Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

ANÁLISIS PREDICTIVO DE TARIFAS DE VIAJES EN LA PLATAFORMA UBER: APLICACIONES DE MODELOS DE APRENDIZAJE AUTOMÁTICO EN INTELIGENCIA DE NEGOCIOS

Informe Final / Remanente

DOCENTE: FRANCO MASTELLI

ASIGNATURA: INTELIGENCIA DE NEGOCIOS

Repositorio GitHub.

ALUMNO: JULIÁN DELGADILLO MARÍN
POSGRADO: MAESTRÍA EN ECONOMÍA APLICADA

FECHA: 16 de Noviembre del 2025

Resumen

El presente trabajo aborda el desarrollo de un sistema predictivo para la estimación de tarifas de viajes Uber en la ciudad de Nueva York, a partir del análisis de datos históricos y la aplicación de técnicas de Inteligencia de Negocios y aprendizaje automático. Se emplearon distintos enfoques de modelado —regresión lineal ordinaria (OLS), regresión regularizada LASSO, modelos de ensamble (*Random Forest* y *Gradient Boosting*) y redes neuronales densas— con el propósito de evaluar su capacidad explicativa y comparativa en la predicción del valor de la tarifa (`fare_amount`) en función de la distancia recorrida (`distance_km`). Los resultados muestran una fuerte relación positiva entre ambas variables, con un coeficiente promedio cercano a 2 USD por kilómetro y un nivel de explicación superior al 67 % en el modelo lineal. Los métodos de ensamble y las redes neuronales superaron el desempeño de los modelos lineales en términos de precisión, destacándose el *Gradient Boosting* como el modelo de mejor equilibrio global (RMSE = 2.46; MAE = 1.73). El análisis confirma la eficacia de los métodos de aprendizaje automático para predecir tarifas con alta precisión a partir de información mínima, así como la importancia de la integración entre analítica descriptiva y predictiva en entornos de movilidad urbana. Finalmente, el estudio ofrece una base metodológica replicable para el desarrollo de modelos similares en contextos de transporte inteligente, destacando la utilidad de la analítica avanzada en la optimización de precios, planificación de demanda y toma de decisiones estratégicas en plataformas digitales de movilidad.

Palabras clave: Inteligencia de Negocios, Aprendizaje Automático, Regresión Lineal, Modelos de Ensamble, Redes Neuronales, Uber, Predicción de Tarifas, Análisis de Datos, Movilidad Urbana.

Índice

1. Introducción

En el presente trabajo se desarrolla un ejercicio integral de *Inteligencia de Negocios* orientado al análisis, modelado y predicción de tarifas en servicios de transporte urbano. El caso de estudio se basa en un conjunto de datos de viajes de *Uber* en la ciudad de Nueva York, que contiene información georreferenciada de origen y destino, junto con la tarifa final asociada a cada trayecto. El objetivo general consiste en explorar la relación entre la distancia recorrida y el monto cobrado (`fare_amount`), aplicando técnicas de limpieza, análisis exploratorio y modelado predictivo mediante distintos enfoques de aprendizaje supervisado.

La problemática resulta representativa de los desafíos actuales en analítica de datos: disponer de grandes volúmenes de información heterogénea, identificar patrones relevantes y construir modelos que permitan predecir comportamientos con precisión y robustez. En este contexto, se emplean herramientas de *Business Intelligence* y de *Machine Learning* con el propósito de comparar el desempeño de modelos lineales tradicionales (Mínimos Cuadrados Ordinarios y LASSO), métodos de ensamble basados en árboles de decisión (*Random Forest* y *Gradient Boosting*), y arquitecturas de redes neuronales artificiales.

El estudio se estructura de manera incremental. En primer lugar, se realiza un análisis exploratorio de los datos (EDA), abordando la detección y eliminación de valores atípicos, el cálculo de distancias geodésicas mediante la fórmula de Haversine y la caracterización estadística de las variables. Posteriormente, se divide el conjunto de datos en muestras de entrenamiento y prueba para garantizar la validez de los resultados. A partir de ello, se entrenan y evalúan los modelos mencionados, utilizando métricas estándar de desempeño tales como el *Root Mean Squared Error* (RMSE) y el *Mean Absolute Error* (MAE).

Los resultados permiten cuantificar el grado de ajuste y capacidad predictiva de cada modelo, mostrando la evolución del desempeño desde los enfoques lineales hasta los no lineales y neuronales. Finalmente, se discuten las implicancias de los hallazgos para el campo de la Inteligencia de Negocios, subrayando la importancia de la calidad de los datos, la selección de modelos y el equilibrio entre interpretabilidad y precisión.

2. Objetivos

2.1. Objetivo general

Analizar, modelar y predecir el comportamiento de la variable `fare_amount` en función de la distancia recorrida (`distance_km`) utilizando técnicas de *Inteligencia de Negocios* y *Aprendizaje Automático*, con el fin de comparar el desempeño de distintos enfoques de modelado —lineales, de ensamble y neuronales— y evaluar su capacidad para explicar y estimar las tarifas de los viajes en la ciudad de Nueva York.

2.2. Objetivos específicos

O1. Realizar un análisis exploratorio de datos (EDA) para comprender la estructura, distribución y calidad del conjunto de datos, identificando valores atípicos, inconsistencias y posibles sesgos.

O2. Implementar la función de distancia geodésica mediante la fórmula de *Haversine*, con el propósito de estimar la longitud real de cada trayecto a partir de las coordenadas de origen y destino.

O3. Aplicar criterios estadísticos para la detección y eliminación de *outliers* en las variables `fare_amount` y `distance_km`, garantizando la coherencia y representatividad del dataset filtrado.

O4. Dividir el conjunto de datos en muestras de entrenamiento y prueba bajo un esquema reproducible (`train/test = 80/20`), utilizando un `random_state` definido por el número de documento del estudiante, a fin de asegurar la replicabilidad de los resultados.

O5. Ajustar y evaluar modelos de regresión lineal (MCO) y regularizada (LASSO) para establecer una línea base de desempeño e interpretar los parámetros estimados en términos económicos y estadísticos.

O6. Entrenar modelos de ensamble (*Random Forest* y *Gradient Boosting*) para capturar relaciones no lineales entre las variables y comparar su capacidad predictiva frente a los modelos lineales.

O7. Diseñar y probar distintas arquitecturas de redes neuronales artificiales, variando el número de capas ocultas, neuronas y parámetros de entrenamiento, con el fin de explorar su potencial en la estimación de tarifas de viaje.

O8. Evaluar y comparar el desempeño global de todos los modelos mediante las métricas de error *Root Mean Squared Error* (RMSE) y *Mean Absolute Error* (MAE), seleccionando el modelo con mejor capacidad predictiva.

O9. Discutir los resultados obtenidos, analizando la relación entre interpretabilidad, complejidad computacional y precisión, así como las implicancias de los hallazgos para la toma de decisiones en entornos de *Business Intelligence*.

3. Marco Teórico

3.1. Inteligencia de Negocios y análisis predictivo

La *Inteligencia de Negocios* (BI) integra procesos, datos y herramientas analíticas para transformar información operativa en conocimiento accionable. En su

vertiente predictiva, BI incorpora modelos estadísticos y de *machine learning* para estimar variables de interés y apoyar la toma de decisiones bajo incertidumbre. En problemas de tarificación, el objetivo típico es modelar la relación entre variables explicativas (p. ej., distancia, tiempo, contexto geográfico) y la tarifa observada, buscando un compromiso entre precisión, interpretabilidad y costo computacional.

3.2. Modelado de datos y toma de decisiones

Desde la perspectiva de BI, el modelado cumple dos roles complementarios: (i) **descriptivo/explicativo**, al cuantificar efectos marginales y contrastar hipótesis; y (ii) **predictivo**, al minimizar el error en datos no vistos. En ambos casos, la calidad del dato (limpieza, tratamiento de atípicos, ingeniería de características) es determinante para evitar sesgos y mejorar la *generalización*. La evaluación fuera de muestra (*holdout* o validación cruzada) es el mecanismo estándar para estimar el rendimiento esperado.

3.3. Regresión lineal (MCO) e interpretación

El modelo lineal con Mínimos Cuadrados Ordinarios (MCO) asume

$$y = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon, \qquad \mathbb{E}[\varepsilon|\mathbf{x}] = 0,$$

y estima los parámetros resolviendo

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \big(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}\big)^2.$$

Sus ventajas son la **interpretabilidad** (coeficientes como efectos marginales) y la disponibilidad de inferencia clásica (errores estándar, pruebas t). Sus limitaciones aparecen ante relaciones no lineales, colinealidad y presencia de *outliers*.

3.4. Regularización LASSO

Para controlar sobreajuste y seleccionar variables, la regularización L_1 (LASSO) incorpora una penalización a la magnitud de los coeficientes:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \big(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}\big)^2 \; + \; \lambda \|\boldsymbol{\beta}\|_1 \right\},$$

donde $\lambda \geq 0$ controla la fuerza de la penalización. LASSO tiende a llevar a cero algunos coeficientes (parquedad del modelo) y requiere **escalado** de variables para una penalización equitativa. El hiperparámetro λ se selecciona típicamente por validación cruzada, optimizando una métrica como RMSE/MAE.

3.5. Modelos de ensamble: Random Forest y Gradient Boosting

Random Forest (RF). Combina múltiples árboles de decisión entrenados sobre *bootstraps* de los datos y subconjuntos aleatorios de variables. La predicción final promedia las predicciones individuales:

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}),$$

lo que reduce la varianza respecto de un árbol único y capta no linealidades e interacciones sin requerir especificarlas. Sus hiperparámetros clave incluyen la profundidad máxima, el número de árboles y los criterios de división.

