# Boston House Pricing

Carlos Milián - 22006103 Yoselin Mejía - 20011412 Kimberly Hernández - 20002617

## Introducción

El presente proyecto tiene como objetivo predecir el precio medio de viviendas en Boston mediante modelos de inteligencia artificial. Se trata de un problema clásico de regresión, donde a partir de variables como el número promedio de habitaciones, nivel de criminalidad y distancia a centros laborales, se busca estimar el valor de una vivienda en miles de dólares. Se entrenaron modelos diferentes para tener diversos puntos de comparación.

## Descripción del Dataset

El dataset **Boston Housing** contiene 506 registros de viviendas y 13 variables predictoras, a continuación describiremos algunas de las variables.

- **RM**: Promedio de habitaciones por vivienda
- **CRIM**: Tasa de criminalidad per cápita
- **DIS**: Distancia a centros de empleo
- **TAX**: Tasa de impuesto a la propiedad
- **LSTAT**: Porcentaje de población con bajos ingresos

Se eliminaron algunas variables con bajo aporte predictivo o impacto ético.

Las variables que se eliminaron son:

- **CHAS**: Cercanía al río Charles (0/1)
- **B**: Proporción transformada de población afroamericana
- **AGE**: Porentaje de casas construidads antes de 1940
- **RAD**: Accesibilidad a autopistas

La variable objetivo es **PRICE**, que representa el valor promedio de una vivienda (en miles de dólares)

## Mejoras a futuro

- Implementar y comparar otros algoritmos como Random Forest o XGBoost.
- Aplicar técnicas de aumento de datos o generación sintética para mejorar la distribución.
- Utilizar validación cruzada y ajuste de hiperparámetros.
- Visualizar la importancia de cada variable para mejorar la interpretabilidad.

## Resultados

Los modelos obtuvieron diversos resultados, a continuación el conjunto de prueba

| Modelo | MAE | RMSE | $R^2$ | $R^2$ ajustado |
|---|---|---|---|---|
| Modelo 1 | 3.08 | 4.26 | 0.64 | 0.53 |
| Modelo 2 | 2.34 | 3.41 | 0.70 | 0.60 |
| Modelo 3 | 2.06 | 3.07 | 0.75 | 0.66 |

## Conclusiones

El modelo 1 usó la regresión lineal la cual permitió establecer una base de comparación sólida para el problema. Aunque el modelo es simple, logró un $R^2$ de 0.64, lo cual indica un ajuste moderado.

Sin embargo, presentó limitaciones en su capacidad de capturar relaciones no lineales o interacciones complejas entre variables, lo cual se refleja en un mayor error absoluto medio (MAE) comparado con las redes neuronales.

El modelo 2 logro explicar el 70% de la variabilidad en el precio de viviendas en Boston, se concluye:

- El preprocesamiento de datos y la selección de variables son fundamentales.
- Dropout ayudó a prevenir el sobreajuste y mejorar la generalización.
- La normalización de variables fue esencial para el buen desempeño de la red neuronal.

El modelo 3 alcanzo alcanzando un $R^2$ de 0.75. La arquitectura más profunda y el uso combinado de Dropout y EarlyStopping ayudaron a prevenir el sobreajuste sin perder capacidad de aprendizaje.

- La mejora en MAE y $R^2$ muestra que la red captura mejor relaciones complejas.
- Es un buen balance entre complejidad del modelo y rendimiento predictivo.
- Validación y regularización fueron clave para el rendimiento final.

Comparación Modelos: Al comparar los tres modelos, el Modelo 2 se destaca por su mejor desempeño general. Obtuvo el menor MAE (2.70) y el mayor $R^2$ (0.7350), dando predicciones más precisas. El Modelo 1, aunque tiene el RMSE es más bajo (4.25), presenta un $R^2$ poco inferior (0.7319). El Modelo 3 es el menos preciso, con el mayor RMSE (5.05) y el $R^2$ más bajo (0.6812). En resumen, el Modelo 2 es el más recomendable para este conjunto de datos.

## Metodología

El Modelo 1 utiliza una regresión lineal multivariable implementada con `scikit-learn`. El dataset fue preprocesado eliminando las variables `CHAS`, `B`, `AGE` y `RAD`, que se consideraron poco útiles o éticamente cuestionables.

Las variables predictoras (`X`) fueron separadas de la variable objetivo (`PRICE`), y los datos fueron divididos en 80% para entrenamiento y 20% para prueba.

El modelo fue entrenado sin normalización explícita, ya que los modelos lineales pueden tolerar rangos variados. Se evaluó con métricas estándar para regresión.

En el Modelo 2 se normalizaron los datos con `MinMaxScaler` y se dividieron en 80% entrenamiento y 20% prueba. El modelo 2 fue una red neuronal con la siguiente arquitectura:

- Capa densa con 64 neuronas y activación `ReLU`
- `Dropout` del 30%
- Capa densa con 32 neuronas y activación `ReLU`
- `Dropout` del 30%
- Capa de salida con 1 neurona (regresión)

Se utilizó el optimizador `Adam`, función de pérdida MSE, y métrica MAE. El modelo fue entrenado durante 200 épocas con validación del 20%.

El modelo 3 es una red neuronal mejorada con regularización adicional y más neuronas:

- Capa densa con 128 neuronas + ReLU
- Dropout del 20%
- Capa densa con 64 neuronas + ReLU
- Dropout del 20%
- Capa densa con 32 neuronas + ReLU
- Capa de salida con 1 neurona (regresión)

Se utilizó el optimizador Adam, pérdida MSE, y métrica MAE. Entrenamiento de 200 épocas con validación del 20%. Se implementó EarlyStopping para evitar sobreentrenamiento.