Pontificia Universidad Javeriana

Taller Regresión

Tecnologías Emergentes

Presentado por:

Nicolás Rincón Ballesteros

Alejandro Suarez Acosta

Andrés García Montoya

Entregado a:

Randy Lancheros

Bogotá D.C

Sábado 06 de abril 2024

Descripción del Dataset

El Dataset de “The Boston Housing Dataset” fue extraído de la página de Kaggle https://www.kaggle.com/datasets/sakshisatre/the-boston-housing-dataset?resource=download. En esta página se encuentra la siguiente información acerca del Dataset:

Sobre el Dataset

El conjunto de datos de Boston Housing, que a menudo se utiliza para análisis de regresión y tareas de modelado predictivo, normalmente no tiene un “subtítulo” oficial. Sin embargo, comúnmente se lo conoce como “conjunto de datos de viviendas de Boston” o “conjunto de datos de precios de viviendas de Boston” debido a que se centra en las características relacionadas con la vivienda y su principal variable objetivo es el valor medio de las viviendas ocupadas por sus propietarios en los suburbios de Boston.

Columnas

- **CRIM:** tasa de criminalidad per cápita por ciudad (numérica)
- **ZN:** proporción de suelo residencial zonificado para lotes superiores a 25.000 pies cuadrados. (numérico)
- **INDUS:** proporción de acres de negocios no minoristas por ciudad (numérica)
- **CHAS:** Variable ficticia de Charles River (1 si el tramo limita con el río; 0 en caso contrario) (categórica)
- **NOX:** concentración de óxidos nítricos (partes por 10 millones) (numérica)
- **RM:** número medio de habitaciones por vivienda (numérico)
- **AGE:** proporción de unidades ocupadas por sus propietarios construidas antes de 1940 (numérica)
- **DIS:** distancias ponderadas a cinco centros de empleo de Boston (numéricas)
- **RAD:** índice de accesibilidad a carreteras radiales (numérico)
- **TAX:** tasa de impuesto a la propiedad de valor total por $10,000 (numérico)
- **PTRATIO:** ratio alumnos-profesor por localidad (numérica)
- **B:** 1000(Bk - 0,63)^2 donde Bk es la proporción de [personas de ascendencia afroamericana] por ciudad (numérica)
- **LSTAT:** % estado inferior de la población (numérico)
- **MEDV:** Valor medio de viviendas ocupadas por sus propietarios en miles de dólares (variable objetivo) (numérico)

Análisis Exploratorio de los Datos

Para elegir cuales variables independientes se van a contrastar con la variable objetivo, primero, se elabora un mapa de calor para identificar la correlación entre las variables.

Figura 1. Mapa de calor dataframe. Fuente: Elaboración propia.

Asimismo, se usa la librería de seaborn para encontrar las relaciones entre todas las variables del dataframe, lo cual muestra lo siguiente:

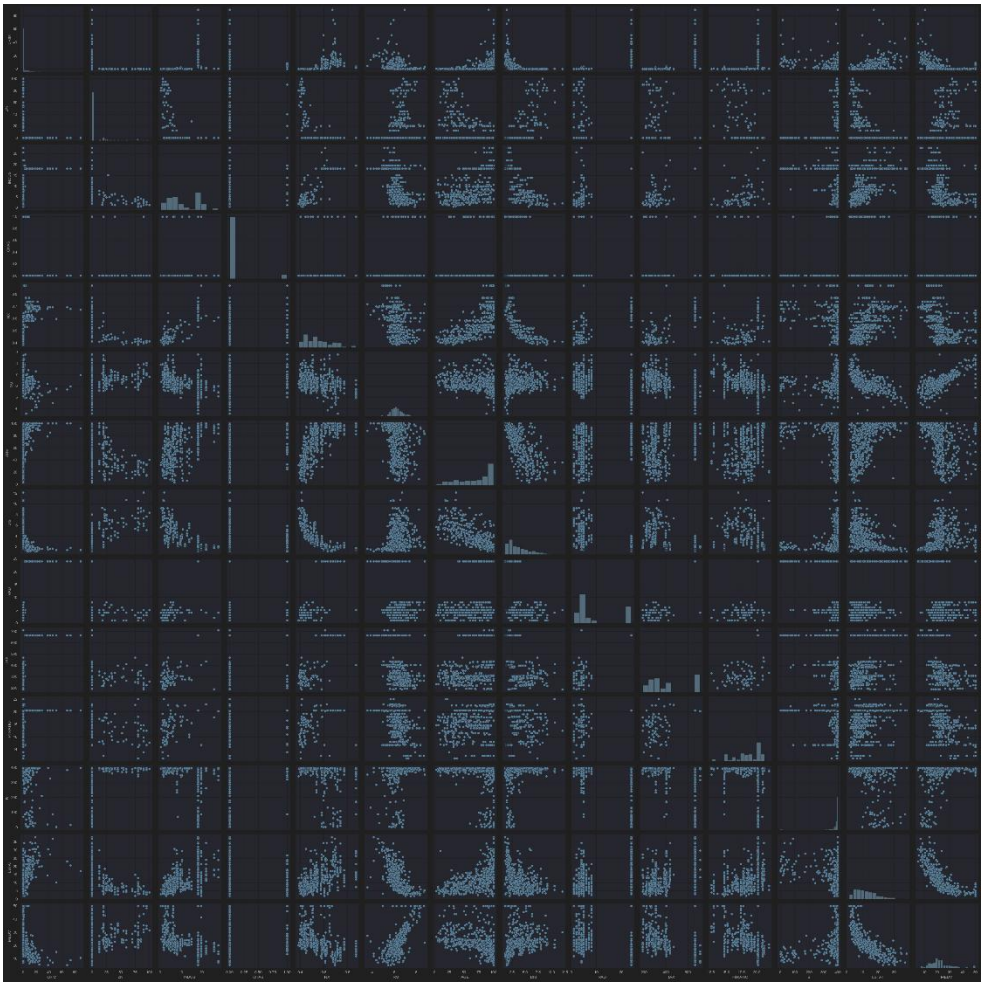

Figura 2. Pairplot dataframe. Fuente: Elaboración propia.

Sin embargo, como en este caso solo importa las variables independientes comparadas con la variable objetivo que es MEDV, por lo que, solo se tiene en cuenta importa la última fila.

Tras entender esto, se encuentra que las tres variables que tienen mayor correlación con la variable objetivo, teniendo en cuenta el mapa de calor y el pairplot, se seleccionan **LSTAT**, **RM** y **NX** para realizar el análisis de regresión.

Resultados Obtenidos

- **Regresión lineal simple**

MEDV vs LSTAT

La siguiente figura ilustra la regresión lineal simple entre la variable MEDV (Valor medio de viviendas ocupadas por sus propietarios en miles de dólares) y LSTAT (% estado inferior de la población).

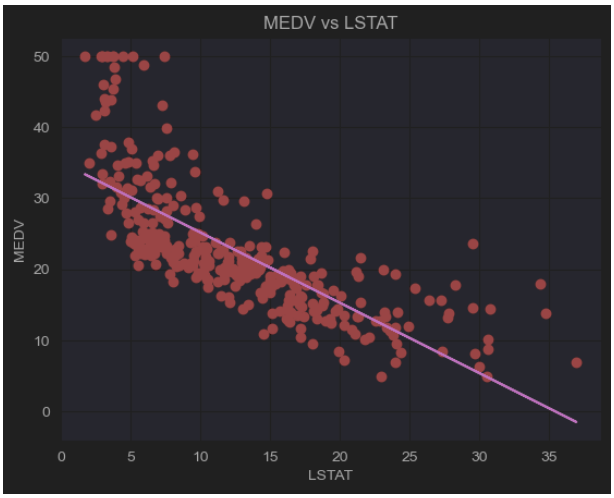

Figura 3. Regresión lineal simple datos de entrenamiento MEDV vs LSTAT. Fuente: Elaboración propia.

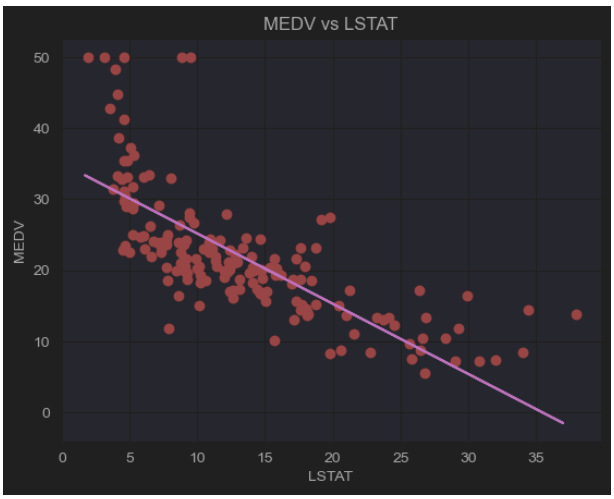

Figura 4. Regresión lineal simple datos de prueba MEDV vs LSTAT. Fuente: Elaboración propia.