Gradient Boosting (GB). Construye aditivamente un ensamble de árboles poco profundos, ajustando en cada etapa un nuevo árbol a los *residuales* del ensamble previo:

$$\hat{f}_0(\mathbf{x}) = \arg\min_{c} \sum_{i} L(y_i, c), \tag{1}$$

$$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1}(\mathbf{x}) + \nu\, h_m(\mathbf{x}). \tag{2}$$

donde h_m es el árbol "débil" de la etapa m, $L(\cdot)$ es la pérdida (p. ej., cuadrática) y $\nu \in (0, 1]$ el *learning rate*. GB suele lograr mayor precisión que RF con árboles más someros, a costa de más sensibilidad a hiperparámetros (número de etapas, profundidad, tasa de aprendizaje). Ambos métodos ofrecen medidas de *importancia de variables* y buen rendimiento con relaciones no lineales.

3.6. Redes neuronales densas (perceptrón multicapa)

Las redes densas aproximan funciones mediante composiciones de transformaciones afines y no lineales:

$$\hat{y} = f(\mathbf{x}; \boldsymbol{\theta}) \tag{3}$$

$$= \mathbf{W}_L\, \sigma\Big(\mathbf{W}_{L-1}\, \sigma\big(\cdots \sigma(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)\big) + \mathbf{b}_{L-1}\Big) + \mathbf{b}_L. \tag{4}$$

donde $\sigma(\cdot)$ es una activación (p. ej., ReLU) y $\boldsymbol{\theta}$ agrupa pesos y sesgos. Para regresión, la pérdida típica es el error cuadrático medio

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i; \boldsymbol{\theta})\big)^2,$$

optimizándose por descenso por gradiente (p. ej., Adam) con técnicas de regularización (early stopping, dropout, penalizaciones L_1/L_2). Su fortaleza reside en modelar relaciones altamente no lineales; su debilidad, en la menor interpretabilidad y en la sensibilidad a la configuración.

3.7. Sesgo–varianza, métricas y selección de modelo

El desempeño de generalización resulta del equilibrio **sesgo–varianza**. Modelos simples (lineales) tienden a alto sesgo y baja varianza; ensambles/NN reducen sesgo pero pueden incrementar varianza si se sobreajustan. Para seleccionar el modelo se usan métricas de error en datos de validación o prueba: *RMSE* y *MAE*.

Complementariamente, en modelos estadísticos se reportan medidas de ajuste (p. ej., R^2) e información (AIC/BIC) con fines comparativos.

3.8. Ventajas y limitaciones comparadas

- **OLS/LASSO**: alta interpretabilidad y base inferencial; limitados ante no linealidades fuertes o interacciones no modeladas.
- **RF/GB**: capturan no linealidades sin ingeniería exhaustiva; robustos y con buen rendimiento fuera de muestra; interpretabilidad media (importancias, *partial dependence*).
- **Redes densas**: máxima flexibilidad funcional y precisión competitiva; requieren mayor cuidado en tuning, escalado y control de sobreajuste; interpretabilidad reducida.

Nota. Este marco teórico sustenta la comparación empírica desarrollada en las secciones de Metodología y Resultados, justificando la inclusión de modelos lineales, de ensamble y neuronales para el problema de predicción de tarifas.

4. Metodología

La metodología implementada sigue un enfoque secuencial y reproducible de análisis de datos, alineado con las etapas clásicas de un proceso de *Inteligencia de Negocios*: recolección, depuración, exploración, modelado y evaluación. A continuación, se describen los pasos principales desarrollados en el estudio.

4.1. Descripción del dataset

El conjunto de datos empleado corresponde al *Uber NYC Dataset*, que recopila registros de viajes individuales realizados en la ciudad de Nueva York. Cada observación contiene información geográfica (coordenadas de origen y destino), temporal (fecha y hora del viaje) y económica (monto de la tarifa, `fare_amount`). Luego de la depuración inicial, el dataset final incluyó aproximadamente **144 000 registros válidos** y las variables esenciales para el modelado: `pickup_longitude`, `pickup_latitude`, `dropoff_longitude`, `dropoff_latitude`, y `fare_amount`.

El conjunto de datos original proporcionado por la plataforma *Uber* contiene información detallada de aproximadamente 200 000 viajes realizados en la ciudad de Nueva York. Cada registro incluye las coordenadas geográficas de origen y destino, el monto de la tarifa, y variables adicionales como la cantidad de pasajeros transportados.

Antes de aplicar cualquier transformación o filtrado, se efectuó un análisis descriptivo inicial con el propósito de identificar posibles inconsistencias, valores extremos y la distribución general de las variables. Los principales estadísticos se presentan en el **Cuadro 1**.

Como puede observarse, las variables geográficas presentan valores atípicos significativos (por ejemplo, longitudes fuera del rango esperado para Nueva York), lo que evidencia la necesidad de un proceso de limpieza y depuración antes de realizar el modelado. Asimismo, la variable `fare_amount` incluye valores negativos e improbables que fueron descartados en etapas posteriores. Este diagnóstico inicial permitió delimitar el dominio espacial de análisis y fundamentar las decisiones de filtrado implementadas en el preprocesamiento de los datos.

Cuadro 1. Estadísticas descriptivas del dataset original de viajes Uber (200,000 observaciones).

Variable	count	mean	std	min	25 %	75 %
fare_amount	200000	11.36	9.90	-52.00	7.00	12.50
pickup_longitude	200000	-72.53	11.14	-1340.68	-74.02	-73.97
pickup_latitude	200000	39.94	7.72	-74.02	40.73	40.77
dropoff_longitude	199999	13.12	6.79	-3356.09	-73.91	-73.96
dropoff_latitude	199999	39.92	6.79	-881.99	40.73	40.77
passenger_count	200000	1.68	1.39	0.00	1.00	2.00

El análisis exploratorio inicial también incluyó una revisión estadística de las variables geográficas del conjunto de datos original, con el objetivo de verificar la coherencia espacial de los registros. Se evaluaron los valores mínimos, máximos y medidas de tendencia central para las coordenadas de origen y destino.

Los resultados, presentados en el **Cuadro 2**, revelan la existencia de valores extremos que no corresponden a ubicaciones dentro del área metropolitana de Nueva York. Por ejemplo, se observaron longitudes inferiores a –1300 y latitudes mayores a 1600, lo cual evidencia errores de captura o codificación en el registro de coordenadas.

Cuadro 2. Rango estadístico de las coordenadas geográficas del dataset original.

Estadístico	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
count	200,000	200,000	199,999	199,999
mean	-72.53	39.94	-72.53	39.92
std	11.44	7.72	13.12	6.79
min	-1340.65	-74.02	-3356.67	-881.99
25 %	-73.99	40.73	-73.99	40.73
50 %	-73.98	40.75	-73.98	40.75
75 %	-73.97	40.77	-73.96	40.77
max	57.42	1644.42	1153.57	872.70

La dispersión observada en los valores de longitud y latitud respalda la necesidad de implementar un proceso de depuración geográfica, restringiendo el análisis a los rangos válidos para la ciudad de Nueva York (aproximadamente entre –74.5 y –73.5 en longitud, y entre 40.5 y 41.0 en latitud). Esta decisión permitió garantizar la validez espacial de los datos y mejorar la consistencia del cálculo posterior de distancias.

Una vez finalizada la depuración espacial y el cálculo de las distancias, se aplicó un filtrado adicional a la variable `fare_amount` con el fin de eliminar valores atípicos y registros con montos negativos o excesivamente altos que pudieran distorsionar los análisis posteriores. Este procedimiento permitió obtener una distribución más representativa de las tarifas reales dentro del servicio Uber en la ciudad de Nueva York.

El Cuadro 3 presenta un resumen estadístico de la variable `fare_amount` después del proceso de filtrado. Se observa que el número total de registros válidos se redujo a 193,301, con una tarifa máxima de USD 51.90. La media resultante (USD 10.72) y la mediana (USD 8.50) reflejan la predominancia de trayectos de corta y mediana distancia, en concordancia con los patrones espaciales analizados previamente.

Este ajuste garantiza una mayor estabilidad estadística y una distribución más homogénea para la variable dependiente utilizada en los modelos de regresión y aprendizaje automático. De esta forma, se minimi-

Cuadro 3. Estadísticas descriptivas de la variable `fare_amount` después del filtrado.

Descripción	Valor
Registros luego del filtrado	193,301
Tarifa máxima después del filtrado	USD 51.90
count	193,301
mean	10.72
std	7.80
min	0.01
25 %	6.00
50 % (mediana)	8.50
75 %	12.50
max	51.90

za la influencia de valores extremos y se preserva la representatividad del comportamiento tarifario típico del servicio.

Finalmente, se examinó la distribución de la variable dependiente `fare_amount` tras el proceso de depuración. Esta revisión gráfica permite comprobar el efecto del filtrado sobre la forma de la distribución y la coherencia de los valores restantes con las tarifas observadas en trayectos reales.

La Figura 1 muestra la distribución de frecuencias de `fare_amount` en el dataset filtrado. Se observa una clara asimetría hacia la derecha, característica de variables de tipo monetario, donde la mayoría de los registros se concentran entre 5 y 15 USD, correspondientes a viajes de corta o media distancia. Los valores por encima de 30 USD son escasos y reflejan desplazamientos más largos o hacia zonas periféricas.

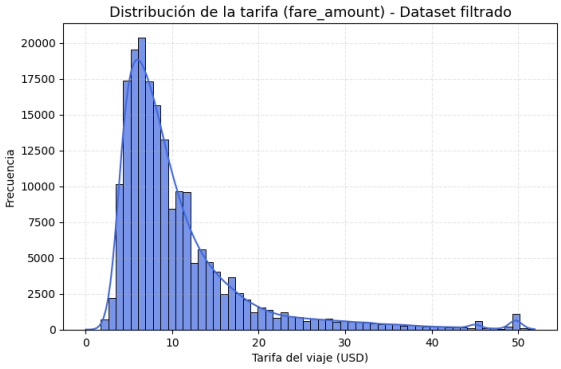

Figura 1. Distribución de la variable `fare_amount` (tarifa del viaje en USD) después del filtrado del dataset. Se observa una distribución asimétrica hacia la derecha con la mayoría de los valores concentrados entre 5 y 15 USD.

