modelos generados, con la metodología de evaluación de Cross Validation, utilizando métricas como MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) y R², con el fin de seleccionar el modelo más preciso y eficiente.
5. Implementar un sistema de recomendación basado en el modelo predictivo, que sugiera la asignación óptima del gestor de seguros y determine si se requiere o no desplazamiento al lugar del incidente.
6. Proponer y generar un documento rfc con los lineamientos para la integración del modelo en los procesos operativos de la aseguradora, orientados a mejorar la eficiencia en la atención de siniestros y la satisfacción del cliente.

# Datos

## Datos originales

Para llevar a cabo el entrenamiento del modelo, capaz de predecir el tiempo de espera en la atención de un asegurado durante un accidente de tránsito, se realizó un análisis exploratorio sobre el conjunto de datos proporcionado por la empresa AsisNet, que contiene información detallada de accidentes de tránsito atendidos por los agentes de la empresa, a partir del mes de abril del 2016 al 11 de mayo del 2025.

El DataSet tiene el nombre DataSet\_Aseguradora.xlsx y se encuentra en formato de excel con extensión .xlsx, con un tamaño de 3.330 Kb el cual contiene 14520 registros y 28 variables, la información del dataset fue obtenida mediante una consulta Transact SQL en un servidor de base de datos SQL Server, que contiene información de las incidencias de tránsito y de las gestiones de atención realizadas por los agentes de las aseguradoras suscritas a la empresa AssisNet .

Para nuestro objetivo inferido que es modelar el tiempo de espera mínimo para el desplazamiento de un agente al lugar del accidente, haremos uso de la siguiente fórmula: tiempo\_espera\_min = fechallegada - fechasiniestro.

Con el fin de construir un conjunto de datos robusto, limpio y relevante para alimentar el modelo procedemos a identificar las variables principales del DataSet

| **Categoría** | **Variables relevantes** | **Descripción** |
| --- | --- | --- |
| **Fechas** | fechasiniestro, fechallegada | Fecha y hora del siniestro y de llegada del gestor al lugar |
| **Localización** | municipio, departamento | Ubicación del incidente |
| **Datos asegurado** | nombreconductor, telefonoasegurado, correo | Información personal (parcialmente anonimizada) |
| **Aseguradora** | Aseguradora | Nombre de la entidad aseguradora |
| **Atención** | tiempo\_espera\_min, diasemana, tipoatencion, instancia | Detalles del tiempo y tipo de atención brindada |
| **Otros campos** | tipovehiculo, ResultadoFallo, tipoevento | Información adicional sobre el vehículo, el siniestro y su resolución |

**Tabla 1.** Principales Variables del DataSet

A continuación, indicamos la caracterización de las columnas del dataset

|  |  |  |
| --- | --- | --- |
| **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 la fecha de reporte de accidente y la fecha de llegada al sitio |  |
| HoraPicoTarde | Si el accidente ocurrió en la hora pico de la tarde (de 5 a 7 Pm) valores (Si, No) | Si |
| HoraPicoManana | Si el accidente ocurrió en la hora pico de la mañana (de 6:30 a 8:30 Am) valores (Si, No) | No |
| InicioNoche | Si el accidente ocurrió entre las 7 de la noche y las 11:59 pm valores (Si, No) | No |
| Amanecer | Si el accidente ocurrió entre las 12 m y las 5:59 am valores (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 se llega entre las partes de los implicados | Transito |
| UsuarioRegistra | Usuario que registra el caso en el sistema | Carolina Garcia Valencia |
| LiberaVehiculo | Si el vehículo fue retenido por el tránsito y después fue 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 |

**Tabla 2.** Caracterización de las columnas

## Datsets

A partir de los datos originales, se construyó un conjunto de datos más específicos para el modelo de machine learning, se aplicaron varios pasos de preprocesamiento a las variables seleccionadas que aplicarán para el modelo:

* + - 1. **Manejo de valores atípicos:** En el análisis de los datos se encontró un aproximado de un 15% de valores atípicos en la variable **TiempoAtencion** como se puede ver en la siguiente imagen.

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

**Figura 1.** BOXPLOT con Outlier

Usamos el método del rango Inter cuartil (IQR) con el 25% en el primer cuartil y el 75% en el tercer cuartil para identificar los valores atípicos

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Figura 2**. Rango inter-cuartil (IQR)

Luego los valores atípicos que identificamos fueron reemplazados por el promedio de los datos de la columna **TiempoAtencion** sin tener en cuenta los valores atípicos en el cálculo de este promedio.

