Informe ejecutivo: predicción de accidentes de tráfico en Madrid (2012-2018)

Analista: Pablo Santilli

Año: 2025

Índice:

1. Introducción

Este informe presenta un análisis exhaustivo de los datos de accidentes de tráfico en la ciudad de Madrid, abarcando el periodo comprendido entre 2012 y 2018. La información utilizada para este análisis proviene del portal de datos abiertos del Ayuntamiento de Madrid (https://visualizadatos.madrid.es/pages/accidentes-de-trafico). Es importante destacar que la base de datos registra un registro por cada persona implicada en un accidente, y para el periodo en estudio (2012-2018), solo se incluyen los accidentes que resultaron en heridos o daños al patrimonio municipal.

El objetivo principal de este análisis es desarrollar un modelo predictivo que permita estimar el número de accidentes de tráfico que ocurrirán en la ciudad en meses futuros. Para lograr este objetivo, se aplicarán técnicas de análisis de datos, incluyendo la limpieza, transformación y análisis exploratorio de los datos históricos, así como la investigación y selección de modelos de predicción adecuados.

2. Carga y tratamiento de datos

En esta etapa inicial del proyecto, se procedió a la carga y tratamiento de los datos de accidentes de tráfico en Madrid. Se utilizaron archivos en formato .csv que contenían la información recopilada entre 2012 y 2018. La carga de los datos se realizó mediante el lenguaje de programación Python, utilizando librerías especializadas para el manejo de datos.

El conjunto de datos presentaba valores nulos solo en la variable N°, que representa el número de calle o ruta en la cual se produjo el accidente. Se modificaron estos valores nulos por 0. Las demás variables no presentaban valores nulos, lo que facilitó el proceso de limpieza. Se realizó una revisión exhaustiva de las variables para asegurar la calidad de la información. Las variables disponibles en el conjunto de datos se clasificaron en las siguientes categorías:

- Variables temporales:FECHA, RANGO HORARIO, DIA SEMANA
- Variables geográficas: DISTRITO ,LUGAR ACCIDENTE
- Variables relacionadas con el accidente: TIPO ACCIDENTE, Nº VICTIMAS *, LESIVIDAD
- Variables relacionadas con las condiciones: CPFA Granizo, CPFA Hielo, CPFA Lluvia, CPSV Mojada, CPSV Aceite, CPSV Barro
- Variables relacionadas con las personas involucradas: TIPO PERSONA, SEXO, Tramo Edad, Tipo Vehiculo
- Consulta realizada correctamente

3. Analítica descriptiva

En esta fase, se realizó un análisis exploratorio de los datos con el objetivo de obtener una visión general de los mismos y comprender las relaciones entre las variables. Para ello, se utilizaron diferentes herramientas de visualización, como histogramas, gráficos de barras y tablas de contingencia.

Visualización de Variables Individuales:

- Se utilizaron histogramas para mostrar la distribución de variables individuales como Nº VICTIMAS * y FECHA.
- Se emplearon gráficos de barras para comparar la frecuencia de accidentes en diferentes categorías, como DISTRITO, TIPO ACCIDENTE, Tipo Vehiculo y RANGO HORARIO.

- Se construyeron tablas de contingencia para observar las relaciones entre pares de variables, como SEXO vs. TIPO ACCIDENTE y DIA SEMANA vs. TIPO ACCIDENTE.

Hallazgos principales:

- Patrones temporales:
  - Los viernes son los días con mayor número de accidentes, mientras que los domingos registran la menor cantidad.
  - El rango horario de 14:00 a 15:00 horas presenta la mayor frecuencia de accidentes.
- Distribución geográfica:
  - El distrito de Salamanca registra el mayor número de accidentes, seguido de Chamartín y Centro.
- Tipología de accidentes:
  - La colisión doble es el tipo de accidente más frecuente, representando más de la mitad de los casos.
- Perfil de las personas involucradas:
  - Los hombres son los principales conductores implicados en accidentes.
  - La mayoría de los accidentes ocurren en condiciones de pavimento seco.

Gráficos relevantes:

Considerando la importancia de la visualización para la comprensión de los datos, se destacan tres gráficos como los más relevantes en este análisis:

Gráfico de barras de accidentes por día de la semana: Este gráfico permite observar claramente la diferencia en la frecuencia de accidentes entre los distintos días de la semana, destacando los viernes como el día con mayor número de accidentes y los domingos con el menor.