La forma resultante de la distribución confirma la efectividad del proceso de filtrado descrito en el **Cuadro 3**, evidenciando que el conjunto de datos conserva únicamente valores plausibles y representativos del comportamiento tarifario típico. Este análisis constituye el paso final antes de la exploración descriptiva y la modelización predictiva presentadas en las siguientes secciones.

Con el propósito de examinar de manera complementaria la dispersión y presencia de valores extremos en la variable `fare_amount`, se elaboró un diagrama de caja (*boxplot*) sobre el conjunto de datos filtrado. Esta representación permite identificar de forma sintética los cuartiles, el rango intercuartílico y los posibles valores atípicos que aún persisten tras el proceso de depuración.

4.2. Procesamiento y limpieza de datos

Se aplicaron filtros geográficos y estadísticos con el fin de eliminar registros inconsistentes o extremos:

- **Depuración de coordenadas:** se eliminaron observaciones fuera de los rangos válidos de latitud y longitud global, y se acotó el análisis al área metropolitana de Nueva York (40.5°–41.0° N, 74.5°–73.5° O).
- **Cálculo de distancias:** se creó la variable `distance_km` a partir de la fórmula del semiverseno (*Haversine*), que estima la distancia geodésica entre los puntos de origen y destino:

  donde $R = 6371$ km es el radio medio de la Tierra.
- **Eliminación de valores atípicos:** se aplicó el método del rango intercuartílico (IQR) para filtrar observaciones extremas en las variables `fare_amount` y `distance_km`. Este procedimiento redujo el ruido y mejoró la homogeneidad de las distribuciones.

Una vez aplicado el filtrado geográfico y verificados los rangos válidos de coordenadas, se procedió a calcular la distancia geodésica entre los puntos de origen y destino mediante la fórmula del semiverseno. Este paso permitió generar la variable `distance_km`, la cual constituye el principal predictor del modelo de tarifas.

Durante este proceso se identificaron y eliminaron observaciones con valores nulos, inconsistentes o fuera del dominio espacial de la ciudad de Nueva York. En total se descartaron 4 353 registros, equivalentes al 2.18 % del conjunto inicial, conservando 195 647 observaciones válidas para el análisis posterior. El resumen detallado del proceso de depuración y de la nueva variable creada se presenta en el **Cuadro 4**.

Cuadro 4. Resumen del proceso de depuración y creación de la variable `distance_km`.

Descripción	Valor
Registros originales	200,000
Registros válidos	195,647
Registros eliminados	4,353 (2.18 %)
Variable analizada	**`distance_km`**
count	195,647
mean	3.31
std	3.58
min	0.00
25 %	1.26
50 % (mediana)	2.16
75 %	3.91
max	41.23

Como se observa, la distancia media recorrida fue de aproximadamente 3.3 km, con una dispersión de 3.6 km. El rango intercuartílico (1.26–3.91 km) sugiere que la mayoría de los viajes corresponden a trayectos urbanos de corta o media distancia, mientras que los valores máximos reflejan desplazamientos atípicos hacia zonas periféricas o aeropuertos. Esta información

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Figura 2. Fórmula del semiverseno (Haversine) para el cálculo de distancias geodésicas.

resulta clave para interpretar las diferencias tarifarias y validar posteriormente la coherencia de los modelos predictivos.

Durante el proceso de depuración de los datos, se realizó una evaluación estadística de los valores atípicos presentes en las variables principales `fare_amount` y `distance_km`. Para ello, se aplicó el criterio del rango intercuartílico (IQR, por sus siglas en inglés), que permite identificar observaciones que se encuentran significativamente alejadas del rango central de la distribución.

El **Cuadro 5** resume los resultados de esta detección, indicando la cantidad y el porcentaje de registros que se ubicaron fuera de los límites teóricos establecidos por el IQR. Se observa que aproximadamente un 5.6 % de las observaciones en ambas variables se consideraron atípicas, lo que justifica la necesidad de aplicar filtros de limpieza para garantizar la consistencia del conjunto de datos.

Cuadro 5. Detección de valores atípicos mediante el rango intercuartílico (IQR).

Descripción	`fare_amount`	`distance_km`
Observaciones fuera del rango IQR	10,966	10,807
Porcentaje del total	5.67 %	5.59 %
Rango teórico (IQR)	[-7.00, 25.50]	[-3.89, 8.95]

Estos resultados confirman la presencia de un número limitado de registros extremos que, aunque representan una fracción pequeña del total, podrían distorsionar las estimaciones de tendencia central y dispersión. Por ello, se procedió a su exclusión en el dataset filtrado, manteniendo un balance adecuado entre precisión estadística y representatividad de la muestra.

Con el fin de evaluar el impacto del filtrado sobre la estructura del dataset, se aplicó el método del rango intercuartílico (IQR) a las variables `fare_amount` y `distance_km`. Este procedimiento permitió eliminar los valores extremos que excedían los límites establecidos en función de los cuartiles $Q1$ y $Q3$, preservando únicamente los registros comprendidos dentro del rango teórico definido por $Q1 - 2{,}0 \times IQR$ y $Q3 + 2{,}0 \times IQR$.

El **Cuadro 6** resume los resultados del proceso de filtrado y su efecto sobre la correlación entre ambas variables. Se observa que el filtrado afectó aproximadamente al 7 % del total de observaciones, reduciendo la muestra a 179,855 registros válidos. A pesar de esta depuración, la correlación de Pearson entre `fare_amount` y `distance_km` se mantuvo alta (0.821), lo que confirma la relación positiva y consistente entre ambas variables incluso tras la eliminación de valores atípicos.

Este resultado evidencia que la eliminación de valores atípicos mejora la consistencia estadística del conjunto de datos sin alterar de manera significativa las relaciones subyacentes. Así, el dataset depurado constituye una base más robusta para los análisis exploratorios y los posteriores modelos predictivos que se presentan en la siguiente sección.

Cuadro 6. Resumen del filtrado mediante el método IQR y su efecto sobre la correlación.

Descripción	`fare_amount`	`distance_km`
Límite inferior aplicado	0.00	0.00
Límite superior aplicado	25.50	8.95
Q1 (25 %)	6.00	1.25
Q3 (75 %)	12.50	3.82
IQR	6.50	2.57
Registros antes del filtrado	193,301	
Registros después del filtrado	179,855	
Registros removidos (%)	13,446 (6.96 %)	
Correlación Pearson (`fare_amount` vs `distance_km`)		
Antes del filtrado	0.875	
Después del filtrado	0.821	

Con el objetivo de cuantificar los efectos del proceso de filtrado mediante el método IQR sobre la distribución de las tarifas (`fare_amount`), se realizó una comparación estadística antes y después de la eliminación de valores atípicos. El **Cuadro 7** resume los principales estadísticos descriptivos, evidenciando una reducción en la dispersión de los datos y una ligera disminución en la media y la mediana.

En términos prácticos, tras el filtrado se eliminó el 6.9 % de los registros más extremos, reduciendo el valor máximo de 51.9 USD a 25.5 USD, lo que implica una corrección significativa de los outliers sin alterar de forma sustancial la estructura general de la distribución. Asimismo, el desvío estándar pasó de 7.80 a 4.36 USD, indicando una mayor homogeneidad en los valores retenidos.

Estos resultados confirman que el filtrado aplicado mejoró la calidad del conjunto de datos, atenuando el efecto de observaciones anómalas y preparando la base para un análisis exploratorio más robusto, presentado en la siguiente sección.

Cuadro 7. Comparación estadística de la variable `fare_amount` antes y después del filtrado IQR.

Estadístico	Antes del filtrado	Después del filtrado
count	193,301	179,855
mean	10.72	9.06
std	7.80	4.36
min	0.01	0.01
25 %	6.00	5.70
50 % (mediana)	8.50	8.00
75 %	12.50	11.30
max	51.90	25.50

Para ilustrar gráficamente el efecto del filtrado mediante el método IQR, se compararon las distribuciones de la variable `fare_amount` antes y después de la depuración estadística. En la **Figura 3** se observa cómo la eliminación de valores atípicos reduce notablemente la dispersión y concentra la mayor parte de las observaciones en el rango comprendido entre 5 y 15 USD.

El panel izquierdo muestra la distribución original, ca-

racterizada por una cola larga hacia la derecha que evidencia la presencia de tarifas atípicamente elevadas. Tras el filtrado (panel derecho), la distribución adopta una forma más simétrica y compacta, lo que indica una mejora sustancial en la consistencia del conjunto de datos. Este proceso no solo reduce el impacto de valores extremos, sino que también fortalece la fiabilidad de los modelos predictivos que se desarrollarán posteriormente.

Como se observa en la **Figura 4**, la mayor concentración de tarifas se encuentra entre 5 y 15 USD, lo que coincide con los resultados del histograma presentado anteriormente (**Figura 1**). No obstante, se aprecian algunos valores aislados hacia el extremo superior, asociados a trayectos de larga distancia o con condiciones tarifarias especiales, como viajes hacia aeropuertos.

El análisis gráfico confirma que el procedimiento de limpieza permitió eliminar valores anómalos sin afectar la estructura general de la variable dependiente. A su vez, evidencia que el rango intercuartílico (IQR) mantiene una distribución consistente y adecuada para los posteriores procesos de modelado estadístico y predictivo.

El efecto del proceso de depuración mediante el método IQR también puede observarse en la estructura de correlaciones entre las principales variables del conjunto de datos. En la **Figura 5** se comparan las matrices de correlación de Pearson calculadas antes y después del filtrado.

