
This is the **published version** of the bachelor thesis:

Kharrat Motabite, Almoujtaba; Oropesa Fisica, Ana, tut. PredictorAutoML :
model basat en Machine Learning per a la predicció de vendes de vehicles. 2024.
(Enginyeria de Dades)

This version is available at <https://ddd.uab.cat/record/317334>

under the terms of the  license

This is the **published version** of the bachelor thesis:

Kharrat Motabite, Almoujtaba; Oropesa Fisica, Ana, tut. PredictorAutoML :
model basat en Machine Learning per a la predicció de vendes de vehicles. 2024.
(Enginyeria de Dades)

This version is available at <https://ddd.uab.cat/record/317334>

under the terms of the  license

PredictorAutoML: modelo basado en Machine Learning para la predicción de venta de vehículos

Almoujtaba Kharrat Motabite

Resumen— Este trabajo desarrolla un modelo de predicción de ventas de automóviles en Estados Unidos utilizando Machine Learning y análisis de datos. La venta de coches está influenciada por factores como la economía y las estrategias comerciales. Para prever la demanda, se analizan datos históricos, considerando variables clave como precios, tasas de interés y promociones, las cuales suelen intensificarse a finales de año para facilitar la liquidación de stock. Se implementan modelos de regresión lineal, series temporales y árboles de decisión, evaluando su precisión. Finalmente, los resultados se visualizan en Tableau, permitiendo identificar tendencias y optimizar la toma de decisiones en el mercado automovilístico, con especial foco en los vehículos eléctricos.

Palabras clave— Machine Learning, Predicción de ventas, Industria automotriz, Análisis de datos, Inteligencia artificial, Python, RStudio, Tableau.

Abstract— This project develops a car sales prediction model in the United States using Machine Learning and data analysis. Car sales are influenced by factors such as the economy and commercial strategies. To forecast demand, historical data is analyzed, considering key variables such as prices, interest rates, and promotions, which tend to intensify at the end of the year to facilitate stock clearance. Linear regression, time series forecasting, and decision tree models are implemented and evaluated for accuracy. Finally, the results are visualized in Tableau, allowing for the identification of trends and optimization of decision-making in the automotive market, with a special focus on electric vehicles.

Index Terms— Machine Learning, Sales Prediction, Automotive Industry, Data Analysis, Artificial Intelligence, Python, RStudio, Tableau.



1 INTRODUCCIÓN

El mercado de los automóviles ha cambiado mucho en los últimos años. La llegada de los coches eléctricos [1], las nuevas normas medioambientales [2] y los cambios en la economía [3] han influido bastante en la manera en que las personas toman decisiones de compra de vehículos. Hoy en día, predecir cuántos coches se venderán en el futuro es muy importante para las marcas y los concesionarios, ya que les ayuda a planificar su producción y mejorar sus estrategias de venta.

Sin embargo, hacer estas predicciones no es sencillo, ya que hay muchos factores que pueden afectar las ventas, como el precio de los vehículos, los cambios en los aranceles [4], los descuentos ofrecidos por los fabricantes, la situación económica [5] y las preferencias de los consumidores.

En este contexto, el uso de herramientas de inteligencia artificial y análisis de datos tiene una gran importancia [6]. A través de la exploración de datos históricos, es posible identificar patrones que permitan prever con mayor precisión las tendencias en la compra de automóviles.

Este trabajo busca analizar el comportamiento de las ventas de automóviles en los últimos años y desarrollar un modelo que permita anticipar la demanda futura. Para ello, se emplearán herramientas de análisis de datos y aprendizaje [7].

2 OBJETIVOS DEL PROYECTO

El objetivo general del proyecto es predecir las ventas futuras de vehículos, con un enfoque especial en los vehículos eléctricos, mediante el uso de modelos de Machine Learning. Para alcanzar este objetivo, se establecen los siguientes objetivos específicos:

- Recopilación de datos históricos sobre las ventas de vehículos, incluyendo variables como precios, tasas de interés y tendencias del mercado de los últimos años, centrándose en los modelos eléctricos por su creciente importancia.
- Análisis sobre los datos para identificar patrones, detectar valores atípicos y seleccionar las variables más relevantes que influyen en la demanda de vehículos eléctricos.

- Implementación de diferentes modelos de Machine Learning con el fin de compararlos entre ellos para ver cuál es el más óptimo para este proyecto, especialmente en el segmento de vehículos eléctricos.
- Evaluación de los modelos mediante métricas de precisión con el objetivo de seleccionar la mejor estrategia de predicción.
- Creación de un dashboard para visualizar y facilitar la interpretación de los resultados.

3 METODOLOGÍA Y PLANIFICACIÓN

El proyecto sigue la metodología CRISP-DM [8], un enfoque ampliamente utilizado en análisis de datos y Machine Learning [9]. Esta metodología permite estructurar el estudio en distintas etapas, facilitando la organización del trabajo y garantizando un desarrollo eficiente del modelo de predicción de ventas de automóviles.

Para la gestión y planificación de las tareas, se ha utilizado Kanban, implementado a través de la herramienta Trello [10]. Este sistema permite visualizar el estado de cada actividad, priorizar tareas y realizar un seguimiento para el avance del proyecto.

Además, se ha diseñado un Diagrama de Gantt [11] que proporciona una visión clara del cronograma del proyecto, asegurando la correcta distribución del tiempo en cada etapa. Este diagrama permite identificar dependencias entre tareas y establecer plazos realistas para su ejecución. La planificación detallada puede consultarse en el Anexo A1.

4 ESTADO DEL ARTE

En los últimos años, el análisis de datos se ha convertido en una herramienta fundamental para anticipar tendencias en sectores tan dinámicos como el automovilístico. Con la digitalización de procesos y la disponibilidad creciente de grandes volúmenes de datos históricos, cada vez más empresas recurren al análisis predictivo para optimizar su producción, campañas comerciales y toma de decisiones estratégicas.