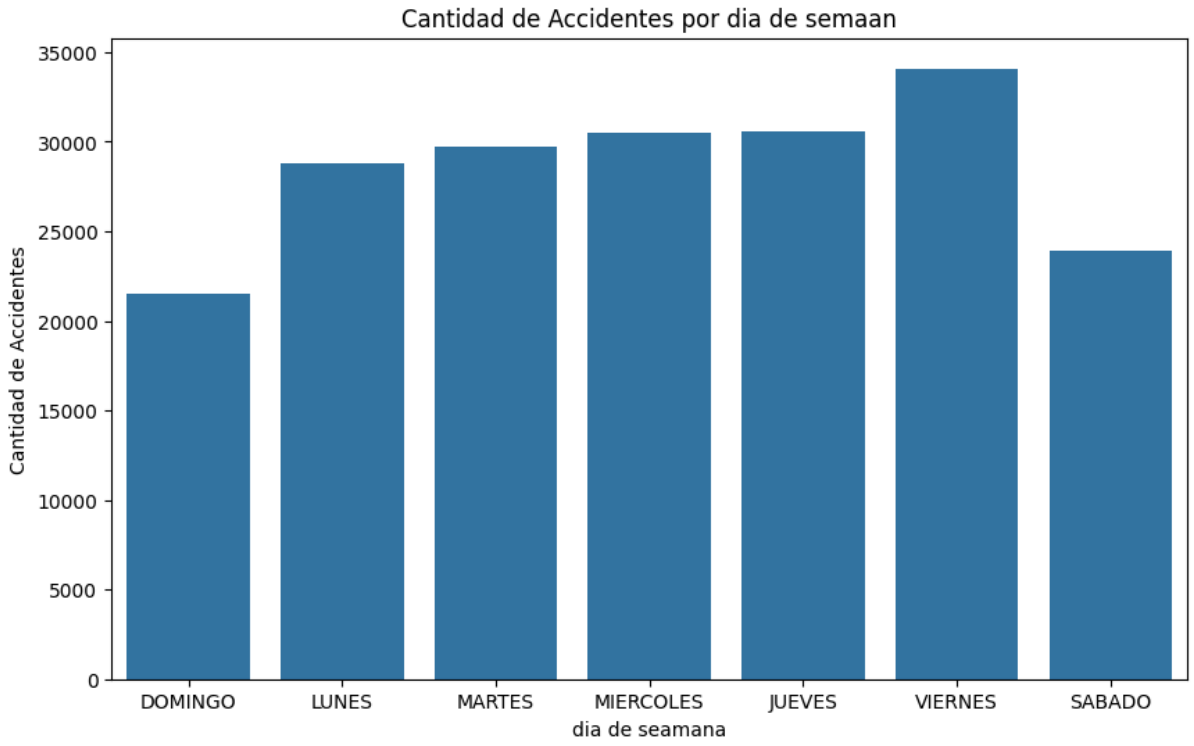


Gráfico de barras de accidentes por distrito: Este gráfico muestra la distribución geográfica de los accidentes, permitiendo identificar los distritos con mayor incidencia.