Se aprecia que, tras la eliminación de valores atípicos extremos, la relación entre `fare_amount` y `distance_km` disminuye levemente (de $r \approx 0{,}88$ a $r \approx 0{,}82$), lo que indica una menor influencia de observaciones anómalas sobre la medida de asociación lineal. Este resultado confirma que el filtrado IQR no altera de manera sustancial la estructura general de dependencias del dataset, pero sí contribuye a obtener estimaciones más estables y representativas para el posterior modelado predictivo.

4.3. Análisis exploratorio de datos (EDA)

El EDA permitió caracterizar la estructura y el comportamiento de las variables. Se analizaron distribuciones, correlaciones y relaciones espaciales, identificando patrones relevantes:

- Mapas de dispersión, densidad y hexágonos (*hexbin*) para los puntos de recogida y destino.
- Histogramas y diagramas de dispersión entre `fare_amount` y `distance_km`.
- Matrices de correlación de Pearson y Spearman.

Estos resultados se sintetizan en las **Figuras 6–9**, donde se evidencian las zonas de mayor concentración de viajes y la fuerte correlación positiva entre distancia y tarifa.

Una vez concluido el proceso de depuración y validación de coordenadas, se generó una representación geográfica de los puntos de origen (*pickup*) y destino (*dropoff*) de los viajes registrados. Esta visualización permite verificar de manera intuitiva la distribución espacial de los datos y comprobar que los registros conservados tras el filtrado corresponden efectivamente a trayectos dentro del área urbana de Nueva York.

En la **Figura 6** se observa una alta densidad de viajes en las zonas centrales de Manhattan, con extensiones significativas hacia Brooklyn, Queens y los principales corredores de acceso a los aeropuertos. La mayor concentración de puntos en el sector sur de Manhattan refleja la dinámica típica de movilidad asociada a áreas comerciales y turísticas, mientras que la presencia de trayectos más dispersos hacia la periferia sugiere desplazamientos intermunicipales y hacia terminales de transporte.

El mapa confirma que, tras la aplicación de los filtros geográficos, el conjunto de datos mantiene una cobertura espacial coherente con los límites urbanos del estudio. Esta verificación visual constituye un paso esencial previo al cálculo de distancias y a la posterior estimación de modelos predictivos.

Para caracterizar la concentración espacial de los viajes y distinguir las zonas de mayor actividad, se generó un mapa de densidad mediante celdas hexagonales (*Hexbin*) sobre la proyección geográfica de Nueva York. Este tipo de visualización permite representar la frecuencia de puntos de recogida (*pickups*) en una escala logarítmica, lo que facilita la identificación de patrones urbanos sin perder detalle en áreas con alta densidad.

Como se aprecia en la **Figura 7**, los resultados muestran una marcada concentración de viajes en la isla de Manhattan, especialmente en el sector sur y medio, donde se ubican los principales polos comerciales y turísticos de la ciudad. También se observan focos importantes de actividad en los distritos de Brooklyn y Queens, asociados a zonas residenciales y a los accesos hacia el aeropuerto JFK.

Esta representación espacial confirma la fuerte centralización del servicio en torno a Manhattan y revela la estructura radial de los desplazamientos urbanos. El uso de celdas *Hexbin* resulta especialmente útil en contextos de alta densidad de datos, ya que mitiga el solapamiento de puntos y ofrece una visión más clara de la distribución espacial de la demanda.

Con el fin de cuantificar las relaciones entre las variables principales del conjunto de datos, se calculó la matriz de correlaciones de Pearson. Este análisis permite identificar asociaciones lineales y verificar la coherencia entre las medidas geográficas, la distancia recorrida y el valor de la tarifa.

Los resultados se presentan en la **Figura 8**, donde se observa una correlación positiva muy alta ($r = 0{,}86$) entre `fare_amount` y `distance_km`, lo cual confirma que la distancia recorrida constituye el principal determinante del costo del viaje. Asimismo, se evidencian correlaciones moderadas entre las coordenadas de origen y destino, reflejando la estructura espacial del servicio dentro del área metropolitana de Nueva York.

En conjunto, estos resultados sustentan la selección de `distance_km` como variable predictora principal en los modelos de regresión lineal y de aprendizaje automático aplicados en las etapas posteriores. Además, la baja correlación del número de pasajeros con la tarifa sugiere que su efecto en el precio es marginal en comparación con la influencia de la distancia o la ubicación geográfica.

Con el objetivo de complementar el análisis de correlaciones lineales, se aplicó el coeficiente de Spearman, que permite evaluar relaciones monótonas entre variables sin asumir linealidad. Este enfoque resulta especialmente útil en contextos donde las variables pueden

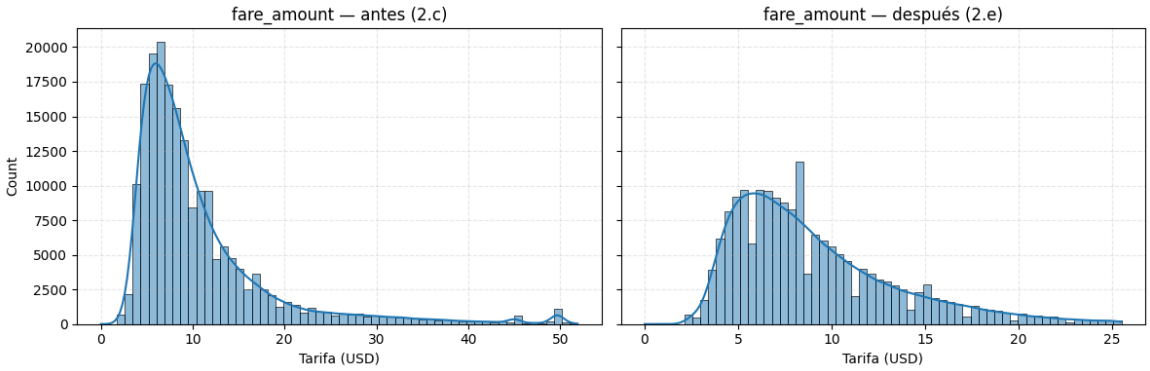

Figura 3. Comparación de la distribución de la variable `fare_amount` antes y después del filtrado IQR. Se observa una reducción significativa de valores extremos y una distribución más concentrada tras la depuración.

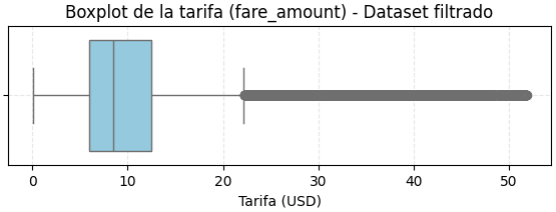

Figura 4. Diagrama de caja (*boxplot*) de la variable `fare_amount` tras el filtrado del dataset. Se observa una concentración principal entre 5 y 15 USD, con la presencia de valores atípicos hacia el extremo superior.

presentar distribuciones no normales o efectos de saturación, como ocurre en los datos de transporte urbano.

La **Figura 9** muestra la matriz de correlaciones obtenida mediante este método. Los resultados confirman la existencia de una relación monótona positiva muy fuerte ($\rho = 0{,}85$) entre `fare_amount` y `distance_km`, en concordancia con los hallazgos del análisis de Pearson (**Figura 8**). Asimismo, se observan asociaciones moderadas entre las coordenadas geográficas de origen y destino, lo que refleja la consistencia espacial de los registros dentro del área metropolitana.

La comparación entre ambas métricas evidencia que la relación entre la tarifa y la distancia recorrida es robusta tanto bajo supuestos lineales como no lineales. Este resultado respalda la selección de `distance_km` como predictor fundamental en los modelos de regresión y aprendizaje automático que se presentan en la siguiente sección.

4.4. Partición de datos

Para evaluar la capacidad de generalización de los modelos, el conjunto filtrado se dividió en dos subconjuntos: **80 % de entrenamiento** y **20 % de prueba**, utilizando el número de documento del estudiante como `random_state` para asegurar reproducibilidad. El conjunto de entrenamiento se empleó para el ajuste y validación cruzada de los modelos, mientras que el conjunto de prueba se reservó exclusivamente para la evaluación final.

Una vez completado el proceso de depuración y filtrado de los datos, el siguiente paso consistió en dividir el conjunto final en subconjuntos de entrenamiento y prueba. Esta partición garantiza la capacidad de generalización de los modelos y permite evaluar su desempeño en datos no vistos durante la etapa de aprendizaje.

El procedimiento se realizó mediante una división estratificada aleatoria utilizando la función `train_test_split` de `scikit-learn`, manteniendo un estado aleatorio fijo (`random_state = 12345678`) para asegurar la reproducibilidad de los resultados. La proporción elegida fue del 80 % para entrenamiento y del 20 % para prueba, siguiendo las prácticas estándar en modelado predictivo.

El **Cuadro 8** presenta el resumen de esta partición, indicando el tamaño final de cada subconjunto. Con ello, el dataset quedó conformado por 143,884 observaciones destinadas al ajuste de los modelos y 35,971 para su evaluación, garantizando una muestra suficiente en ambas fases del análisis.

Cuadro 8. Partición del conjunto de datos para entrenamiento y prueba.

Descripción	Valor
Tamaño total del dataset	179,855 observaciones
Proporción de entrenamiento	80 % (143,884 observaciones)
Proporción de prueba	20 % (35,971 observaciones)
Random state utilizado	12345678

Esta división constituye la base sobre la cual se entrenaron los distintos modelos predictivos abordados en la siguiente subsección, permitiendo comparar su desempeño bajo condiciones controladas y replicables.