Texto

Descripción generada automáticamente

**Figura 3.** Reemplazar atípicos por la media

Llegando finalmente al siguiente resultado donde podemos evidenciar que los valores presentan una mejor distribución con respecto al tiempo de atención

Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamente

**Figura 4.** Distribución de los tiempos después del tratamiento de atípicos

* + - 1. **Imputación de valores faltantes:** 
         1. En la variable **TiempoAtencion** se encontraron varios valores faltantes los cuales fueron imputados mediante un valor aleatorio entre el valor mínimo y máximo de la columna **TiempoAtencion**, con el fin de mantener una base de datos completa y reducir el sesgo.
         2. La variable **fechallegada** fue calculada con el valor de la variable de **fecha** más la suma en minutos del valor calculado previamente en la variable **TiempoAtencion.**
         3. Las variables de **annoatencion, mesantencion, diaatencion, horaantencion, diasemana** fueron generados desde la variable **fechallegada** calculada previamente.

Texto, Escala de tiempo

Descripción generada automáticamente

**Figura 5**. Imputación de valores faltantes

* + - 1. **Agrupación de valores:**
         1. En la variable **Intsancia** se agruparon los datos con cantidades inferiores a 500 asistencias llegando finalmente a este resultado.

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

**Figura 6.** Listado de instancias

* + - * 1. En la variable de **Municipio** se agruparon los datos con cantidades inferiores a 350 asistencias llegando finalmente a este resultado.

Tabla

Descripción generada automáticamente

**Figura 7**. Listado de municipios

* + - * 1. En la variable de **Aseguradora** se agruparon los datos con cantidades inferiores a 350 asistencias llegando finalmente a este resultado.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Figura 8.**Listado de aseguradoras

* + - * 1. En la variable de **Acuerdo** se agruparon los datos con cantidades inferiores a 1000 asistencias llegando finalmente a este resultado.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Figura 9.** Listado de acuerdos

Columnas innecesarias para el modelo: se eliminaron las columnas que según los análisis de los datos no se consideran que sean útiles para modelo que se va a desarrollar como son: 'Abogadounico', 'ResultadoFallo', 'LiberaVehiculo', 'UsuarioRegistra', 'idproceso', 'placa'

## Analítica descriptiva

En nuestro dataset contamos con diferentes tipos de datos, float64(5), int64(5), object(18), observamos que predominan datos tipo Object , los cuales requieren tratamiento por ejemplo, campos de tipo fecha como **fecha** y **fechallegada** requieren conversión a tipo de dato DateTime.

Haciendo un análisis de los valores nulos, encontramos que las variables **ResultadoFallo**, **LiberaVehiculo** y **clienteimportante** tienen gran cantidad de valores de Null por lo tanto es poco útil mantenerlas en el análisis o modelo. la variable **Acuerdo** también se presenta gran cantidad de nulos, durante el análisis evaluaremos si procedemos a eliminarla del conjunto de datos ya que no aportaría mucho valor; las variables **diaatencion, mesatencion, annoatencion, fechallegada, Horaatencion, TiempoAtencion** presentan una cantidad moderada de valores nulos, lo que nos sugiere posiblemente una imputación con valores representativos como la moda o la mediana.

Imagen que contiene Texto

Descripción generada automáticamente

**Figura 10.** Porcentaje de datos Null

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Figura 11.** Valores nulos por columna

A continuación, se presenta un análisis de las variables con presencia significativa de valores nulos en el conjunto de datos:

Las columnas ResultadoFallo y LiberaVehiculo presentan un nivel de vacíos cercano al 99%, lo cual indica que casi no contienen información útil. Por tanto, podrían eliminarse si no son variables esenciales para el análisis. La variable clienteimportante también muestra un porcentaje elevado de valores nulos (~97%), y dado que muy pocos registros están marcados como "importantes", podría considerarse su transformación en una variable binaria o incluso su eliminación si no aporta valor predictivo.

La columna Acuerdo tiene aproximadamente un 60% de datos faltantes. Dado este porcentaje, es recomendable revisar si los valores presentes contienen información clave antes de decidir su conservación. Por otro lado, las variables diaatencion, mesatencion, annoatencion, fechallegada, Horaatencion y TiempoAtencion presentan alrededor de un 10% de valores ausentes. Esta ausencia parece estar relacionada con registros en los que no se brindó atención, por lo cual se debe analizar el contexto de estos casos antes de imputar valores o descartarlos.