La ecuación de la recta que mejor se ajusta a la regresión lineal simple entre las dos variables es la siguiente:

$$y = 35.07 - 0.99x$$

Como se puede ver, existe una relación inversa entre la variables, esto quiere decir que mientras la variable LSTAT aumenta, la variable MEDV disminuye. Esto se refleja en el coeficiente de la variable LSTAT en la ecuación de la recta, que es -0.99. Esto indica que por cada aumento unitario en LSTAT, se espera una disminución de aproximadamente 0.99 en el valor medio de las viviendas.

El coeficiente de intersección (constante) en la ecuación de la recta es 35.07. Este valor representa el valor esperado de la variable MEDV cuando la variable LSTAT es igual a cero. Esto significa que cuando el porcentaje de población de bajos ingresos en un área (LSTAT) es cero, el valor medio de las viviendas (MEDV) se estima en 35.07 unidades.

Ahora, respecto al coeficiente R^2, que en este caso es 0.5151425477523661. Este valor indica que aproximadamente 51.5% de la variabilidad en la variable MEDV puede explicarse por la variable LSTAT utilizando este modelo de regresión lineal simple.

MEDV vs RM

La siguiente figura ilustra la regresión lineal simple entre la variable MEDV (Valor medio de viviendas ocupadas por sus propietarios en miles de dólares) y RM (número medio de habitaciones por vivienda).

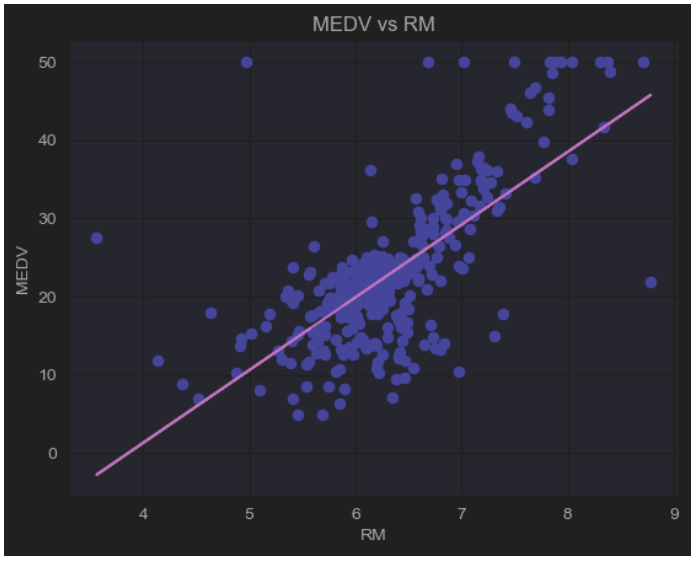

Figura 5. Regresión lineal simple datos de entrenamiento MEDV vs RM. Fuente: Elaboración propia.

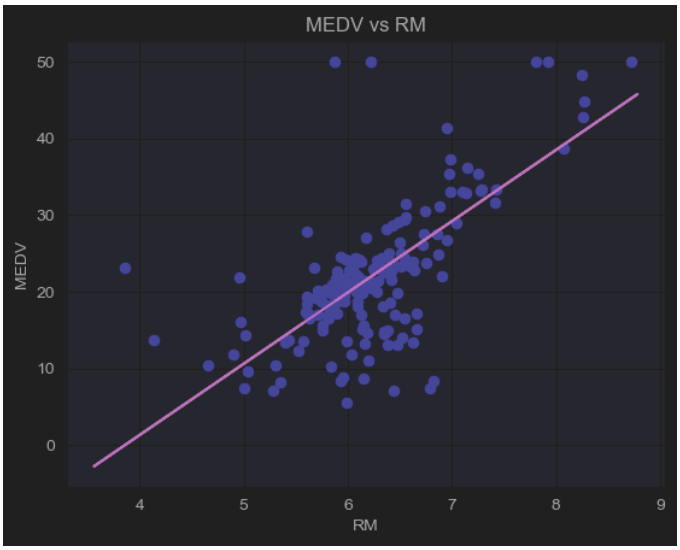

Figura 6. Regresión lineal simple datos de prueba MEDV vs RM. Fuente: Elaboración propia.

La ecuación de la recta que mejor se ajusta a la regresión lineal simple entre las dos variables es la siguiente:

$$y = -35.83 + 9.29x$$

Como se puede ver, existe una relación directa entre las variables, esto quiere decir que mientras la variable RM aumenta, la variable MEDV aumenta también. Esto se refleja en el coeficiente de la variable RM en la ecuación de la recta, que es 9.29. Esto indica que por cada aumento unitario en LSTAT, se aumenta en aproximadamente 9.29 el número medio de habitaciones por vivienda.