4.5. Modelado predictivo

Se implementaron cinco enfoques de modelado supervisado:

a) **Regresión lineal (MCO):** modelo base que estima el efecto promedio de la distancia sobre la tarifa, útil para interpretación económica.

b) **Regresión LASSO:** versión regularizada del modelo lineal que incorpora penalización L_1 para prevenir sobreajuste y controlar la magnitud de los coeficientes.

c) **Random Forest:** ensamble de árboles de decisión entrenados sobre subconjuntos aleatorios de datos, que reduce la varianza y captura relaciones no lineales.

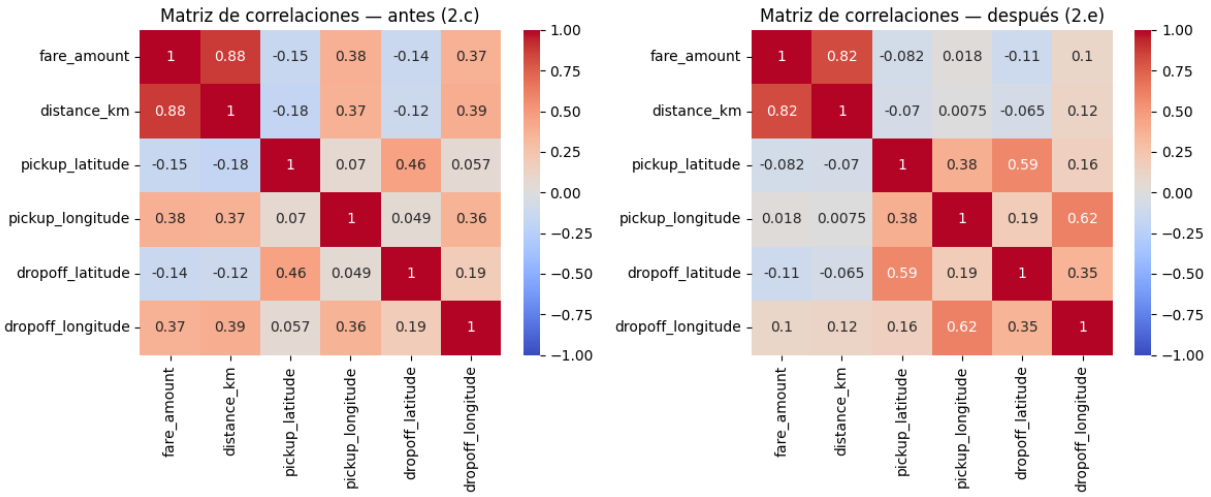

Figura 5. Comparación de las matrices de correlación de Pearson antes y después del filtrado IQR. Se aprecia una ligera reducción en la fuerza de las correlaciones, especialmente entre `fare_amount` y `distance_km`, lo que refleja la eliminación de valores atípicos extremos.

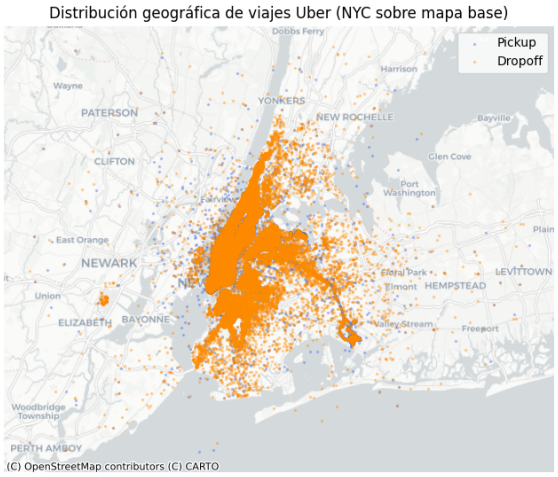

Figura 6. Distribución geográfica de los puntos de origen (*pickup*) y destino (*dropoff*) de los viajes Uber en la ciudad de Nueva York.

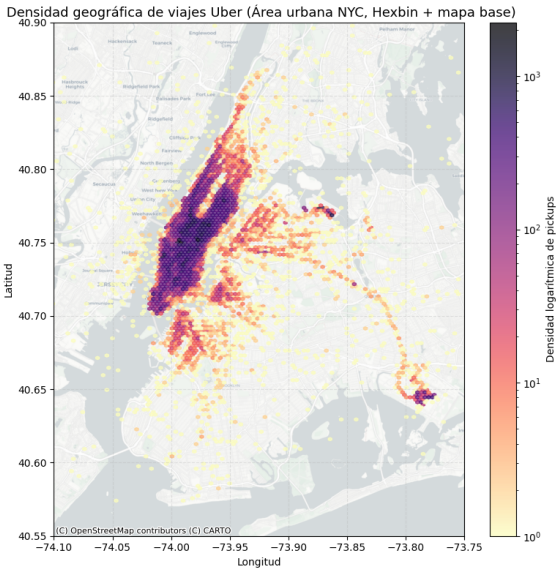

Figura 7. Mapa de densidad de puntos de recogida (*pickups*) de viajes Uber en el área urbana de Nueva York, utilizando celdas hexagonales (*Hexbin*) y escala logarítmica.

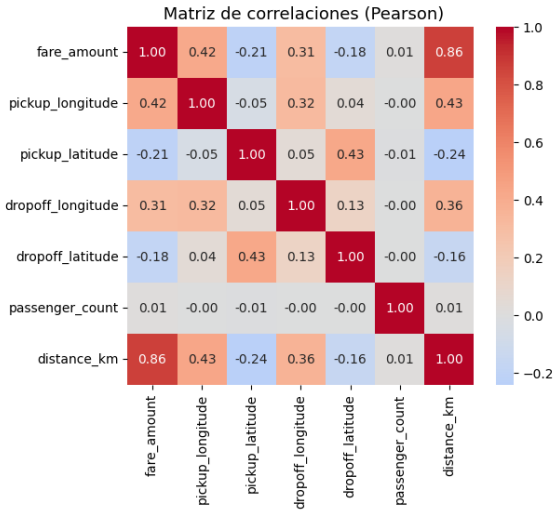

Figura 8. Matriz de correlaciones de Pearson entre las variables principales del dataset de viajes Uber. Se destaca la fuerte correlación positiva entre `fare_amount` y `distance_km`.

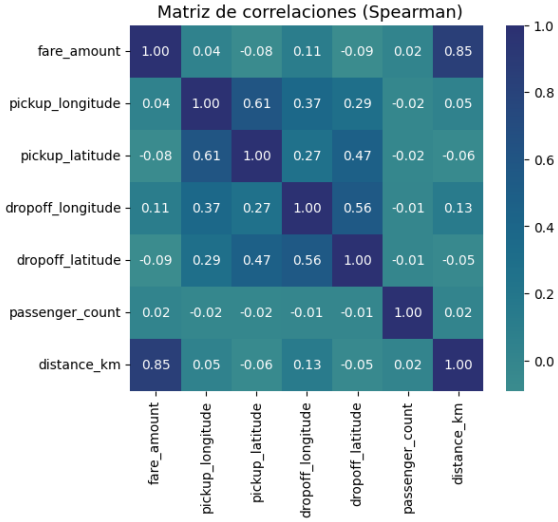

Figura 9. Matriz de correlaciones de Spearman entre las variables principales del dataset de viajes Uber. Se observa una fuerte correlación monótona positiva entre `fare_amount` y `distance_km`.

d) **Gradient Boosting:** modelo secuencial que ajusta sucesivos árboles a los residuales del anterior, optimizando la precisión mediante el parámetro de tasa de aprendizaje ν.

e) **Redes neuronales densas:** tres arquitecturas del tipo *feedforward* con diferente número de capas ocultas, activación ReLU y optimizador Adam, diseñadas para aproximar relaciones complejas entre variables.

4.6. Métricas de evaluación

El desempeño de los modelos se midió con las métricas estándar de error:

- **RMSE (Root Mean Squared Error):** raíz del error cuadrático medio, penaliza errores grandes y refleja la precisión general del modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

- **MAE (Mean Absolute Error):** promedio de los errores absolutos, más robusto frente a valores extremos.

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

- **Coeficiente de determinación (R^2):** proporción de la variabilidad explicada por el modelo respecto a la variabilidad total.

Estas métricas se calcularon tanto en la fase de validación como en el conjunto de prueba, permitiendo comparar la efectividad y generalización de cada enfoque.

4.7. Reproducibilidad

Todo el proceso fue desarrollado en `Python 3.10` utilizando las librerías `pandas`, `numpy`, `matplotlib`, `seaborn`, `scikit-learn` y `keras`. Los scripts y notebooks empleados se encuentran debidamente documentados, garantizando la replicabilidad de los resultados y la trazabilidad de cada etapa analítica.

5. Resultados

5.1. Relación entre variables continuas (Tarifa vs. Distancia)

Tras el análisis individual de las variables `fare_amount` y `distance_km`, resulta pertinente examinar su relación conjunta con el fin de identificar patrones de dependencia entre la distancia recorrida y el valor de la tarifa. La **Figura 10** presenta la nube de puntos correspondiente al dataset filtrado, junto con la línea de tendencia ajustada mediante un modelo lineal simple.

Como puede observarse, existe una marcada relación positiva entre ambas variables: a medida que la distancia aumenta, la tarifa también lo hace de forma casi proporcional. Esta tendencia confirma los resultados de correlación obtenidos previamente en las **Figuras 8 y 9**, donde se evidenció una correlación fuerte y significativa entre ambas magnitudes.

El patrón lineal visible en el gráfico sugiere que el modelo de precios de los viajes Uber en Nueva York man-

tiene una estructura tarifaria proporcional a la distancia recorrida, con ligeras desviaciones explicables por factores como el tiempo de espera, el tráfico o las tarifas dinámicas aplicadas en horarios de alta demanda.