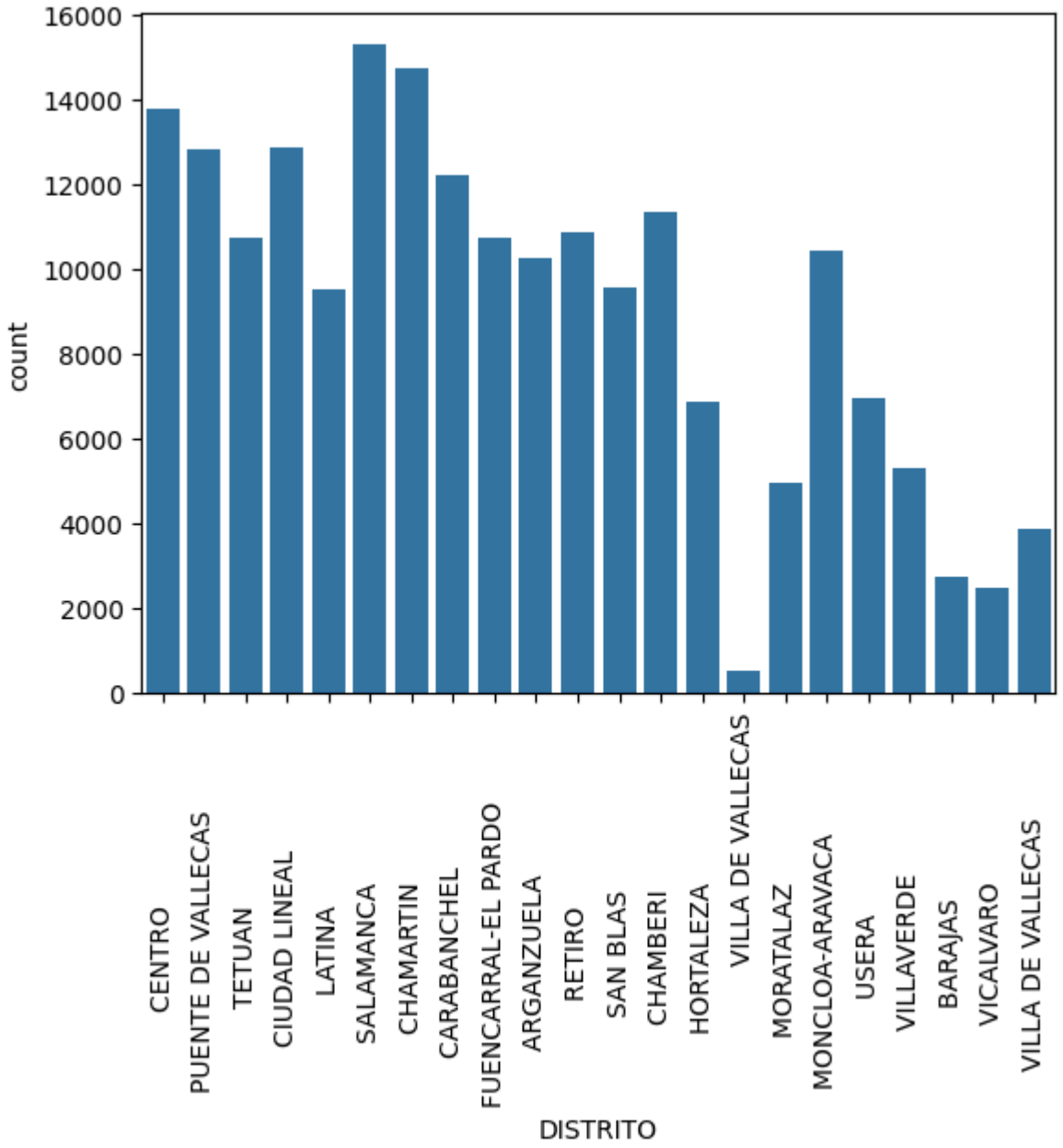
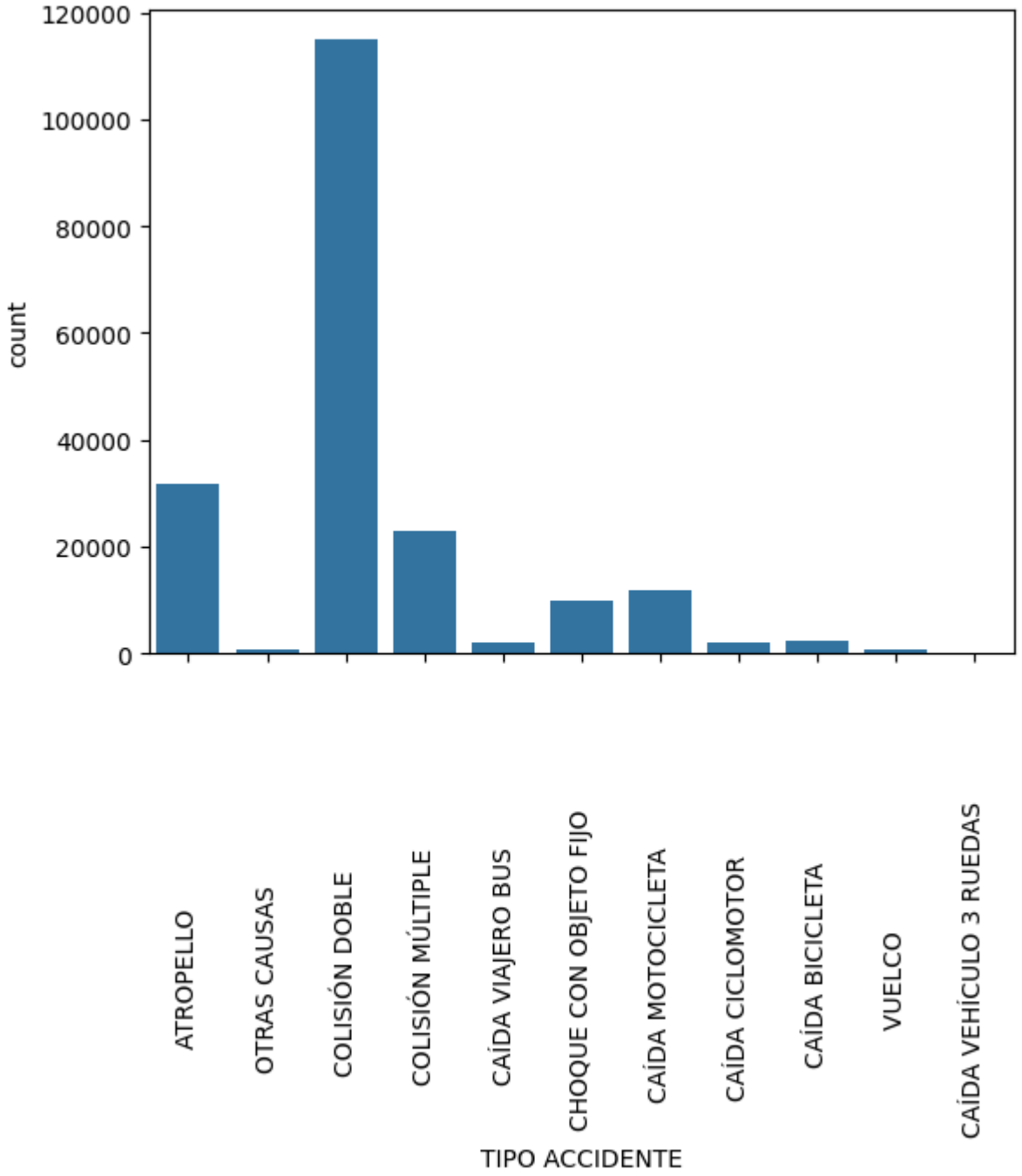