4.1. Predicción de ventas en el sector del automóvil

La predicción de ventas es una de las aplicaciones más frecuentes del aprendizaje automático (Machine Learning) en el ámbito empresarial. Modelos como la regresión lineal [12], los árboles de decisión o los modelos de bosque aleatorio (Random Forest) [13] permiten predecir el número de vehículos vendidos en base a variables como el precio, tipo de coche, fabricante o fecha de compra. Estas técnicas son especialmente útiles para identificar patrones temporales

y de comportamiento del comprador, ayudando a las empresas a planificar su producción y sus estrategias de stock.

Estudios recientes han demostrado que el uso de modelos de aprendizaje supervisado, entrenados con datos históricos de ventas, mejora significativamente la precisión en la estimación de la demanda futura, lo que permite reducir pérdidas por sobrestock.

4.2. Series temporales y análisis estacional

Además de los modelos predictivos clásicos, las series temporales son herramientas muy utilizadas cuando se trata de analizar tendencias mensuales o anuales. Modelos como ARIMA [14] son habituales en contextos donde las ventas están influenciadas por factores estacionales como promociones, fin de año o lanzamientos de nuevos modelos. Este tipo de enfoque permite tener en cuenta tanto las tendencias generales como los picos de demanda específicos.

4.3. Visualización y exploración de datos

Antes de entrenar cualquier modelo de predicción, es fundamental realizar un análisis exploratorio para entender el comportamiento de los datos. En este trabajo, se han generado visualizaciones que muestran la evolución mensual de ventas, los tipos de vehículos más vendidos, la distribución por tipo de energía, el top de fabricantes y los colores preferidos por los clientes.

Estas visualizaciones no solo ayudan a comprender el contexto, sino que también permiten detectar variables relevantes para la predicción y posibles relaciones entre ellas. Por ejemplo, se observa que los SUV y Hatchback son los más vendidos, o que los coches eléctricos ya representan una cuota significativa del mercado, lo que refleja una visión más hacia energías más sostenibles.

4.4. Casos de uso reales

Marcas líderes como Toyota, Hyundai o Tesla han adoptado modelos de predicción para ajustar su producción y responder de manera más ágil a las variaciones del mercado. Estas empresas utilizan algoritmos de Machine Learning [15] para anticipar picos de demanda, adaptar campañas de marketing y gestionar sus cadenas de suministro de forma más eficiente.

5 Desarrollo

Durante esta sección se detallan las decisiones tomadas y los pasos realizados a lo largo del proyecto, justificando cada uno en base al análisis previo y los objetivos definidos.

5.1. Elección del Dataset y Preparación Inicial

Para este proyecto se evaluaron varios datasets disponibles en Kaggle y finalmente, se optó por un único dataset que recoge más de 10.000 registros de ventas de vehículos entre 2015 y 2024, incluyendo variables como la fecha de compra, tipo de coche, combustible, color, fabricante, precio y ubicación. Este dataset fue suficiente por su nivel de detalle y temporalidad adecuada para observar tendencias.

Los datos fueron importados y visualizados mediante Python y Pandas. Se realizó una limpieza básica:

- conversión de la variable de fecha.
- creación de la columna Year-Month.
- conversión de ciudades a estados
- eliminación de columnas irrelevantes (como ID o valoraciones del vendedor).
- selección del periodo entre 2015 y 2024.

El conjunto de datos incluye las siguientes variables relevantes para el análisis:

Variable	Descripción	Tipo de dato
Purchased_Date	Fecha de adquisición del vehículo	Fecha (datetime)
Car_Type	Categoría del coche	Categórica (texto)
Energy	Tipo de energía	Categórica (texto)
Color	Color del coche	Categórica (texto)
Manufacturer_Name	Marca del coche	Categórica (texto)
Price-\$	Precio de venta	Númérica (float)
Mileage-Km	Kilometraje registrado	Númérica (entero)
Location	Ciudad de venta	Categórica (texto)
Number_of_Seats	Número de asientos del vehículo	Númérica (entero)
Number_of_Doors	Número de puertas del vehículo	Númérica (entero)
Año	Extraído de Purchased_Date	Númérica (entero)
Year-Month	Combinacion de año/mes	Fecha (periodo)

Figura 1. Variables relevantes del dataset: nombre, descripción y tipo de dato utilizado en el análisis entre 2015 y 2024

Estas variables se han utilizado tanto en el análisis exploratorio como en la preparación del modelo predictivo.

5.2. Preparación para Modelado Predictivo

A partir de las variables disponibles, se estructuró un nuevo dataset mensualizado con una fila por mes y el total de coches vendidos en ese periodo. Este conjunto servirá de base para aplicar modelos de predicción de ventas.

La variable objetivo es el número de coches vendidos por mes y las variables explicativas se basarán en el tratamiento futuro de:

- estacionalidad (mes).
- tipo de coche.
- tipo de energía.
- marca.
- color del vehículo.
- número de asientos y puertas.
- ubicación geográfica (estado).
- año de la venta.

Cabe destacar que, a partir del conjunto de datos original, se generaron distintos datasets derivados adaptados a los objetivos específicos del análisis exploratorio, las visualizaciones y los modelos predictivos.