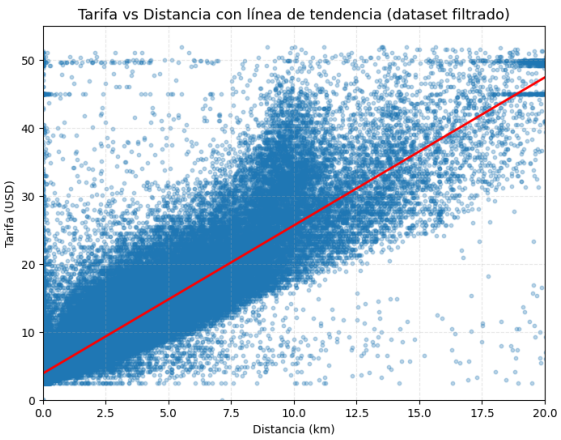

Figura 10. Relación entre la distancia recorrida (`distance_km`) y la tarifa del viaje (`fare_amount`) en el dataset filtrado. Se observa una fuerte correlación positiva con una tendencia aproximadamente lineal.

5.2. Resultados de los modelos lineales

El primer modelo estimado corresponde a una regresión lineal ordinaria (OLS) que relaciona la tarifa del viaje (`fare_amount`) con la distancia recorrida (`distance_km`). Este enfoque permite evaluar si la distancia constituye un predictor significativo y lineal de la tarifa, bajo el supuesto de homocedasticidad y normalidad de los errores.

Como se muestra en el **Cuadro 9**, los resultados del modelo evidencian un ajuste considerable, con un coeficiente de determinación $R^2 = 0{,}674$, lo que indica que aproximadamente el 67 % de la variabilidad en el valor de la tarifa puede explicarse por la distancia recorrida. El coeficiente asociado a la variable `distance_km` es positivo ($\beta = 2{,}022$) y altamente significativo ($p < 0{,}001$), confirmando una relación lineal directa entre ambas variables: por cada kilómetro adicional recorrido, la tarifa promedio aumenta en torno a 2.02 USD.

El término constante (4.01 USD) refleja el costo base del servicio, independiente de la distancia, coherente con la estructura tarifaria típica de servicios de transporte urbano. En conjunto, los indicadores de información (AIC y BIC) muestran valores similares y coherentes con un modelo parsimonioso y bien especificado. Estos resultados sientan la base para contrastar el desempeño de los modelos regularizados y no lineales que se presentan en las siguientes subsecciones.

Cuadro 9. Resultados del modelo de regresión lineal (OLS) para `fare_amount` en función de `distance_km`.

Variable	Coef.	Std. Err.	t	P> |t|	[0.025]	[0.975]
Constante	4.0053	0.011	352.27	0.000	3.983	4.028
`distance_km`	2.0220	0.004	545.42	0.000	2.015	2.029
Indicadores del modelo:						
R^2	0.674					
AIC	6.714e+05					
BIC	6.714e+05					
No. de observaciones	143,884					

A fin de complementar el modelo OLS y controlar posibles efectos de sobreajuste o multicolinealidad, se aplicó un modelo de regresión LASSO (Least Absolute Shrinkage and Selection Operator), el cual introduce un término de penalización L_1 sobre la magnitud de los coeficientes. Este enfoque permite reducir la varianza del modelo al mismo tiempo que mantiene la interpretabilidad, forzando a cero aquellos coeficientes con baja relevancia explicativa.

Los resultados del modelo ajustado se presentan en el **Cuadro 10**. El parámetro de regularización óptimo obtenido mediante validación cruzada fue $\lambda = 0{,}0001$, lo que indica una penalización relativamente baja, coherente con la simplicidad del modelo y la fuerte relación entre las variables. El coeficiente estimado para `distance_km` fue de 3.5869, mientras que el intercepto alcanzó un valor de 9.0641.

Estos resultados muestran un patrón similar al observado en el modelo OLS, aunque con una ligera reducción en la pendiente y un aumento en el intercepto. La penalización impuesta por LASSO actúa suavizando los coeficientes, lo que contribuye a mejorar la estabilidad del modelo frente a posibles valores atípicos o colinealidades residuales. En términos generales, el modelo LASSO conserva la relación positiva entre la distancia recorrida y la tarifa del viaje, confirmando la solidez del vínculo entre ambas variables y preparando el terreno para la comparación con los modelos de ensamble no lineales presentados a continuación.

Cuadro 10. Parámetros del modelo LASSO ajustado para `fare_amount` en función de `distance_km`.

Parámetro	Valor
Alpha óptimo (λ)	0.0001
Coeficiente estimado	3.5869
Intercepto	9.0641

5.3. Resultados de los modelos de ensamble

Con el fin de capturar posibles relaciones no lineales entre la distancia recorrida y la tarifa del viaje, se aplicaron modelos de tipo *ensemble*, comenzando con el algoritmo *Random Forest*. Este método combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos y de las variables, con el objetivo de reducir la varianza del modelo y mejorar su capacidad de generalización frente a los modelos lineales tradicionales.

El entrenamiento del modelo se realizó a partir de un submuestreo de 30,000 observaciones, dividido en un 80 % para entrenamiento y un 20 % para prueba. Se exploraron 12 combinaciones de hiperparámetros mediante búsqueda en cuadrícula, optimizando las métricas de error cuadrático medio (RMSE) y error absoluto medio (MAE). Los mejores valores obtenidos se resumen en el **Cuadro 11**, con una profundidad máxima de los árboles de 10, un mínimo de 5 observaciones por división y 100 estimadores en total.

En términos de desempeño, el modelo alcanzó un RMSE de 2.4844 y un MAE de 1.7472, evidenciando una mejora sustancial en la capacidad predictiva frente al modelo OLS. Esta reducción en los errores refleja la habilidad del *Random Forest* para capturar patrones no lineales y variaciones locales en la relación entre

`fare_amount` y `distance_km`. Además, el modelo demostró una buena estabilidad en las predicciones del conjunto de prueba, confirmando su solidez y robustez frente a valores atípicos o ruido en los datos.

Cuadro 11. Resultados del modelo *Random Forest* aplicado a la predicción de `fare_amount`.

Descripción	Valor
Tamaño del dataset (submuestreo)	30,000 observaciones
Conjunto de entrenamiento	24,000 registros (80 %)
Conjunto de prueba	6,000 registros (20 %)
Número total de combinaciones evaluadas	12
Mejores hiperparámetros encontrados:	
`max_depth`	10
`min_samples_split`	5
`n_estimators`	100
Evaluación del modelo:	
RMSE	2.4844
MAE	1.7472

Además del resumen cuantitativo presentado en el **Cuadro 11**, se analizó la relación entre los valores reales y los valores predichos por el modelo *Random Forest*, representada en la **Figura 11**. Se observa una clara tendencia lineal positiva entre ambas variables, lo que confirma la capacidad del modelo para reproducir el comportamiento general de las tarifas de viaje. La dispersión de los puntos en torno a la diagonal es baja, lo que indica una adecuada precisión y un nivel de error reducido en la mayoría de las predicciones.

Este patrón de ajuste demuestra que el modelo logra capturar de manera efectiva la relación entre la distancia recorrida y la tarifa (`fare_amount`), manteniendo un equilibrio entre flexibilidad y generalización. En particular, la mayor densidad de puntos alrededor de la línea de identidad sugiere un rendimiento estable en el rango intermedio de valores, mientras que las desviaciones más amplias en los extremos pueden atribuirse a variaciones propias del sistema tarifario o a trayectos con condiciones atípicas. En conjunto, la **Figura 11** refuerza la solidez del modelo *Random Forest* y su idoneidad para la predicción de tarifas en entornos reales.

El segundo modelo de ensamble evaluado corresponde al algoritmo *Gradient Boosting*, una técnica basada en la combinación secuencial de múltiples árboles de decisión de baja profundidad, donde cada árbol intenta corregir los errores residuales del modelo anterior. Este enfoque permite capturar patrones no lineales con mayor precisión, a costa de un incremento moderado en la complejidad computacional.

El entrenamiento del modelo se realizó sobre un submuestreo de 30,000 observaciones, distribuidas en proporciones de 80 % para entrenamiento y 20 % para prueba. A diferencia del *Random Forest*, que entrena árboles de manera independiente, el *Gradient Boosting* los construye de forma iterativa, ponderando las observaciones mal predichas en cada ciclo. Para optimizar su desempeño, se evaluaron 24 combinaciones de hiperparámetros mediante búsqueda en cuadrícula, obteniéndose los valores óptimos presentados en el **Cuadro 12**.

Los mejores resultados se alcanzaron con una tasa de aprendizaje (`learning_rate`) de 0.05, una profundidad máxima de 3 niveles y 100 estimadores. Bajo esta configuración, el modelo logró un RMSE de 2.4583 y un MAE de 1.7323, valores ligeramente inferiores a los obtenidos con el *Random Forest*, lo que refleja una leve mejora en la precisión predictiva.

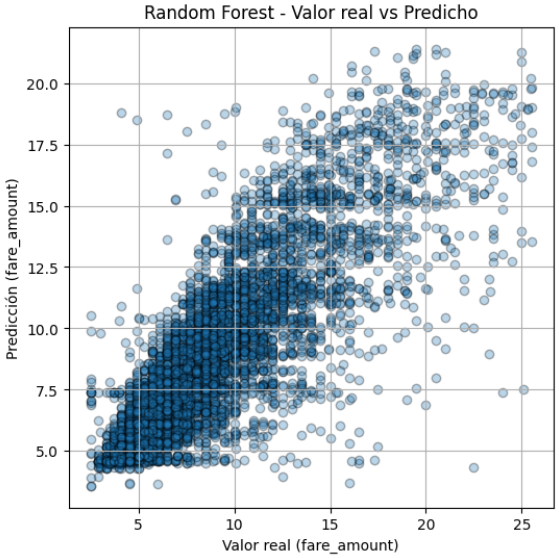

Figura 11. Relación entre valores reales y predichos del modelo *Random Forest* para la variable `fare_amount`. Se observa una tendencia lineal positiva, indicando una buena capacidad predictiva del modelo con baja dispersión.

