Predicción de propinas de taxi en NYC

# Descripción del proyecto

Este proyecto se centra en el desarrollo de un modelo de Machine Learning capaz de predecir la propina que un cliente podría dejar al solicitar un viaje. Dicha predicción se basa en diversas variables, como la distancia del trayecto, las zonas de origen y destino, la hora del día, el día de la semana, el estado del tráfico y más factores relevantes. Este modelo es una propuesta preliminar que podría implementarse en una futura aplicación dirigida a taxistas para mejorar la estimación de propinas durante sus servicios en función a las características del viaje, facilitando así la toma de decisiones estratégicas y mejorando la eficiencia del servicio.

Para elaborar el modelo, se utiliza un dataset descargado de la plataforma Kaggle, que contiene información sobre todos los viajes en taxi realizados en la ciudad de Nueva York durante el mes de septiembre de 2023.

# Descripción de los datos

El dataset de Kaggle consiste en dos archivos:

* Yellow\_trip\_data\_2023-09: un archivo .parquet con 19 columnas y 2.846.721 registros que contiene las características de cada uno de los viajes en taxi realizados en dicho periodo
* Taxi\_zone\_geo: un archivo .csv con 4 columnas y 263 de registros que contiene las coordenadas que delimitan cada zona y su nombre

A continuación, se muestra la info del dataset. Para ampliar información sobre qué significa cada columna, se puede consultar el archivo data\_dictionary\_trips\_yellow.pdf en la carpeta info del dataset.

Escala de tiempo

Descripción generada automáticamente con confianza baja

# Limpieza y preparación de los datos

Se llevaron a cabo una serie de acciones para garantizar la **calidad y relevancia de los datos**, eliminando inconsistencias y filtrando información poco útil para el modelo. Las principales acciones fueron:

* Eliminación de todas las filas con algún valor null, dado que el volumen de datos era suficientemente grande
* Eliminación de filas con algún valor negativo, ya que no tienen sentido en el contexto del dataset.
* Eliminación de todas las filas con PULocationID o DOLocationID no comprendidas entre 1 y 263, ya que no existen en taxi\_zone\_geo.
* Extracción de información temporal. Se generaron nuevas columnas a partir de las columnas tipo datetime: día de la semana, hora, duración del viaje y velocidad media.
* Eliminación de los viajes pagados en cash, ya que en la documentación de Kaggle se especifica que el dataset no recoge las propinas en cash.
* Para centrarnos en viajes "normales" y evitar distorsionar el modelo, se seleccionaron solo los viajes con duración entre 3 y 90 minutos, y distancia entre 1 y 100km, y se eliminaron los viajes con 0 en passenger\_count y velocidad media superior a 100km/h.
* También se eliminó la variable total\_amount, ya que contiene la propina, y la variable improvement\_surcharge, ya que no aporta nada al no tener variabilidad.
* La variable Passenger\_count estaba muy desbalanceada. Mediante el método de **Kruskal-Wallis**, se determinó que la mejor transformación consistía en agrupar todos los valores mayores a 3 en una sola categoría con valor 3.

Tras estas acciones, el número de registros se redujo a 1.959.279, manteniendo únicamente la información relevante y de calidad para el desarrollo del modelo de predicción.

# Exploratory Data Analysis

En el **análisis exploratorio de datos (EDA)**, se realizó una evaluación detallada del conjunto de datos para comprender su estructura, distribución y relación entre variables clave. El foco principal fue la variable objetivo **tip\_amount** (cantidad de propina en $). Resumidamente, se obtuvo la siguiente información:

1. **Distribución del target (tip\_amount)**:
   * La variable está **altamente sesgada a la izquierda**, con una larga cola hacia valores más altos. Esto indica que la mayoría de los viajes tienen propinas pequeñas o nulas.
   * Se analizaron las categorías de propinas dividiendo los valores en rangos (sin propina, hasta $5, hasta $10, etc.) para visualizar mejor su frecuencia y comportamiento.
2. **Resumen estadístico y correlaciones**:
   * Se generó un resumen estadístico de las variables, mostrando sus **medias, desviaciones estándar** y rangos.
   * Se calcularon correlaciones entre las características numéricas y el target para identificar variables potencialmente relevantes para el modelo.
3. **Visualización de outliers**:
   * Mediante **histogramas** y **boxplots**, se identificaron outliers en las propinas y otros valores clave.
   * Se filtraron o resaltaron valores extremos en variables como tip\_amount para evitar distorsiones en futuros modelos.
4. **Visualización de mapas**:
   * Con la herramienta folium se crearon mapas para visualizar la distribución de los viajes y las propinas según la zona de NYC

Estas acciones permitieron extraer información valiosa sobre la naturaleza de los datos y sentar las bases para la fase de modelado.

# Machine Learning

En un inicio, el proyecto se planteó con el siguiente enfoque:

1. **Modelo de clasificación**: Diseñar un modelo para clasificar los viajes en dos categorías: **propinas bajas** y **propinas altas**, estableciendo una frontera en **10 dólares**.
2. **Modelos de regresión**: Posteriormente, desarrollar **dos modelos de regresión**, uno para cada clase identificada en la etapa de clasificación, con el fin de predecir de manera más precisa el monto de la propina en cada grupo.

Dado que el objetivo principal del proyecto es **identificar viajes con propinas potencialmente altas**, se seleccionó el **F2-score** como métrica de evaluación para el clasificador inicial. Esta métrica, derivada del **F1-score**, utiliza un coeficiente **β** que otorga al **recall** (tasa de aciertos en la identificación de viajes con propinas altas) **cuatro veces más importancia que a la precisión**, priorizando así la minimización de falsos negativos (es decir, evitar no detectar viajes con propinas altas).