5.3. Análisis Exploratorio de los Datos (EDA)

Se llevó a cabo un análisis exploratorio [16] para entender el comportamiento de las variables. Se generaron diferentes visualizaciones, entre ellas:

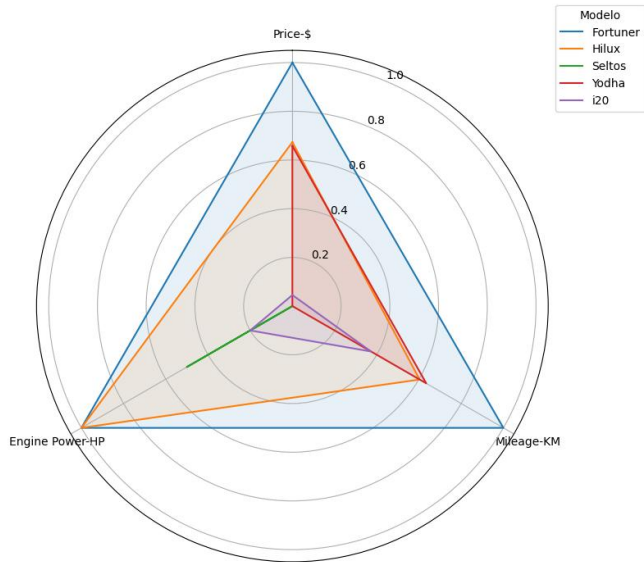


Figura 2. Comparativa de los cinco coches eléctricos más vendidos entre 2015 y 2024

Esta representación ayuda a identificar perfiles distintos entre los modelos, desde los más premium hasta los más urbanos o económicos.

- **Precio-\$**, que representa el coste para el comprador.
- **Potencia (Engine Power-HP)**, como indicador del rendimiento.
- **Kilometraje (Mileage-KM)**, que puede reflejar el uso o la durabilidad del vehículo.

Con estas variables, la comparación entre modelos se vuelve más clara. A partir del gráfico, se pueden distinguir algunos perfiles interesantes:

- Fortuner destaca claramente en las tres variables, lo que sugiere que es un modelo de gama alta, potente y con un uso intensivo.
- Hilux también obtiene buenos valores, algo por debajo del Fortuner, lo que puede indicar un buen equilibrio entre rendimiento y precio.
- Seltos llama bastante la atención por su potencia, por lo que podría atraer a quienes buscan un coche más potente sin irse a los modelos más caros.
- Yodha mantiene valores intermedios enfocado a un perfil más conservador.
- i20 se sitúa en la parte baja de las tres variables, lo que sería un coche más sencillo y pensado para ciudad

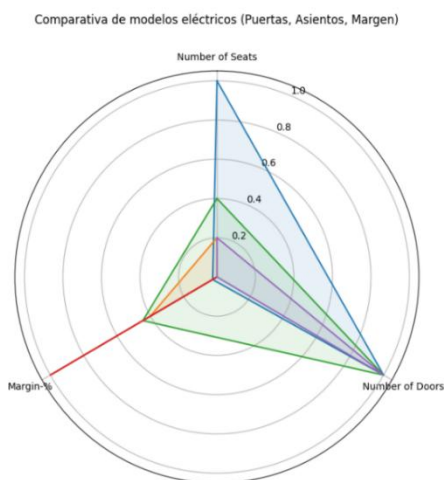


Figura 3. Comparativa de modelos eléctricos según número de puertas, asientos y margen de beneficio

En esta visualización también se comparan los cinco coches eléctricos más vendidos del periodo 2015–2024, pero esta vez analizando tres variables distintas que están más relacionadas con el diseño y la rentabilidad del vehículo:

- **Number of Seats**, número de asientos.
- **Number of Doors**, número de puertas del vehículo.
- **Margen de beneficio (Margin-%)**, que da una idea de la rentabilidad para el fabricante.

Estas variables permiten explorar cómo se posicionan los modelos en cuanto a configuración y estrategia comercial:

- Fortuner vuelve a destacar con el mayor número de asientos y puertas, lo que sugiere que es un modelo más espacioso y familiar.
- Seltos también ofrece buena capacidad interior y, además, muestra un margen de beneficio relativamente alto, por lo que podría resultar interesante para los distribuidores.
- Yodha sobresale claramente por su margen de beneficio, lo que demuestra ser una estrategia más enfocada en rentabilidad, aunque con menos capacidad de pasajeros.
- Hilux y i20 muestran un equilibrio más discreto entre las variables, lo que podría responder a diseños más compactos o a una estrategia orientada a modelos funcionales y asequibles.

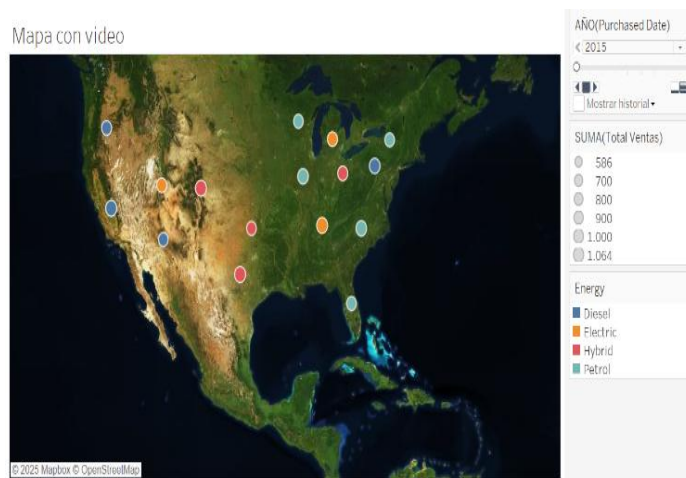


Figura 4. Distribución geográfica de las ventas por tipo de energía en Estados Unidos entre 2015 y 2024

Esta visualización ha sido elaborada con Tableau utilizando la funcionalidad de mapa interactivo.

Se muestra la distribución de ventas de vehículos según el tipo de energía (diésel, gasolina, híbrido, eléctrico) por estado en EE. UU.

Para construir este gráfico, se transformó la variable de ciudad en su correspondiente estado, con el objetivo de agrupar las ventas a un nivel geográfico superior y más representativo para análisis regionales.

Cada punto en el mapa representa un estado y está codificado por color en función del tipo de energía predominante en las ventas de ese año. El tamaño de los círculos representa el volumen total de ventas en cada ubicación.

Además, se incorporó un control deslizante de año que permite observar la evolución del patrón de ventas por tipo de energía a lo largo del tiempo.

Esta visualización permite identificar fácilmente dónde se están adoptando más rápidamente los vehículos eléctricos o híbridos, lo que puede tener implicaciones en estrategias de marketing.