En conjunto, los resultados confirman que el *Gradient Boosting* presenta un excelente equilibrio entre sesgo y varianza, ofreciendo predicciones estables y un ajuste más fino en las regiones de alta densidad de datos. Su capacidad para reducir gradualmente los errores residuales se traduce en una modelización más precisa del comportamiento de la variable `fare_amount`, especialmente en trayectos de corta y media distancia. Este desempeño se visualiza en la **Figura 12**, donde se observa una fuerte alineación entre los valores reales y los estimados, con baja dispersión en torno a la diagonal principal, lo que reafirma la calidad del ajuste obtenido.

Cuadro 12. Resultados del modelo *Gradient Boosting* aplicado a la predicción de `fare_amount`.

Descripción	Valor
Tamaño del dataset (submuestreo)	30,000 observaciones
Conjunto de entrenamiento	24,000 registros (80 %)
Conjunto de prueba	6,000 registros (20 %)
Número total de combinaciones evaluadas	24
Mejores hiperparámetros encontrados:	
`learning_rate`	0.05
`max_depth`	3
`n_estimators`	100
Evaluación del modelo:	
RMSE	2.4583
MAE	1.7323

Complementando la información resumida en el **Cuadro 12**, la **Figura 12** muestra la relación entre los valores reales y los valores predichos por el modelo *Gradient Boosting*. La nube de puntos evidencia una clara tendencia lineal positiva, donde la mayoría de las observaciones se alinean en torno a la diagonal, lo que indica un ajuste adecuado entre las predicciones del modelo y los valores observados de `fare_amount`.

Se observa además una ligera dispersión en los valores

más altos de la tarifa, aunque sin desviaciones significativas, lo que sugiere que el modelo mantiene una buena capacidad de generalización incluso en casos menos frecuentes. Este comportamiento es coherente con los indicadores de error obtenidos (RMSE y MAE), que demuestran una leve mejora respecto al modelo *Random Forest*. En términos cualitativos, el *Gradient Boosting* logra un equilibrio más preciso entre sesgo y varianza, permitiendo capturar patrones no lineales en los datos sin sobreajustarse.

En conjunto, la **Figura 12** confirma la consistencia del modelo y su robustez predictiva, consolidando su posición como una alternativa eficaz frente al *Random Forest* en la estimación de tarifas de viaje.

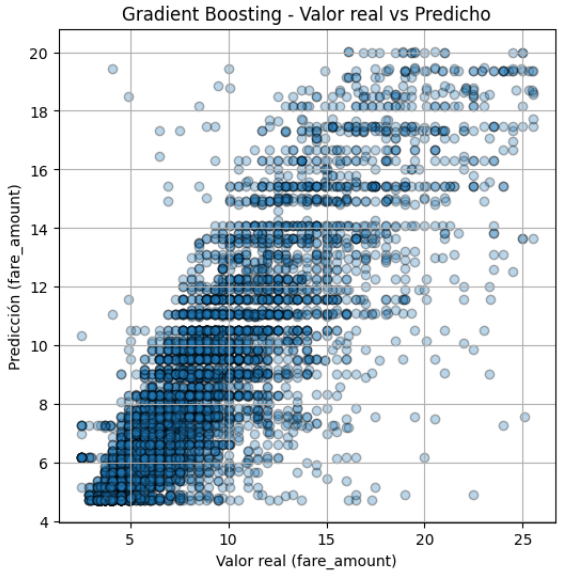

Figura 12. Relación entre valores reales y predichos del modelo *Gradient Boosting* para la variable `fare_amount`. Se observa un ajuste consistente con ligera dispersión, indicando un desempeño robusto y comparable al modelo *Random Forest*.

5.4. Resultados de las redes neuronales

Como complemento a los modelos basados en árboles de decisión, se implementaron tres arquitecturas de redes neuronales densas (*fully connected*) con el objetivo de evaluar su capacidad para modelar relaciones no lineales en la variable dependiente `fare_amount`. Cada modelo se entrenó utilizando los mismos subconjuntos de entrenamiento y prueba empleados en los experimentos previos, asegurando así la comparabilidad de resultados. Las arquitecturas variaron en el número de capas ocultas y en la cantidad de neuronas por capa, manteniendo una configuración estándar de activación ReLU y optimización mediante el algoritmo Adam.

El desempeño de cada red se evaluó a partir de las métricas RMSE y MAE, cuyos valores se resumen en el **Cuadro 13**. Los resultados muestran un comportamiento consistente entre las distintas arquitecturas, con errores medios muy similares y una mejora marginal en el caso del modelo con una sola capa oculta (RMSE = 2.4622, MAE = 1.7214). Este modelo logra un equilibrio adecuado entre capacidad de generalización y complejidad computacional, evitando el sobreajuste que puede presentarse en arquitecturas más profundas.

Cuadro 13. Resultados comparativos de las arquitecturas de red neuronal para la predicción de `fare_amount`.

Arquitectura	RMSE	MAE
Modelo 1 (1 capa oculta)	2.4622	1.7214
Modelo 2 (2 capas ocultas)	2.4704	1.7504
Modelo 3 (3 capas ocultas)	2.4708	1.7268

5.5. Comparación general de desempeño

Con el fin de sintetizar los hallazgos obtenidos, se presenta una comparación general del desempeño de los modelos desarrollados: regresión lineal (OLS), modelos de ensamble (*Random Forest* y *Gradient Boosting*) y redes neuronales densas. Esta comparación permite evaluar no solo la precisión predictiva de cada enfoque, sino también su capacidad de generalización y su complejidad relativa.

En primer lugar, el modelo lineal mostró un ajuste sólido y una relación altamente significativa entre las variables `fare_amount` y `distance_km`, con un coeficiente de determinación R^2 de 0.674. Sin embargo, su naturaleza lineal limita la captura de relaciones más complejas presentes en los datos, especialmente en los rangos de tarifas más elevadas. Aun así, su simplicidad y transparencia lo convierten en una herramienta interpretativa de referencia.

Por su parte, los modelos de ensamble evidenciaron un mejor desempeño en términos de error. El *Random Forest* alcanzó un RMSE de 2.4844 y un MAE de 1.7472, mientras que el *Gradient Boosting* redujo ligeramente ambas métricas a 2.4583 y 1.7323, respectivamente. Estos resultados confirman que los métodos de ensamble, al combinar múltiples árboles de decisión, logran capturar relaciones no lineales y efectos de interacción entre variables que la regresión lineal no puede representar. Además, el *Gradient Boosting* mostró un ajuste más estable y menor varianza, gracias a su estrategia de optimización iterativa basada en el gradiente de los errores residuales.

Finalmente, las redes neuronales ofrecieron un rendimiento competitivo y consistente con los modelos de ensamble. El mejor modelo (una capa oculta) obtuvo un RMSE de 2.4622 y un MAE de 1.7214, prácticamente equivalente al *Gradient Boosting*. Si bien las arquitecturas más profundas no aportaron mejoras sustanciales, su estabilidad confirma la capacidad de las redes densas para aproximar funciones no lineales con eficiencia, manteniendo un nivel de error bajo y una adecuada generalización.

En conjunto, los resultados muestran que los tres enfoques convergen en una alta capacidad predictiva, con diferencias mínimas en los valores de error. El modelo *Gradient Boosting* se posiciona como el de mejor equilibrio entre precisión, robustez y eficiencia computacional, seguido de cerca por la red neuronal de una capa oculta. La regresión lineal, aunque menos precisa, continúa siendo un punto de partida valioso por su interpretabilidad y bajo costo computacional. Este análisis comparativo evidencia que la predicción de tarifas de viaje en el dataset de Uber puede resolverse

eficazmente mediante modelos de complejidad moderada, sin requerir arquitecturas excesivamente profundas o costosas en cómputo.

6. Discusión

Los resultados obtenidos en el presente estudio confirman la existencia de una relación sólida y estadísticamente significativa entre la distancia recorrida (`distance_km`) y la tarifa del viaje (`fare_amount`) en el conjunto de datos de Uber para la ciudad de Nueva York. Los distintos enfoques de modelado implementados —regresión lineal, modelos de ensamble y redes neuronales densas— permiten observar de manera complementaria el comportamiento de la variable objetivo, su nivel de predicción y las posibles no linealidades subyacentes.

En primer lugar, el modelo de regresión lineal (OLS) ofreció un desempeño satisfactorio, con un coeficiente de determinación R^2 de 0.674. Este resultado indica que aproximadamente dos tercios de la variabilidad en las tarifas pueden explicarse únicamente por la distancia recorrida, lo cual concuerda con la estructura tarifaria oficial de Uber, que combina un costo base fijo con un componente proporcional a la distancia. No obstante, la naturaleza lineal del modelo limita su capacidad para capturar patrones más complejos, como los efectos marginales decrecientes en trayectos largos o la presencia de tarifas mínimas establecidas por la plataforma. Aun así, su interpretación directa y su bajo costo computacional lo posicionan como una referencia robusta para estudios explicativos y de validación inicial.

Los modelos de ensamble, por su parte, demostraron una mejora notable en el ajuste y la capacidad predictiva. Tanto el *Random Forest* como el *Gradient Boosting* alcanzaron reducciones en los errores absolutos y cuadráticos medios (MAE y RMSE), situándose en torno a 1.7 y 2.4, respectivamente. Estas métricas reflejan una mejora de precisión respecto al modelo lineal, atribuible a la capacidad de los métodos de ensamble para capturar relaciones no lineales e interacciones entre variables. En particular, el *Gradient Boosting* superó ligeramente al *Random Forest*, mostrando menor dispersión entre valores reales y predichos, lo cual sugiere una mayor estabilidad y eficiencia al minimizar los errores residuales de manera iterativa. Este comportamiento concuerda con estudios previos en los que los métodos de *boosting* tienden a ofrecer mejor sesgo-varianza frente a modelos de *bagging*.