La variable Abogadounico presenta entre un 7 y 8% de registros nulos, posiblemente relacionados con procesos sin abogado asignado. Finalmente, Aseguradora muestra una proporción mínima de valores faltantes (alrededor del 1%), lo que sugiere que estos podrían imputarse sin afectar significativamente la calidad del conjunto de datos.

* + 1. **Estadísticas Descriptivas**

A partir del análisis de las estadísticas descriptivas, se pueden extraer varias conclusiones relevantes sobre las variables incluidas en el conjunto de datos.

Las variables idproceso, annoregistro, mesregistro, diaregistro y horaregistro, relacionadas con el momento del registro del siniestro, están completas con 14,519 observaciones cada una.

El mes y el día de registro (mesregistro y diaregistro) se distribuyen de manera uniforme a lo largo del año, con medianas de 7 y 16, respectivamente. La hora de registro (horaregistro) tiene una media cercana a las 14 horas, con valores mínimos y máximos entre las 0 y las 23 horas, lo que sugiere una distribución regular a lo largo del día.

En cuanto a las variables asociadas al momento de la atención (annoatencion, mesatencion, diaatencion, Horaantencion), estas tienen alrededor de 12,955 observaciones, lo que refleja que aproximadamente el 10% de los procesos registrados no llegaron a ser atendidos. Las distribuciones de año, mes, día y hora de atención son similares a las de registro, lo que indica cierta coherencia temporal entre ambas etapas del proceso.

Por su parte, la variable **TiempoAtencion** muestra un comportamiento particular. Aunque su media es de aproximadamente 4,090 minutos (unas 68 horas), presenta una desviación estándar extremadamente alta (alrededor de 40,000 minutos), y un valor máximo que supera los 1.7 millones de minutos (más de 3 años), lo que sugiere la existencia de valores atípicos extremos. La mediana del tiempo de atención es de solo 28 minutos, y el 75% de los casos se resuelven en menos de 230 minutos (menos de 4 horas), lo que refuerza la hipótesis de una distribución altamente asimétrica. Este hallazgo es clave, ya que indica la necesidad de aplicar técnicas de tratamiento de **outliers** o transformación de datos para que la variable sea útil en modelos predictivos.

En conclusión, las variables temporales se comportan de forma esperada y coherente, aunque la variable **TiempoAtencion** requiere especial atención debido a su alta variabilidad y presencia de valores extremos. Además, se evidencia una pérdida de información en los campos relacionados con la atención que serán consideradas en las etapas de limpieza y modelado.

Tabla

Descripción generada automáticamente

**Figura 12.** Estadísticas Descriptivas

**Texto

Descripción generada automáticamente con confianza media**

**Figura 13.** Estadísticas Descriptivas de la variable DiferenciaHoras

* + 1. **Validación de la normalidad**

Como parte del análisis exploratorio, se evaluó la normalidad de las distribuciones de la variable **TiempoAtencion**, tanto visualmente mediante gráficos Q-Q (ver Figura 5), como estadísticamente mediante las pruebas de Shapiro-Wilk y Kolmogorov-Smirnov.

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

**Figura 14**. Distribución Normal TiempoAtención

El gráfico Q-Q (Quantile-Quantile) correspondiente a la variable **TiempoAtencion** permite evaluar visualmente la adecuación de su distribución con respecto a una distribución normal teórica.

Al analizar la gráfica, se observa una marcada desviación de los puntos respecto a esta línea, especialmente en los extremos, lo que indica la presencia de asimetría y valores atípicos. En particular, el extremo superior derecho muestra una curvatura ascendente pronunciada, lo que evidencia una fuerte asimetría positiva (sesgo a la derecha) y la existencia de valores extremos elevados.

Este comportamiento sugiere que la variable **TiempoAtencion** no sigue una distribución normal.

Se sugiere analizar y tratar los outliers detectados, ya que pueden influir negativamente en el desempeño de los modelos predictivos.