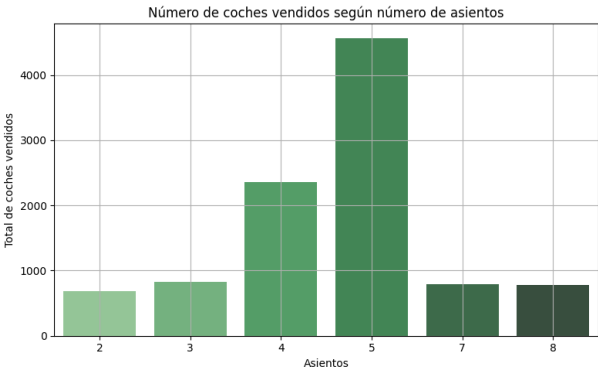


Figura 5. Distribución de ventas de vehículos según el número de asientos

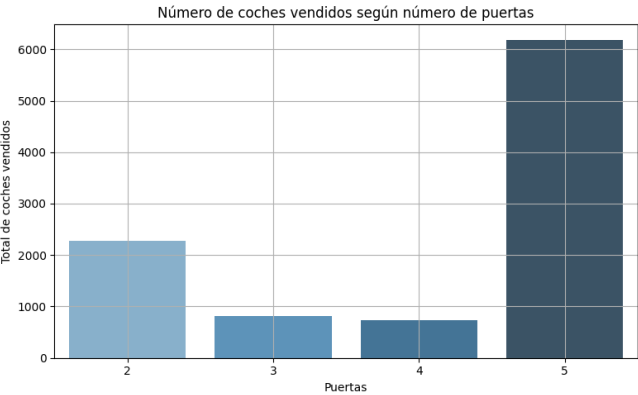


Figura 6. Distribución de ventas de vehículos según el número de puertas

En estas visualizaciones se analiza la distribución de las ventas en función del número de asientos y de puertas por vehículo entre 2015 y 2024.

Se observa que los coches con 5 asientos concentran la mayoría de las ventas, seguido por el de 4 plazas. Este patrón refleja una clara preferencia por vehículos de tamaño familiar estándar, posiblemente por su equilibrio entre capacidad, consumo y precio. Este dato es coherente con la popularidad de los SUV compactos y hatchbacks, que suelen ofrecer esta configuración.

En cuanto al número de puertas, los coches con 2 y 5 puertas son los más vendidos con diferencia. Esto sugiere una fuerte valoración por parte de los consumidores hacia la funcionalidad y accesibilidad, especialmente en contextos urbanos o familiares. Los modelos de 3 o 4 puertas quedan reservados para usos más concretos, como deportivos o utilitarios económicos [17].

Este tipo de análisis permite entender la evolución de la demanda desde una perspectiva estructural. Dado que tanto el número de asientos como el de puertas responden a decisiones de uso, estas variables pueden considerarse indicadores indirectos del perfil del consumidor.

Asimismo, estas variables podrían integrarse en futuros modelos predictivos con el objetivo de afinar las estimaciones de ventas en función de diferentes perfiles de consumidores, contextos económicos o cambios sociales.

Por ejemplo, cosas como que la población está envejeciendo, que hay familias más pequeñas o que cada vez hay más gente viviendo sola pueden hacer que la gente prefiera coches más compactos, funcionales o adaptados a nuevas formas de moverse.

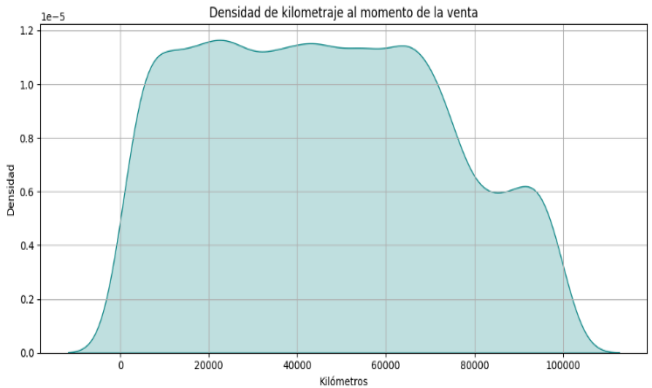


Figura 7. Distribución de densidad del kilometraje al momento de la venta

Esta gráfica muestra la distribución de los vehículos vendidos en función del kilometraje acumulado.

Se observa que la mayoría de los coches se comercializan entre los 10.000 y 70.000 kilómetros, siendo este rango el más denso en cuanto a operaciones de compraventa.

Este comportamiento sugiere que los usuarios tienden a vender sus vehículos en una etapa intermedia del ciclo de vida del coche, cuando aún mantienen un valor razonable de mercado, pero han superado su periodo inicial de mayor depreciación.

A partir de los 70.000 km, la curva de densidad desciende significativamente, lo que podría reflejar una menor demanda de vehículos con mayor desgaste percibido.

Esta información resulta útil para futuros modelos predictivos, ya que permite identificar el rango de kilometraje con mayor probabilidad de transacción, lo cual puede ser determinante a la hora de ajustar la predicción según el perfil del vehículo.

Distribución de ventas por localidad

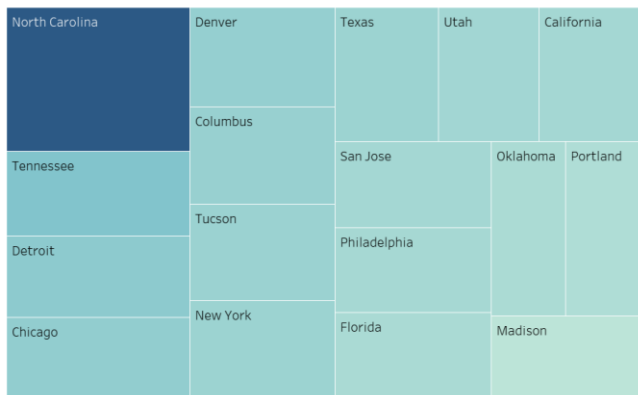


Figura 8. Distribución de ventas por localidad entre 2015 y 2024

En esta visualización destaca claramente North Carolina como la localidad con el mayor volumen de ventas acumuladas, casi duplicando al resto de regiones durante el periodo analizado.