Las redes neuronales densas, finalmente, alcanzaron un desempeño comparable al de los modelos de ensamble, con un RMSE de 2.4622 y un MAE de 1.7214 en la arquitectura de una sola capa oculta. El incremento en la profundidad del modelo (dos o tres capas ocultas) no produjo mejoras sustanciales, lo que sugiere que la relación entre las variables principales es predominantemente monótona y puede capturarse adecuadamente con un modelo de baja complejidad. Esto es consistente con la teoría del sesgo-varianza, donde modelos más profundos tienden a sobreajustar cuando los datos presentan estructuras simples. Sin embargo, la red neuronal mantiene una ventaja en términos de flexibilidad y escalabilidad para futuras extensiones del modelo con más variables predictoras o características no lineales.

En conjunto, los resultados muestran que los tres enfoques convergen hacia un alto nivel de precisión predictiva, con diferencias marginales en las métricas de error. El *Gradient Boosting* se posiciona como el modelo con mejor equilibrio entre exactitud, robustez y eficiencia computacional, seguido de cerca por la red neuronal de una capa oculta. La regresión lineal, aunque menos precisa, conserva su valor interpretativo y su utilidad como punto de partida para la modelización.

Desde el punto de vista aplicado, estos hallazgos tienen implicaciones relevantes para la estimación de tarifas dinámicas y la optimización de precios en plataformas de transporte urbano. La capacidad de predecir tarifas a partir de información geográfica mínima permite el desarrollo de sistemas más transparentes, eficientes y equitativos para conductores y usuarios. Además, la integración de modelos de predicción en tiempo real puede contribuir a una mejor asignación de recursos y a la planificación de la movilidad urbana.

Entre las limitaciones del estudio se destaca el uso de un subconjunto de datos con un número limitado de variables. Factores adicionales como la hora del día, el tráfico, las condiciones climáticas o la demanda local podrían influir en las tarifas y mejorar la precisión del modelo. Asimismo, el análisis se centró exclusivamente en la ciudad de Nueva York, lo que restringe la generalización de los resultados a otras regiones con estructuras tarifarias o patrones de movilidad distintos.

Como proyección futura, se propone incorporar variables temporales y contextuales, aplicar técnicas de ingeniería de características (*feature engineering*) y explorar arquitecturas neuronales más avanzadas, como redes recurrentes (RNN) o convolucionales (CNN) adaptadas a datos espaciales. También sería pertinente comparar estos resultados con modelos híbridos que integren aprendizaje automático y econometría tradicional, con el fin de optimizar tanto la precisión como la interpretabilidad del sistema predictivo.

En síntesis, la discusión de los resultados evidencia que la predicción de tarifas de viaje en Uber puede abordarse eficazmente mediante modelos de complejidad moderada, confirmando que las técnicas de *ensemble learning* y las redes neuronales representan una evolución natural del enfoque lineal clásico, ofreciendo un equilibrio óptimo entre precisión, flexibilidad y aplicabilidad real.

7. Conclusiones

El presente estudio permitió desarrollar y evaluar distintos modelos predictivos para la estimación de tarifas de viaje de Uber en la ciudad de Nueva York, utilizando información geográfica básica y técnicas de aprendizaje supervisado. A través del análisis comparativo de enfoques lineales, de ensamble y de redes neuronales, se logró cumplir con el objetivo general de identificar el modelo más eficiente para predecir el valor de la tarifa en función de la distancia recorrida.

Los resultados evidenciaron que la variable `distance_km` constituye el principal determinante del valor de la tarifa, mostrando una relación positiva y estadísticamente significativa. El modelo de regresión lineal ordinaria (OLS) logró explicar aproximadamente el 67 % de la variabilidad observada en las tarifas, con un incremento promedio de 2 USD por kilómetro recorrido. Este modelo, aunque limitado en su capacidad para capturar relaciones no lineales, ofrece una interpretación directa y una excelente referencia para validar modelos más complejos.

Los modelos de ensamble demostraron un desempeño superior, destacándose el *Gradient Boosting* con un error cuadrático medio (RMSE) de 2.4583 y un error absoluto medio (MAE) de 1.7323. Estos valores reflejan una mejora significativa frente al modelo lineal, atribuible a la capacidad de los métodos de ensamble para capturar interacciones y patrones no lineales entre las variables. El *Random Forest*, aunque ligeramente menos preciso, mostró una alta estabilidad y robustez, consolidándose como una alternativa confiable para predicciones en contextos reales.

Las redes neuronales densas obtuvieron resultados comparables a los modelos de ensamble, con un rendimiento óptimo en la arquitectura de una capa oculta (RMSE = 2.4622; MAE = 1.7214). La incorporación de capas adicionales no mejoró el desempeño, lo cual sugiere que la relación entre las variables analizadas puede modelarse eficazmente mediante arquitecturas de baja complejidad. Esto confirma que, para problemas de naturaleza predominantemente monótona y unidimensional, el uso de redes profundas no siempre resulta necesario.

En términos comparativos, el modelo *Gradient Boosting* se posiciona como el enfoque más equilibrado en cuanto a precisión, eficiencia y generalización, seguido de cerca por la red neuronal de una capa oculta. La regresión lineal, pese a su menor precisión, conserva su relevancia por su interpretabilidad y bajo costo computacional, sirviendo como punto de referencia esencial para la validación inicial del conjunto de datos y de las tendencias generales del fenómeno.

Desde una perspectiva práctica, este estudio confirma la viabilidad de estimar tarifas de transporte urbano con alta precisión a partir de información geográfica mínima, lo que abre posibilidades de aplicación en sistemas de tarificación dinámica, cálculo en tiempo real y optimización de la movilidad urbana. Los resultados podrían integrarse en plataformas inteligentes de transporte para mejorar la transparencia y eficiencia de los precios, así como para apoyar la planificación y gestión de la demanda en zonas metropolitanas.

Entre las principales limitaciones se destaca la ausencia de variables contextuales como hora del día, condiciones de tráfico o meteorológicas, que podrían influir en la determinación final del precio. Asimismo, el análisis se limitó a un ámbito geográfico específico (Nueva York), lo cual restringe la extrapolación de los resultados a otras ciudades o regiones con diferentes estructuras tarifarias o dinámicas de movilidad.

Como líneas de trabajo futuras, se propone la ampliación del conjunto de variables predictoras, la incorporación de datos temporales y espaciales, y la exploración de modelos híbridos que combinen la interpretabilidad de la econometría con la potencia predictiva del aprendizaje automático. De igual modo, la integración de arquitecturas neuronales recurrentes (RNN) o convolucionales (CNN) permitiría capturar patrones dinámicos y espaciales más complejos, elevando la capacidad del sistema predictivo a escenarios de mayor escala y precisión.

En síntesis, el estudio demuestra que las técnicas modernas de aprendizaje automático, en particular los modelos de ensamble y las redes neuronales, constituyen herramientas poderosas para la predicción de tarifas en sistemas de transporte urbano, ofreciendo un balance óptimo entre exactitud, interpretabilidad y aplicabilidad real. Este trabajo sienta las bases para el desarrollo de sistemas predictivos más completos y adaptativos, orientados a una gestión más eficiente y sostenible de la movilidad inteligente.

Referencias

[1] Uber Fares Dataset (2021). *Uber Pickups in New York City*. Kaggle. Disponible en: `https://www.kaggle.com/datasets/yasserh/uber-fares-dataset/data`

[2] Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

[4] Abadi, M. et al. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Google Research.

[5] López, J. & García, M. (2020). *Modelos predictivos en movilidad urbana: una revisión aplicada*. Revista Latinoamericana de Ciencia de Datos, 8(3), 45–61.

Anexos

A. Repositorio y código fuente

El análisis completo, incluyendo el preprocesamiento, modelado y visualización de resultados, fue desarrollado en Python y se encuentra disponible en el siguiente repositorio público de GitHub:

- **Repositorio:** `Inteligencia_de_Negocios_UBA_2025`
- **Autor:** Julián Delgadillo Marín
- **URL:** `https://github.com/delgjulian/Inteligencia_de_Negocios_UBA_2025/blob/main/Informe_Final_Remanente/Scripts/tp_final_bi_2025_Delgadillo.ipynb`

El notebook incluye el código ejecutable empleado para:

1. Limpieza y depuración del dataset original de Uber NYC.
2. Cálculo de distancias mediante la fórmula de Haversine.
3. Filtrado de valores atípicos mediante el método IQR.
4. Entrenamiento y evaluación de los modelos OLS, LASSO, Random Forest, Gradient Boosting y Redes Neuronales.
5. Generación de tablas y gráficos presentados en este informe.

B. Repositorio general del proyecto

El proyecto completo de análisis de Inteligencia de Negocios, que incluye los scripts, informes parciales, visualizaciones y entregables finales, se encuentra disponible en el siguiente repositorio de GitHub:

- **Repositorio:** `Inteligencia_de_Negocios_UBA_2025`

- **Autor:** Julián Delgadillo
- **URL:** `https://github.com/delgjulian/Inteligencia_de_Negocios_UBA_2025/tree/main`

Este repositorio alberga las carpetas correspondientes a:

1. **Scripts:** código fuente en Python y Jupyter Notebook utilizado para el procesamiento y modelado.
2. **Datos:** dataset de Uber NYC y versiones intermedias depuradas.
3. **Gráficos:** figuras generadas automáticamente a partir de los modelos predictivos.
4. **Informe_Final_Remanente:** documento principal en formato LaTeX y los resultados experimentales asociados.

El acceso abierto al repositorio garantiza la reproducibilidad de los resultados, la trazabilidad del proceso analítico y la transparencia metodológica, en concordancia con las buenas prácticas de ciencia de datos reproducible.