Gráfico de barras de accidentes por tipo de accidente: Este gráfico ilustra la frecuencia de los diferentes tipos de accidentes, destacando la colisión doble como la más común.



4. Investigación documental

En esta fase del proyecto, se realizó una profunda investigación documental con el objetivo de explorar diferentes modelos de predicción aplicables al problema de los accidentes de tráfico en Madrid. Se consideraron dos grandes categorías de modelos: modelos de aprendizaje automático (Machine Learning) y modelos de series temporales.

Modelos de aprendizaje automático:

Se estudiaron diversos algoritmos de Machine Learning, los cuales se pueden clasificar en:

Aprendizaje supervisado: En este tipo de aprendizaje, se entrena un modelo con datos etiquetados, donde se conoce la salida deseada para cada entrada. El objetivo es que el modelo aprenda a generalizar a partir de estos datos y pueda predecir la salida para nuevas entradas.

Aprendizaje no supervisado: En este caso, se entrena un modelo con datos no etiquetados, es decir, sin información sobre la salida deseada. El objetivo es que el modelo descubra patrones y estructuras en los datos por sí mismo.

Aprendizaje semi-supervisado: Se entrena un modelo con una combinación de datos etiquetados y no etiquetados.

Aprendizaje por refuerzo: Un agente aprende a interactuar con un entorno tomando acciones y recibiendo recompensas o penalizaciones.

Modelos específicos para la predicción de accidentes:

Dentro de los modelos de Machine Learning, se identificaron algunos que son particularmente relevantes para la predicción de accidentes de tráfico:

Deep learning: Las redes neuronales profundas son capaces de analizar grandes volúmenes de datos, incluyendo imágenes (ej. de cámaras de tráfico), datos de sensores y series temporales, para identificar patrones complejos y realizar predicciones precisas.