* + 1. **Identificación valores únicos**

El conjunto de datos incluye diversas variables relacionadas con identificadores, fechas, tiempos y ubicación. Entre los identificadores y fechas, se encuentra idproceso, que representa un identificador único para cada caso o proceso, sin repeticiones en los 14,519 registros. La variable placa corresponde a la placa del vehículo, con 13,245 valores, lo que indica que algunos vehículos aparecen en más de un caso. La variable fecha, también con 14,519 registros únicos, probablemente representa el momento en que ocurrió el evento asociado al proceso. Por su parte, fechallegada, con 12,955 registros, indica la fecha de llegada del vehículo, aunque puede haber valores repetidos o ausentes.

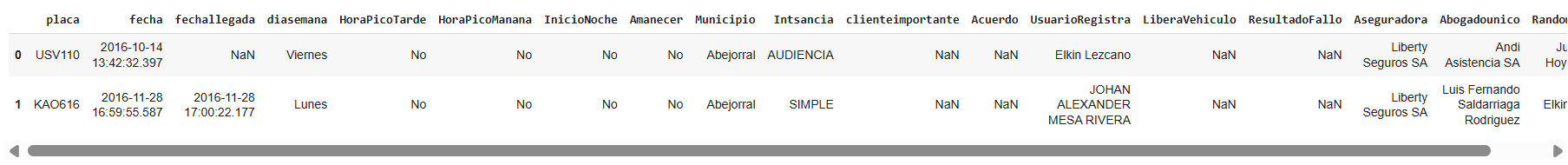
En cuanto a las variables temporales adicionales, TiempoAtencion, con 2,659 registros, parece indicar el tiempo que se tardó en atender un caso; esta variable presenta alta variación, por lo que podría tratarse como continua o categorizarse en rangos. Adicionalmente, las variables HoraPicoTarde, HoraPicoManana, InicioNoche y Amanecer, con dos categorías cada una (probablemente valores binarios 0 o 1), indican si el evento ocurrió dentro de determinadas franjas horarias específicas.

Finalmente, en lo relacionado con la ubicación, la variable Municipio contiene registros correspondientes a 116 municipios distintos, lo que la convierte en una variable categórica clave para realizar análisis geográficos y entender la distribución espacial de los casos.

* + 1. **Variables Categóricas**

En el análisis de las variables categóricas se identifica que hay presencia de datos faltantes y presentan valores nulos, con lo que podemos llegar a un posible descarte en nuestro análisis **fechallegada, clienteimportante,Acuerdo, LiberaVehiculo y ResultadoFallo,** se puede presentar datos faltantes debido a que se tienen datos en varias columnas tienen NaN, posibles outliers en las columnas **fechallegada** y **Aseguradora**.

De acuerdo con la revisión anterior Procedimos a eliminar las columnas 'Abogadounico', 'ResultadoFallo', 'LiberaVehiculo', 'UsuarioRegistra', 'idproceso', 'placa' no aportan valor en el análisis incial de los datos y convertir las variables fecha a tipo datatime

****

**Figura 15.** Variables categóricas

**Diagrama, Esquemático

Descripción generada automáticamente**

**Figura 16.** Frecuencia de instancias variables categóricas

Después de un análisis de frecuencia de variables categóricas podemos concluir que la información de las columnas **acuerdo, aseguradora, instancia y municipios** están un poco desbalanceada entre sus datos y es necesario hacer agrupaciones.

Debemos realizar algunas agrupaciones de columnas para tener mejor comprensión de los datos como son:

* Agrupar columna **horapicomañana con horapicotarde**
* Agrupar columna **inicionoche con amanecer**

Las columnas de **Abogadounico, LiberaVehiculo y Resultadofallo** las podemos borrar por la cantidad tan baja de datos

En la **gráfica de Distribución de Casos por Día de la Semana** identificamos que los martes y viernes suelen tener más incidentes de tránsito. Los fines de semana (sábado y domingo) tienden a mostrar menos eventos.

Esta tendencia puede relacionarse con los patrones de tráfico y movilidad urbana, posiblemente haya más flujo vehicular entre semana, especialmente al inicio y al cierre.

En la **gráfica de Las aseguradoras** se identifica que Liberty Seguros S.A y IKE ASIST concentran una cantidad significativamente mayor de incidentes reportados. Esto puede deberse a mayor cuota de mercado, problemas estructurales en ciertos procesos internos.

**Consideramos pertinente** hacer análisis personalizados por aseguradora para entender su comportamiento y ajustar estrategias de servicio.