Este resultado sugiere que North Carolina ha sido un foco clave del mercado, posiblemente por una demanda más activa por parte de los consumidores.

Este tipo de análisis geográfico permite detectar zonas prioritarias y orientar con mayor precisión las estrategias comerciales vinculadas a la compra de vehículos.

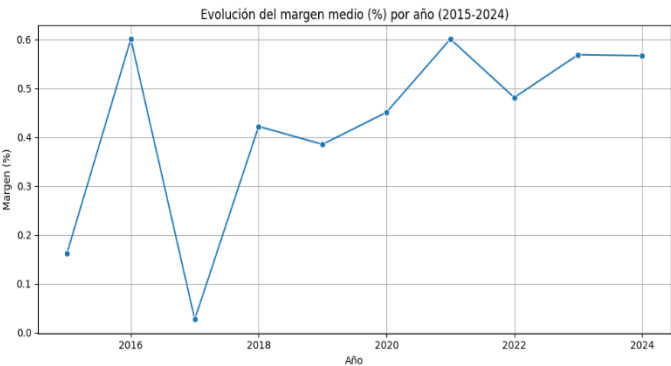


Figura 9. Evolución del margen medio (%) por año entre 2015 y 2024

En esta gráfica se representa cómo ha evolucionado el margen medio por venta a lo largo del tiempo. Se observa una importante variabilidad en los primeros años, destacando especialmente 2016 con el margen más alto (0.60 %) y 2017 con el más bajo (0.03 %). Esta caída podría deberse a descuentos agresivos, liquidación de stock o cambios en la estrategia comercial.

A partir de 2018, el margen empieza a estabilizarse y muestra una tendencia ascendente moderada. Desde 2020, los valores se mantienen altos y estables, por encima del 0.45 %, lo que puede indicar una mejora en la rentabilidad de las ventas o una consolidación del modelo de negocio.

Este análisis resulta útil para evaluar la sostenibilidad del negocio en el tiempo, y sugiere que en los últimos años las operaciones se han vuelto más eficientes desde el punto de vista comercial.

5.4 Implementación de modelos de predicción

Se implementaron cuatro modelos de predicción para estimar la evolución mensual de las ventas de vehículos entre 2015 y 2024.

Estos modelos se entrenaron sobre el mismo conjunto de datos mensualizado, utilizando como variable temporal explicativa el mes ordinal (es decir, cada mes numerado de forma acumulativa).

Los modelos seleccionados fueron los siguientes:

- Regresión lineal: un enfoque clásico que permite detectar una tendencia lineal en la evolución de las ventas a lo largo del tiempo.
- Random Forest: un modelo de ensamble basado en árboles de decisión, capaz de capturar relaciones no lineales entre múltiples variables sin necesidad de supuestos previos.

- SARIMAX: un modelo estadístico autoregresivo con estacionalidad, diseñado específicamente para series temporales, que incorpora efectos del pasado y patrones cíclicos [18].
- XGBoost: un potente algoritmo de boosting por gradiente que se ha destacado por su eficiencia y capacidad de generalización en tareas de regresión estructurada [19].

Cada uno de estos modelos fue entrenado y posteriormente evaluado utilizando métricas como el RMSE (Root Mean Squared Error) y el coeficiente de determinación R².

Esta comparación permitió identificar cuál de ellos se ajusta mejor al comportamiento real del mercado, tal como se presenta en el siguiente apartado.

5.5 Evaluación del modelo

Además de ver los gráficos, también se calcularon las métricas RMSE y R² para evaluar el rendimiento de cada modelo.

El error cuadrático medio (RMSE) permite conocer cuánto se desvían en promedio las predicciones respecto a los valores reales. El coeficiente de determinación R² indica el grado de explicación que ofrece el modelo sobre la variabilidad de los datos. Cuanto más cercano a 1 sea este valor, mejor será el ajuste del modelo a los datos reales.

Comparativa inicial: todos los vehículos

En primer lugar, se aplicaron los cuatro modelos seleccionados: Regresión Lineal, Random Forest, SARIMAX y XGBoost, sobre la serie completa de ventas mensuales de todos los vehículos, independientemente de su tipo de energía.

Esto permitió evaluar el rendimiento de cada modelo sobre el comportamiento global del mercado.

Modelo	RMSE	R²
Regresión lineal	8.03	0.02
Random Forest	4.12	0.74
SARIMAX	10.90	-0.79
XGBoost	1.93	0.94

Figura 10. Comparativa del rendimiento de modelos predictivos sobre toda la serie de datos

Como se puede observar, el modelo XGBoost ofreció un rendimiento significativamente mejor, tanto en precisión (RMSE) como en capacidad de explicación (R²).

Una vez identificados los modelos con mejor rendimiento sobre la serie completa, se decidió aplicar el análisis exclusivamente al subconjunto de vehículos eléctricos por

su creciente importancia en el sector automovilístico.

Enfoque específico: vehículos eléctricos

A partir de estos resultados, se decidió centrar el análisis exclusivamente en los vehículos eléctricos, dado que representan el futuro del sector automovilístico y son clave para la transición energética.

Este enfoque permite obtener predicciones más relevantes para decisiones estratégicas, como planificación de producción o inversión.

Al aplicar los mismos modelos sobre el subconjunto de vehículos eléctricos, se obtuvieron los siguientes resultados:

Modelo	RMSE	R²
Regresión lineal	4.46	0.01
Random Forest	2.04	0.79
SARIMAX	5.90	-0.72
XGBoost	0.79	0.97

Figura 11. Resultados de los modelos aplicados al subconjunto de vehículos eléctricos

El modelo XGBoost volvió a destacar como el más eficaz, con un error muy bajo y una capacidad explicativa muy alta.