XGBoost: Es un algoritmo de aprendizaje automático basado en árboles de decisión que se destaca por su eficiencia y rendimiento en competiciones de Machine Learning.

Modelos de series temporales:

Además de los modelos de Machine Learning, se investigaron los modelos de series temporales, que son especialmente útiles para analizar datos que varían en el tiempo, como es el caso de los accidentes de tráfico. Uno de los modelos más conocidos en esta categoría es ARIMA (Autoregressive Integrated Moving Average).

ARIMA es un modelo estadístico que utiliza tres componentes para analizar y predecir series temporales:

- Autoregresivo (AR): Este componente examina cómo los valores actuales de la serie están relacionados con sus valores pasados.
- Integrado (I): Este componente se encarga de la estacionariedad de la serie.
- Media Móvil (MA): Este componente considera los errores de predicción pasados para mejorar las predicciones futuras.

ARIMA es un modelo flexible que puede adaptarse a una amplia variedad de patrones de series temporales, pero requiere que la serie sea estacionaria o pueda transformarse en estacionaria.

5. Analítica predictiva

En esta fase, se seleccionó un modelo de series temporales para la predicción del número de accidentes en los meses futuros. El conjunto de datos se dividió en dos partes:

- Conjunto de entrenamiento: Incluyó los datos desde el inicio del conjunto de datos original hasta septiembre de 2018.
- Conjunto de prueba: Este conjunto incluyó los datos a partir de octubre de 2018.

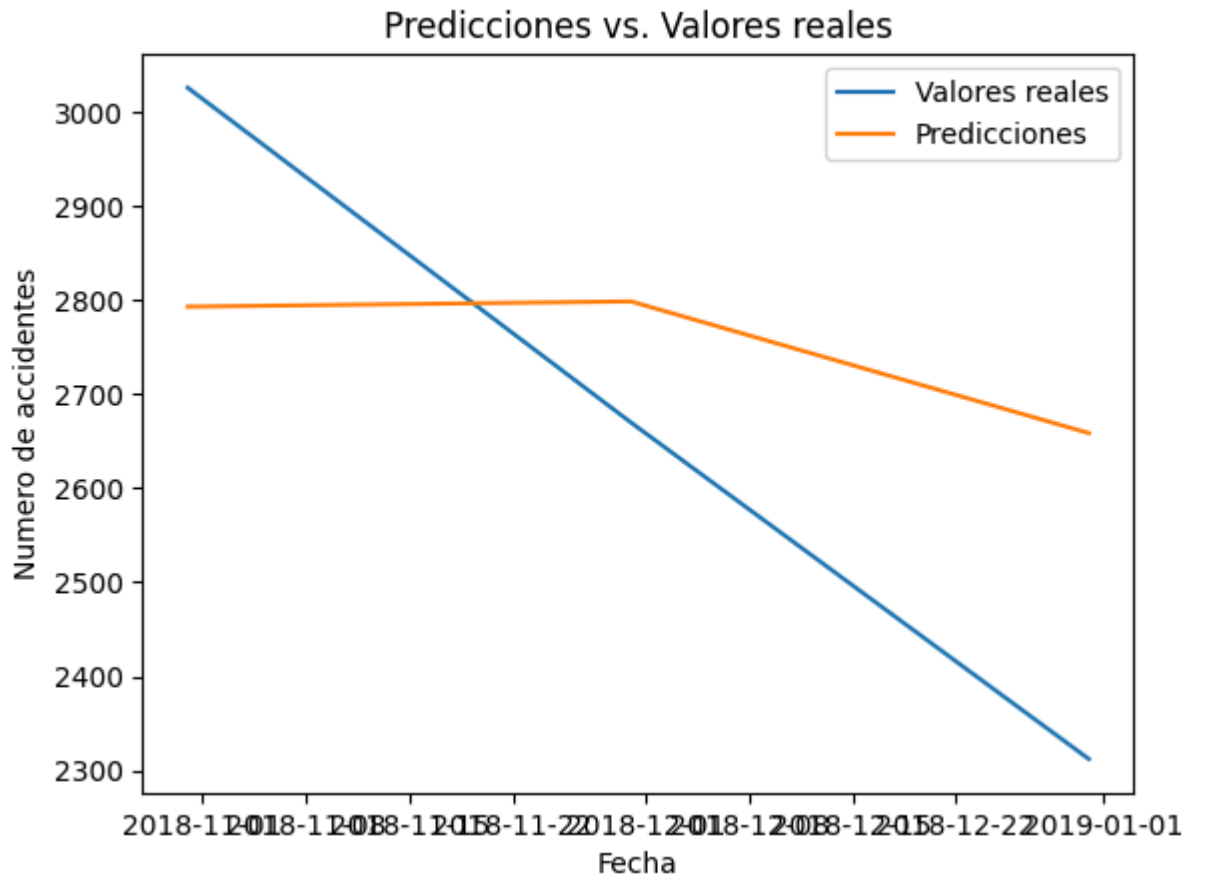
Modelo ARIMA: Se utilizó el modelo ARIMA (Autoregressive Integrated Moving Average) para la predicción. Este modelo es una herramienta estadística que se utiliza para analizar y pronosticar series temporales. ARIMA funciona modelando la serie temporal como una combinación lineal de sus propios valores pasados y errores de predicción pasados. El modelo se define por tres parámetros:

- p: El orden del componente autorregresivo (cuántos valores pasados se utilizan).
- d: El número de diferencias aplicadas para hacer la serie estacionaria.
- q: El orden del componente de media móvil (cuántos errores de predicción pasados se utilizan).

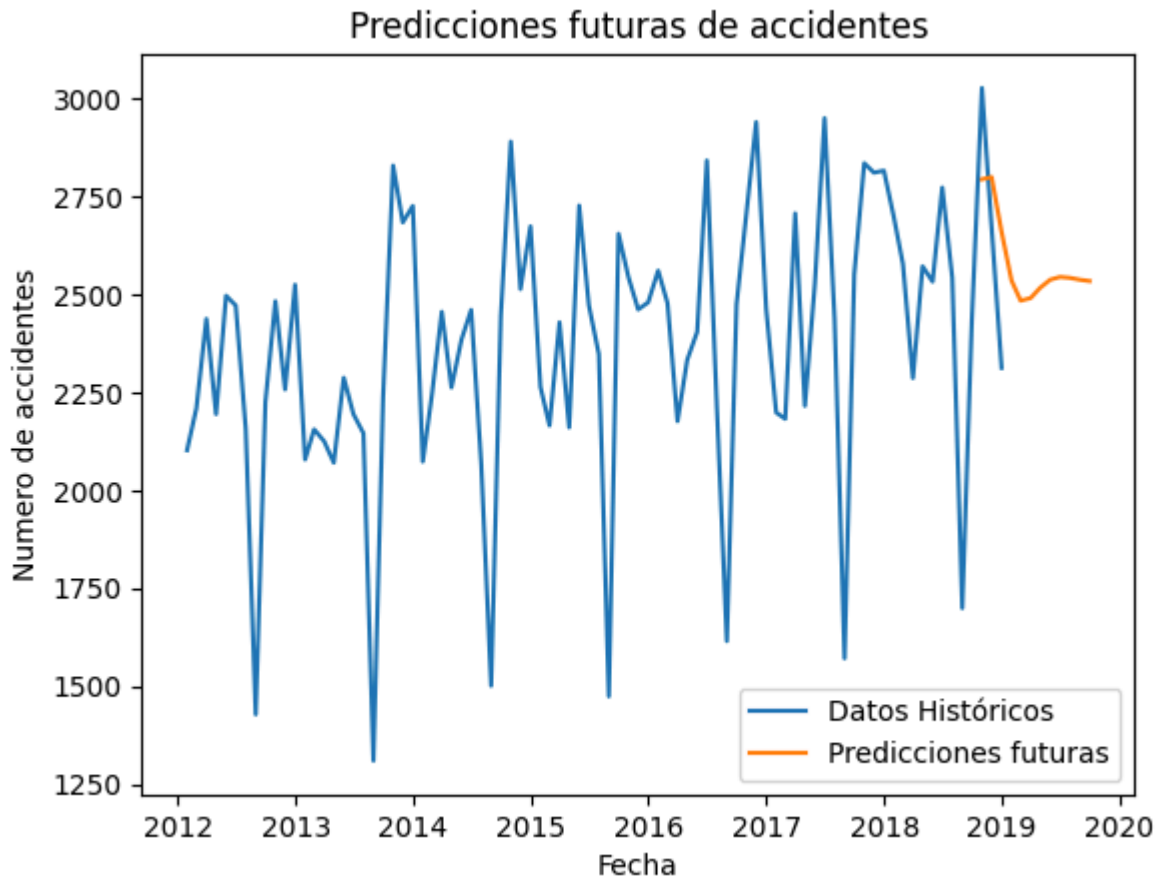
Ajuste del modelo: Se utilizó la función auto_arima de la librería pmdarima para encontrar los parámetros óptimos del modelo ARIMA de forma automática. Los parámetros (p, d, q) seleccionados fueron (2, 1, 2). Con estos parámetros, se entrenó el modelo ARIMA utilizando el conjunto de entrenamiento.