En la interpretación de la gráfica de Instancia y Acuerdo vemos hay relación entre el acuerdo transito con la instancia Simple, parece ser que las atenciones simples requieren alguna asistencia del tránsito. Para validar el anterior hallazgo, procedimos a hallar la correlación entre estas dos variables, como se observa en la imagen 9.

Con la gráfica podemos evidenciar que a diferencia de lo que uno pensaría que en instancias simples no se requiere tránsito de las 8900 de asistencias simples hay unas 2500 que se hizo finalmente la petición del transito

Gráfico

Descripción generada automáticamente

**Figura 17**.Relación entre Instancia y Acuerdo

* + 1. **Variables numéricas**

Realizamos la gráfica de distribución y dispersión para las variables numéricas tal como se muestra a continuación

**Diagrama

Descripción generada automáticamente**

**Figura 18**. Relación entre las variables numéricas

En la gráfica anterior identificamos la distribución de cada variable numérica individual, identificadas en color verde, mediante la curva de densidad, fuera de la diagonal encontramos el diagrama de dispersión entre dos variables, lo que nos permite identificar relación lineal entre ellas.

Como parte del análisis exploratorio de datos, se elaboró una matriz de pares (pairplot) con el fin de evaluar visualmente la relación entre las variables numéricas del conjunto de datos. Esta representación gráfica permite observar tanto la distribución individual de cada variable (ubicadas en la diagonal de la matriz) como la relación bivariada entre pares de variables mediante diagramas de dispersión.

Uno de los principales hallazgos corresponde a una alta correlación entre las variables relacionadas con fechas. Específicamente, se evidencian patrones lineales entre annoregistro y annoatencion, mesregistro y mesatencion, así como entre diaregistro y diaatencion. Esta asociación sugiere que la atención a los eventos tiende a ocurrir el mismo día o en fechas muy próximas al registro inicial. Desde la perspectiva del modelado estadístico, esta redundancia puede resultar problemática, ya que puede introducir multicolinealidad. Por ello, se recomienda considerar la eliminación de una de las variables en cada par fuertemente correlacionado en etapas posteriores del análisis.

En contraste, la variable **TiempoAtencion** no presenta relaciones lineales claras con el resto de las variables numéricas. Los diagramas de dispersión muestran una distribución altamente dispersa, en particular frente a **horaregistro** y **Horaatencion**, lo cual sugiere una escasa o nula correlación lineal. Esta observación indica que el tiempo de atención podría estar más influenciado por variables categóricas o por factores contextuales no reflejados en las variables cuantitativas actuales.

Por último, la distribución de **TiempoAtencion** presenta una fuerte asimetría positiva, con una alta concentración de valores en el rango bajo, acompañada por la presencia de valores extremos (outliers) significativamente elevados. Esta característica sugiere la necesidad de aplicar estrategias para el tratamiento de valores atípicos y, eventualmente, transformar la variable para mejorar su normalidad antes de utilizarla en modelos de predicción que asuman distribuciones simétricas o gaussianas.

* + 1. **Hallazgo Valores atípicos**

En este apartado identificamos cuántos registros tenemos con tiempos de atención superior a 1 hora, donde encontramos que se tiene 1005 casos de la cantidad de 14519 del total. Existen muchos **outliers extremos**, que superan por mucho el rango intercuartílico

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

**Figura 19**. BOXPLOT con Outlier

* + 1. **Análisis variables fechas**

En la gráfica **“Atención de incidencias por mes”** observamos que entre el mes de marzo del 2024 y marzo de este año no hay datos. En validación de esta información con los abogados de la empresa AsisNet nos indican que en sus servidores se presentó durante el mes de febrero del año 2025 un ataque informático y la información de esas fechas no se logró recuperar, por lo tanto no se tiene información en este rango de tiempo.

**Gráfico

Descripción generada automáticamente**

**Figura 20**. Atención de incidencias por mes

# Referencias

Colombia, C. d. (2002, septiembre 13). *Diario Oficial No. 44.932, de 13 de septiembre de 2002* . From Código de Tránsito de Colombia: https://www.medellin.gov.co/es/wp-content/uploads/2023/05/Codigo-nacional-de-transito.pdf

Wang, H. N. (2023). *Building a machine learning-based ambulance dispatch triage model for emergency medical services*. From Health Data Science: https://spj.science.org/doi/10.34133/hds.0008

Zhang, G. J. (2020). *An Analytical Approach to Predict Auto Insurance*. From https://www.ijisrt.com/assets/upload/files/IJISRT24JUL1269.pdf