Por otra parte, según distintos informes y medios especializados, se espera un crecimiento acelerado de las ventas de vehículos eléctricos en los próximos años, impulsado por:

- Políticas medioambientales más estrictas.
- Subvenciones para vehículos sostenibles.
- Reducción de precios por economías de escala.
- Nuevos lanzamientos de modelos competitivos.

Estas proyecciones permiten simular un escenario de crecimiento del 20% anual acumulado, reflejando una adopción rápida del mercado. Bajo esta hipótesis, se ha hecho una comparación entre ambos escenarios [20].

Escenario	Ventas 2025-2030	Diferencia vs base
proyección base	5.386 unidades	0%
Proyección con +20% anual	9.269 unidades	+71.2%

Figura 12. Comparativa de proyecciones de ventas entre 2025 y 2030 bajo distintos escenarios

Como se observa en la tabla, el volumen proyectado bajo un escenario de crecimiento acelerado sería un 71% superior al estimado por el modelo base. Esta diferencia refleja la magnitud que podría alcanzar el mercado de vehículos eléctricos si se cumplen las condiciones esperadas de impulso institucional, tecnológico y económico.

Estas estimaciones ayudan a hacerse una idea del posible crecimiento del sector y pueden ser útiles como base para planificar inversiones, definir estrategias comerciales o valorar políticas públicas enfocadas en una movilidad más sostenible.

7 CONCLUSIONES

En esta sección se resumen las principales conclusiones obtenidas tras el desarrollo del proyecto. Se incluye lo que ha funcionado, lo que no, lo que se ha aprendido y qué mejoras podrían implementarse en trabajos futuros.

7.1 Conclusiones Generales

El proyecto comenzó con la idea de predecir las ventas de coches en general, pero conforme se avanzó y se comprendieron mejor los datos, se identificó que centrarse en los vehículos eléctricos tenía mucho más sentido por su peso creciente en el mercado. Por ello, se tomó la decisión de enfocar el análisis hacia ese tipo de vehículos.

Se probaron cuatro modelos diferentes: regresión lineal, Random Forest, SARIMAX y XGBoost. En ese momento, esta combinación resultó adecuada porque ofrecía diversidad: modelos más sencillos, modelos estadísticos y otros más potentes como XGBoost. Finalmente, XGBoost fue el que mejor funcionó, especialmente con los coches eléctricos, mostrando un buen ajuste y precisión.

Una de las principales dificultades fue obtener un dataset realmente completo. En varios momentos se detectó la ausencia de variables importantes, como el perfil del comprador, información geográfica detallada o datos económicos de contexto. Esto obligó a trabajar con los datos disponibles y a adaptar las ideas a las posibilidades que ofrecía el dataset. Se pudo comprobar que, fuera del ámbito académico, los datos reales rara vez se presentan “limpios” o completos.

Tampoco se siguió al 100% la planificación inicial. Aunque se comenzó con un diagrama de Gantt claro y se utilizó Trello con metodología Kanban, durante el proceso fue necesario introducir cambios. El más relevante fue el giro hacia los vehículos eléctricos, que no estaba previsto desde el inicio. No obstante, este cambio se considera acertado.

El análisis exploratorio (EDA) permitió extraer varias conclusiones útiles, por ejemplo:

- La mayoría de los coches vendidos tienen 5 puertas y entre 4 y 5 plazas, lo que refleja un claro interés por vehículos prácticos o familiares.
- La mayor parte de las ventas se producen cuando los coches tienen entre 10.000 y 70.000 km, lo que parece un punto óptimo para vender antes de que el valor se deprecie más.
- Modelos como el Fortuner o el Hilux destacan por su potencia, precio y tamaño, mientras que otros como el i20 o el Seltos se ajustan más a un cliente urbano con menor presupuesto.
- North Carolina fue, con diferencia, el estado donde más se vendieron vehículos, lo cual puede ser útil para estrategias comerciales regionales.
- Se observó que el año 2017 presentó el margen medio más bajo de todo el periodo (0.03 %), lo que podría deberse a campañas de liquidación, descuentos fuertes o situaciones similares. A partir de 2020, los márgenes se han mantenido estables por encima del 0.45 %, lo que sugiere que las operaciones son ahora más rentables.

Otro aspecto relevante es que, aunque no se modeló directamente, el impacto de la pandemia del COVID-19 se aprecia en los datos. Es probable que haya afectado tanto a las ventas como a los márgenes en determinados años. En futuros estudios, resultaría interesante incluir variables dummy para capturar este tipo de efectos externos.

En resumen, este proyecto no solo ha permitido poner en práctica distintos modelos de predicción, sino también comprender cómo funcionan realmente los datos en un contexto de negocio, qué decisiones deben tomarse durante el análisis y lo importante que resulta adaptarse a las circunstancias.

7.2 Limitaciones y Mejoras Futuras

Aunque el proyecto ha resultado satisfactorio, existen múltiples aspectos que podrían mejorarse:

En primer lugar, sería útil añadir variables externas como el PIB, los precios del combustible o los incentivos a la compra de coches eléctricos, ya que tienen una influencia significativa en la demanda.

- También sería recomendable explorar modelos más avanzados como las redes neuronales recurrentes (RNN)[21], que pueden ofrecer mayor precisión en contextos temporales complejos.
- Otra mejora consistiría en segmentar el análisis por fabricante, tipo de cliente o zona geográfica, con el fin de realizar predicciones más personalizadas.

- Además, si se quisiera aplicar este proyecto en un entorno empresarial, sería necesario conectar los resultados con dashboards o herramientas visuales para facilitar la toma de decisiones.

- Por último, resultaría interesante incluir variables que representen eventos externos (como el COVID), para que el modelo tenga en cuenta situaciones excepcionales que afectan directamente al mercado.

