

Facultad de Ciencias Exactas, Ingeniería y Agrimensura

# Tecnicatura en Inteligencia Artificial

## Aprendizaje Automático 1

Trabajo Práctico N°1: modelo predictivo de tarifas de Uber

**Objetivos**

Familiarizarse con la biblioteca scikit-learn y las herramientas que brinda para el pre-procesamiento de datos, la implementación de modelos de regresión lineal con diversos hiper parámetros y la evaluación de métricas de regresión.

**Dataset**

El dataset se llama uber\_fares.csv y contiene información de tarifas de viajes realizados, además de distintas variables características, como se detallan a continuación.

Características de entrada:

**key** - un identificador único para cada viaje.

**pickup\_datetime** - fecha y hora en que se activó el taxímetro.

**passenger\_count** - el número de pasajeros en el vehículo (valor ingresado por el conductor).

**pickup\_longitude** - la longitud donde se activó el taxímetro.

**pickup\_latitude** - la latitud donde se activó el taxímetro.

**dropoff\_longitude** - la longitud donde se desactivó el taxímetro.

**dropoff\_latitude** - la latitud donde se desactivó el taxímetro.

Variable de salida (target):

**fare\_amount** - el costo de cada viaje en USD

Este trabajo comprende el modelado predictivo para estimar tarifas de viajes Uber en Nueva York. Sobre el dataset original se realizó una exhaustiva limpieza y preprocesamiento, tratando de asegurar la calidad y relevancia de los datos mediante:

* Eliminación de valores anómalos,
* Imputación robusta
* Transformación de variables.
* Incorporación de atributos geográficos y temporales
* Aplicación de métodos de escalado y detección de outliers para optimizar el desempeño de los modelos.

Respecto de las variables temporales (hora de inicio del viaje), se decidió no usarlas en el análisis porque generaban métricas MAPE excesivamente grandes

Se evaluaron distintos enfoques de regresión: desde el modelo OLS clásico, pasando por variantes de gradiente descendente (batch, mini batch y estocástico), hasta técnicas de regularización como Ridge, Lasso y Elastic Net. Cada modelo fue analizado tanto en su capacidad de ajuste como en su generalización, utilizando métricas como RMSE y R², y se compararon sus resultados en el conjunto de prueba.

Los resultados muestran que la variable **distancia** es el principal predictor de la tarifa, y que los modelos regularizados (especialmente Ridge y Elastic Net) logran un buen equilibrio entre ajuste y robustez, evitando el sobreajuste y manteniendo la relevancia de las variables. El análisis de residuos y la comparación de métricas confirman la estabilidad y eficacia de los modelos desarrollados.

**Conclusiones finales**

Como conclusión final del trabajo destacamos la importancia de los siguientes puntos:

1. Realizar un pipeline de datos bien estructurado
2. Aplicar técnicas adecuadas de modelado y validación.
3. Aclarar desde un comienzo cuál es el alcance real del modelo. ¿sobre qué universo de eventos aplica? ¿Cuál es el tamaño de variables explicativas dentro del cual son relevantes los resultados que entregue? ¿Hay variables explicativas que son redundantes? En nuestro caso nos concentramos en los Viajes en Uber realizados del 2009 al 2016 con vehículos con capacidad de hasta seis pasajeros dentro del perímetro delimitado por las coordenadas indicadas en <https://data.cityofnewyork.us>
4. La limpieza de datos puede organizarse en dos ámbitos. El primero de ellos (ámbito externo) debe basarse en las condiciones de contorno del problema que permiten detectar anomalías o datos absurdos. P. Ej, precios o distancias negativas, cantidades ridículas de pasajeros, etc. El segundo ámbito de la limpieza (interno) ya precisa el uso de estadísticos calculados a partir de los datos que quedaron: medias y medianas para decidir sobre atípicos. Nunca generar estadísticos para decidir sobre el conjunto de entrenamiento que hayan sido creados con datos que incluyan a los de prueba.
5. Antes de descartar filas hacer lo posible para ver si se pueden aprovechar haciendo imputaciones de los valores faltantes. Para lo cual es importante que en los pasos previos de limpieza (interna y externa) se hagan marcados como NaN en vez de descartar filas.