Tras probar distintos modelos (clasificadores del 1 al 5 en la carpeta pruebas de modelos de github), principalmente con **XGBoost** y diversas configuraciones, se obtuvieron resultados muy satisfactorios en la clasificación, alcanzando un **F2-score** cercano a **0,94**. Sin embargo, al analizar las **importancias de las variables** (feature importances), se observó que el modelo basaba prácticamente toda su predicción en la **tarifa** y la **distancia del viaje.** Esto no era lo deseado, ya que el objetivo no es únicamente clasificar correctamente, sino también **identificar las características** que puedan influir en la propina y ayudar a entender mejor el comportamiento de los pasajeros.

Para reducir la dependencia del modelo en las variables de tarifa y distancia, se decidió crear una nueva característica denominada tip\_per\_km, dividiendo la propina por la distancia recorrida. Esta transformación permite **normalizar la influencia de la distancia** y resaltar mejor la **generosidad del pasajero** independientemente de la longitud del trayecto. De este modo, el modelo puede hilar más fino y detectar patrones más sutiles en otras variables y mejorar su capacidad de generalización.

Tras este cambio de enfoque, se diseñó un clasificador que dividiese los viajes en propinas normales y propinas generosas, estableciendo la frontera en 2$/km. Se obtuvo un  **F2-score aproximado de 0,82**, lo que representa un empeoramiento en comparación con el modelo inicial (0,94). No obstante, esta reducción en rendimiento se vio compensada con una **mayor explicabilidad** del modelo, al lograr identificar otras características relevantes además de la tarifa y la distancia del viaje.

Posteriormente, se ajustaron dos modelos de regresión específicos para cada categoría. Aunque ambos predicen la propina por km, **el error se evaluó en términos de la cantidad total de propina**, multiplicando la predicción por la distancia correspondiente al viaje. Los resultados obtenidos fueron los siguientes:

* **Propinas normales (<2$/km):**
  + El modelo obtuvo un **MAE (Error Absoluto Medio) de 1,45** y un **R² de 0,665**.
  + Aunque la predicción aún depende en gran medida de la **tarifa por kilómetro**, se logró identificar la **congestión del tráfico** como una variable secundaria de gran importancia, lo cual proporciona una mejor comprensión del fenómeno en este segmento.

Gráfico

Descripción generada automáticamente

Gráfico 1. Feature importances modelo de regresión para propinas standard

* **Propinas generosas (>2$/km):**
  + En este caso, el modelo alcanzó un **MAE de 0,65** y un **R² de 0,67**, lo que refleja un mejor ajuste y menor error en comparación con el modelo de propinas normales.
  + Las **importancias de las variables** estuvieron más equilibradas. Este modelo pudo aprovechar de manera más efectiva múltiples características para realizar predicciones precisas.

Gráfico, Gráfico de embudo

Descripción generada automáticamente

Gráfico 2. Feature importances modelo de regresión para propinas generosas

Adicionalmente, se probó un modelo de regresión aplicado a todo el conjunto de datos, sin clasificar previamente entre propinas normales y generosas. Este modelo logró un MAE de 1,45, similar al modelo de propinas normales, pero mostró limitaciones al no aprovechar la riqueza de características presentes en el conjunto de datos. Esto sugiere que la segmentación previa es fundamental para mejorar tanto el ajuste como la explicabilidad de las predicciones.

En la figura presentada a continuación se muestra la comparación entre los valores reales de propina y las predicciones realizadas por cada uno de los modelos de regresión, representados como puntos en el gráfico. El eje **x** corresponde a las propinas reales ($) y el eje **y** a las propinas predichas ($).

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico 3. Propina real vs predicción del modelo de propinas standard (<2$/km)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico 4. Propina real vs predicción del modelo para propinas generosas (>= 2$/km)

Ambos regresores obtienen resultados aceptables para valores normales de propina, pero tienen dificultades para predecir valores extremos.

# Conclusión

El cambio de enfoque del proyecto permitió alcanzar un equilibrio entre rendimiento predictivo y **explicabilidad del modelo**, logrando una mejor comprensión de los factores que influyen en la propina. Aunque el clasificador mostró una reducción en su desempeño (F2-score de 0,82), logró segmentar los viajes en dos grupos relevantes: propinas normales y generosas.

Los resultados de los modelos de regresión evidencian diferencias significativas entre ambos segmentos. En los viajes con **propinas normales o bajas** (menos de 2$/km), **la tarifa y la congestión del tráfico** parecen ser los factores predominantes, lo que sugiere que los pasajeros tienden a dejar este tipo de propinas cuando perciben incomodidades durante el trayecto, como atascos o perciben que la tarifa es demasiado elevada.

Por el contrario, en los viajes con **propinas elevadas** (más de 2$/km), los pasajeros parecen basar su decisión en una **combinación más amplia de factores**, lo que indica una mayor complejidad en el comportamiento asociado a propinas generosas.

Es decir, para propinas generosas, el modelo requiere más variables y un análisis más complejo, mientras que las propinas normales dependen casi exclusivamente de la tarifa y el estado del tráfico. Este comportamiento puede resumirse con la frase: “*es más fácil criticar que agradecer”.*

En resumen, esta estrategia ofrece una herramienta más explicativa y práctica, que no solo predice el monto de las propinas, sino que también permite identificar las características más relevantes que podrían influir en la generosidad del cliente. Esto sienta las bases para futuras mejoras en el modelo, orientadas a reducir la dependencia de las variables predominantes y capturar patrones más complejos en los datos.