AGRADECIMIENTOS

Quiero agradecer a la Universidad Autónoma de Barcelona por ofrecerme un entorno de aprendizaje cómodo a lo largo de estos años. En especial, agradezco a mi tutora Ana Oropesa por su dedicación, su apoyo constante y sus consejos que me han ayudado a enfocar correctamente este trabajo.

También quiero dar las gracias a mis compañeros de clase, con quienes he compartido dudas, ideas y aprendizajes durante todo el proceso. Su ayuda ha sido clave para avanzar, sobre todo en los momentos más difíciles.

Y por supuesto, a mi familia y amigos, por animarme siempre, confiar en mí y estar a mi lado mientras preparaba este proyecto. Sin su apoyo y motivación, todo esto habría sido mucho más difícil.

BIBLIOGRAFIA

[1] La Sexta, "Hay fecha para la llegada de coches eléctricos con más de 1.000 km de autonomía que cambiarán todo," 2024. Disponible en: https://www.lasexta.com/motor/noticias/hay-fecha-llegada-coches-electricos-mas-1000-autonomia-que-cambiaran-todo_202411016724d7e5d8f8950001d74fc7.html [Accedido: 08-marzo-2025].

[2] Reale Seguros, "Nuevas normas DGT 2025: todas las novedades," 2024. Disponible en: <https://blog.reale.es/nuevas-normas-dgt-2025-todas-las-novedades/> [Accedido: 08-marzo-2025].

[3] WikiGeografía, "Evolución de la economía mundial: crecimiento, desigualdades y etapas clave del siglo XX y XXI," 2024. Disponible en: <https://www.wikigeografia.net/evolucion-de-la-economia-mundial-crecimiento-desigualdades-y-etapas-clave-del-siglo-xx-y-xxi/> [Accedido: 08-marzo-2025].

[4] AutoBild, "Los aranceles a los coches chinos ya son un hecho: entran en vigor hoy mismo," 2024. Disponible en: <https://www.autobild.es/noticias/aranceles-coches-chinos-ya-son-hecho-entran-vigor-hoy-mismo-1415267> [Accedido: 08-marzo-2025].

[5] Cadena SER, "¿Por qué los españoles sienten que están en crisis mientras la economía crece a ritmo de récord?" 2024. Disponible en: <https://cadenaser.com/nacional/2024/12/20/por-que-los-espanoles-sienten-que-estan-en-cri-sis-mientras-la-economia-crece-a-ritmo-de-record-cadena-ser/> [Accedido: 08-marzo-2025].

[6] ROISense, "Inteligencia artificial en el análisis de datos," 2024. Disponible en: <https://www.roisense.com/inteligencia-artificial-en-el-analisis-de-datos/> [Accedido: 08-marzo-2025].

[7] Innovación Digital 360, "Análisis de datos: técnicas y metodologías para la aplicación de analytics," 2024. Disponible en: <https://www.innovaciondigital360.com/big-data/analisis-de-datos-tecnicas-y-metodologias-para-la-aplicacion-de-analytics/> [Accedido: 08-marzo-2025].

[8] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. Disponible en: <https://www.the-modeling-agency.com/crisp-dm.pdf> [Accedido: 08-marzo-2025].

[9] ROISense, "Inteligencia artificial en el análisis de datos," 2024. Disponible en: <https://www.roisense.com/inteligencia-artificial-en-el-analisis-de-datos/> [Accedido: 08-marzo-2025].

[10] Trello, "Herramienta de gestión de proyectos con Kanban," 2024. Disponible en: <https://trello.com> [Accedido: 08-marzo-2025].

[11] Asana, "Fundamentos de los diagramas de Gantt," disponible en: <https://asana.com/es/resources/gantt-chart-basics>, [Accedido: 08-marzo-2025].

[12] A. Zhang, "Linear Regression," Dive into Deep Learning, 2023. Disponible en: https://d2l.ai/chapter_linear-regression/linear-regression.html [Accedido: 15-marzo-2025].

[13] GeeksforGeeks, "Random Forest Algorithm in Machine Learning," 2023. Disponible en: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/> [Accedido: 26-marzo-2025].

[14] DataCamp, "ARIMA Model in Python: Complete Guide to Time Series Forecasting," 2023. Disponible en: <https://www.datacamp.com/es/tutorial/arima> [Accedido: 30-marzo-2025].

[15] RAutomation, "Machine Learning revoluciona la industria automovilística," 2024. Disponible en: <https://www.rautomation.es/2024/02/06/machine-learning-revoluciona-industria-automovilistica/> [Accedido: 01-abril-2025].

[16] MachineLearningPlus, "Exploratory Data Analysis (EDA) in Python," Disponible en: <https://www.machinelearningplus.com/machine-learning/exploratory-data-analysis-eda/> [Accedido: 10-abril-2025].

[17] TopCoches, "Coches de 3 puertas: los más vendidos y recomendables en 2024," Disponible en: <https://topcoches.es/coches-3-puertas/> [Accedido: 18-mayo-2025].

[18] Ciencia de Datos, "Modelos ARIMA y SARIMAX en Python," Disponible en: <https://cienciadedatos.net/documentos/py51-modelos-arima-sarimax-python> [Accedido: 18-mayo-2025].