Evaluación del modelo: Se evaluó la calidad del modelo utilizando el conjunto de prueba. Se calculó el error cuadrático medio (RMSE) entre las predicciones del modelo y los valores reales. El RMSE obtenido fue de 252.29.

```
RMSE: 252.29103357457382
```



Predicciones: Finalmente, se realizaron predicciones para los meses de octubre, noviembre y diciembre de 2018. También se generaron predicciones para los 12 meses siguientes a octubre de 2018.



Conclusión: El modelo ARIMA, aunque con limitaciones, permitió realizar predicciones del número de accidentes para los meses futuros. Se sugiere explorar la incorporación de variables adicionales y la afinación del modelo para mejorar la precisión de las predicciones

6. Conclusiones generales

El análisis realizado sobre los datos de accidentes de tráfico en Madrid entre 2012 y 2018 ha proporcionado información valiosa sobre los patrones y características de estos eventos en la ciudad. El análisis exploratorio permitió identificar patrones temporales, como la mayor frecuencia de accidentes los viernes y la menor los domingos, así como el rango horario de mayor riesgo. También se observaron patrones geográficos, con el distrito de Salamanca registrando la mayor cantidad de accidentes. Se identificó la colisión doble como el tipo de accidente más común y se encontró que los hombres son los principales conductores implicados.

La investigación documental permitió explorar diferentes modelos de predicción, incluyendo modelos de aprendizaje automático y series temporales. Se seleccionó el modelo ARIMA para la predicción del número de accidentes en los meses futuros, el cual, aunque con limitaciones, ofreció predicciones para los meses de octubre, noviembre y diciembre de 2018.

7. Próximos pasos

1. Ajustar los parámetros del modelo ARIMA: Se deben explorar diferentes configuraciones de los parámetros (p, d, q) para encontrar la que mejor se ajuste a los datos y mejore la precisión de las predicciones.
2. Considerar la estacionalidad: Se debe analizar la presencia de patrones estacionales en los datos y, si los hay, incorporarlos al modelo ARIMA para mejorar su capacidad predictiva.
3. Incorporar variables exógenas: Se deben investigar y seleccionar variables adicionales que puedan influir en el número de accidentes, como factores climáticos, del suelo, de tráfico y eventos especiales, para enriquecer el modelo y obtener predicciones más precisas.
4. Explorar modelos alternativos: Se deben evaluar otros modelos de predicción, como XGBoost, para comparar su rendimiento con el modelo ARIMA y determinar si ofrecen mejores resultados.
5. Integrar el modelo en un sistema de alerta temprana: Se debe evaluar la viabilidad de integrar el modelo predictivo en un sistema de alerta temprana que permita a las autoridades tomar medidas preventivas para reducir el número de accidentes.

8. Referencias

[IBM] (https://www.ibm.com/es-es/topics/machine-learning-algorithms)

[MIOTI]
(https://mioti.es/es/blog-deep-learning-y-seguridad-vial-como-salva-vidas-esta-tecnologia/)

[Public Health, Columbia University]
(https://www.publichealth.columbia.edu/research/population-health-methods/box-jenkins-methodology)

9. Link a código utilizado:

https://colab.research.google.com/drive/1SOllyj7BAsbIKzgUwwSy4aEFgntrqLhg?usp=sharing