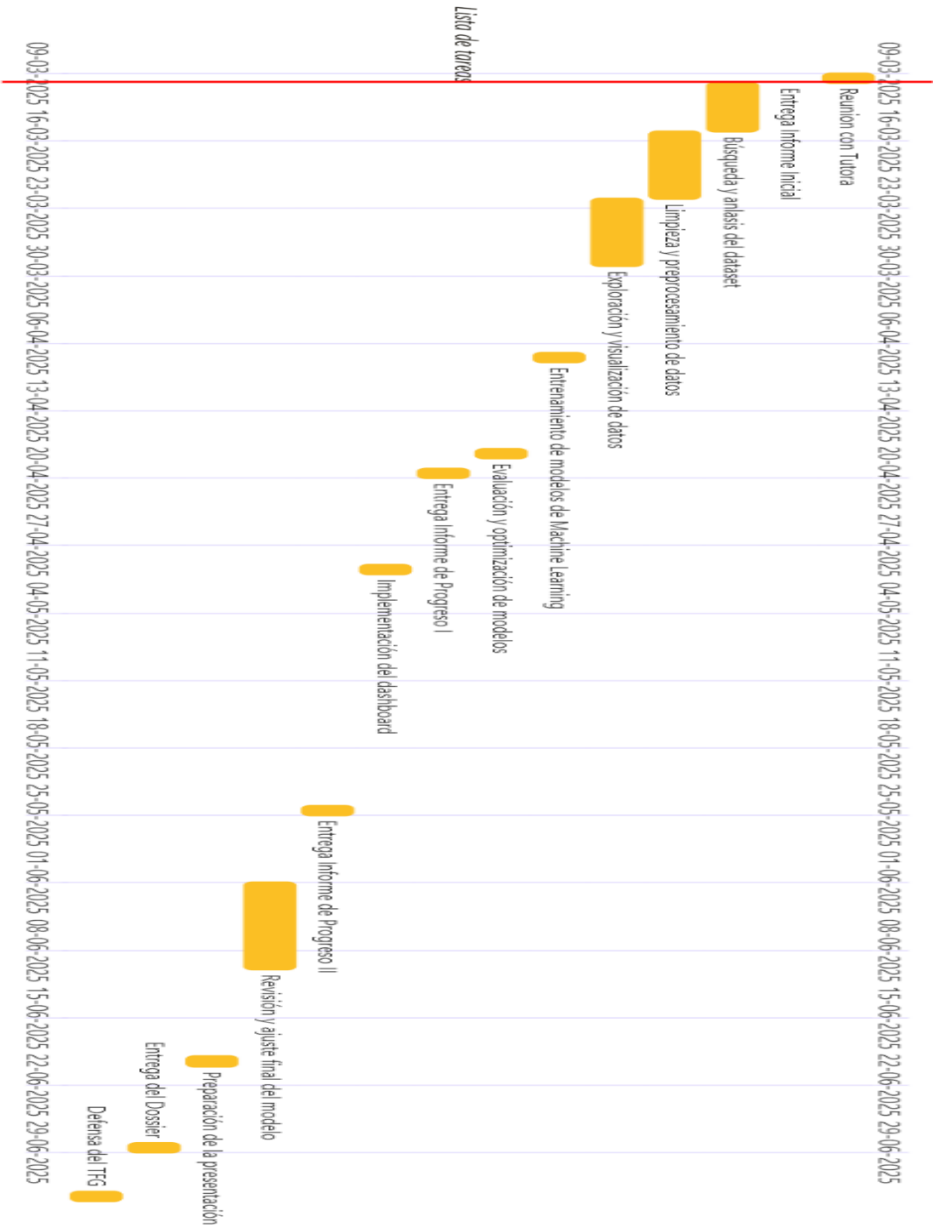
[19] DataCamp, "XGBoost in Python: A Step-by-Step Tutorial," 2024. Disponible en: <https://www.datacamp.com/tutorial/xgboost-in-python> [Accedido: 23-mayo-2025].

[20] International Energy Agency, "More than 1 in 4 cars sold worldwide this year is set to be electric as EV sales continue to grow," IEA, 2024. Disponible en: <https://www.iea.org/news/more-than-1-in-4-cars-sold-worldwide-this-year-is-set-to-be-electric-as-ev-sales-continue-to-grow> [Accedido: 25-mayo-2025].

[21] LovTechnology, "Redes neuronales recurrentes (RNN): Entendiendo el procesamiento de secuencias," 2024. Disponible en: <https://lovtechnology.com/redes-neuronales-recurrentes-rnn-entendiendo-el-procesamiento-de-secuencias/> [Accedido: 22-mayo-2025].

A1. ANEXO

DIAGRAMA DE GANTT



A2. GLOSARIO

1. Machine Learning (ML): Rama de la inteligencia artificial que permite a los ordenadores aprender de los datos y hacer predicciones sin estar programados explícitamente para cada tarea.
2. XGBoost: Algoritmo de aprendizaje supervisado basado en árboles de decisión, optimizado para alto rendimiento y precisión, especialmente eficaz en problemas de regresión y clasificación.
3. Random Forest: Conjunto de árboles de decisión entrenados de forma aleatoria que mejora la precisión del modelo al reducir el sobreajuste.
4. Regresión Lineal: Modelo estadístico que establece una relación lineal entre una variable dependiente y una o más variables independientes.
5. SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors): Modelo de series temporales que incorpora estacionalidad, tendencia y variables externas para hacer predicciones más precisas.
6. RMSE (Root Mean Squared Error): Métrica que calcula la raíz cuadrada del error cuadrático medio entre los valores reales y los predichos por el modelo.
7. R^2 (Coeficiente de determinación): Métrica que indica el grado de ajuste de un modelo a los datos. Su valor varía entre 0 (sin ajuste) y 1 (ajuste perfecto).
8. EDA (Exploratory Data Analysis): Conjunto de técnicas estadísticas y visuales utilizadas para comprender mejor la estructura y patrones de un conjunto de datos antes de aplicar modelos predictivos.
9. Ventas mensualizadas: Transformación de los datos de ventas para agregarlos por mes, facilitando así el análisis de tendencias temporales.
10. Vehículo eléctrico: Tipo de coche que funciona total o parcialmente con electricidad, en contraposición a los tradicionales de combustión interna.
11. Predicción: Estimación o proyección de valores futuros a partir de patrones aprendidos por el modelo sobre datos históricos.
12. Escenario de crecimiento: Supuesto en el que se modela un aumento anual acumulado de un 20% en las ventas, para simular una evolución acelerada del mercado.