El coeficiente de intersección (constante) en la ecuación de la recta es -35.83. Este valor representa el valor esperado de la variable MEDV cuando la variable RM es igual a cero. Esto significa que cuando el número medio de habitaciones por vivienda (RM) es cero, el valor medio de las viviendas (MEDV) se estima en -35.83 unidades.

Ahora, en cuanto al coeficiente R^2, que en este caso es 0.446846413877101. Este valor indica que aproximadamente 44.7% de la variabilidad en la variable MEDV puede explicarse por la variable RM utilizando este modelo de regresión lineal simple. Esto implica que aunque el modelo proporciona cierta capacidad para predecir el valor medio de las viviendas basado en el número medio de habitaciones por vivienda, existe una cantidad significativa de variabilidad que no está siendo explicada por esta relación lineal simple.

MEDV vs NX

La siguiente figura ilustra la regresión lineal simple entre la variable MEDV (Valor medio de viviendas ocupadas por sus propietarios en miles de dólares) y NX (concentración de óxidos nítricos, partes por 10 millones).

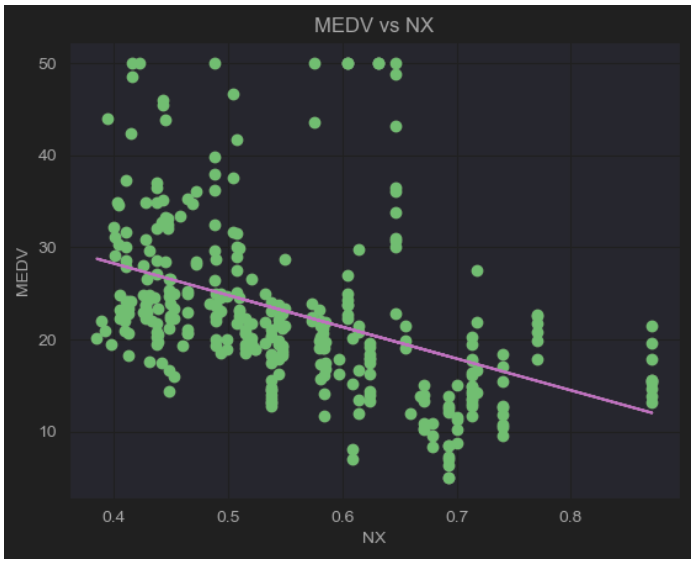

Figura 7. Regresión lineal simple datos de entrenamiento MEDV vs NX. Fuente: Elaboración propia.

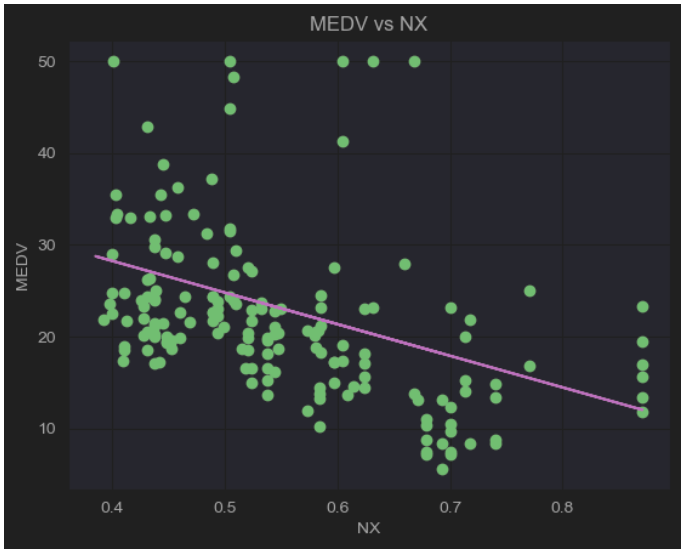

Figura 8. Regresión lineal simple datos de prueba MEDV vs NX. Fuente: Elaboración propia.

La ecuación de la recta que mejor se ajusta a la regresión lineal simple entre las dos variables es la siguiente:

$$y = 42.00 - 34.42x$$

Como se puede observar, existe una relación inversa entre las variables, lo que implica que mientras la variable NX (concentración de óxidos nítricos) aumenta, la variable MEDV (valor medio de las viviendas) tiende a disminuir. Esto se refleja en el coeficiente de la variable NX en la ecuación de la recta, que es -34.42. Indica que por cada aumento unitario en la concentración de óxidos nítricos, se espera una disminución de aproximadamente 34.42 en el valor medio de las viviendas.

El coeficiente de intersección (constante) en la ecuación de la recta es 42.00. Este valor representa el valor esperado de la variable MEDV cuando la variable NX es igual a cero. Esto significa que cuando la concentración de óxidos nítricos es cero, el valor medio de las viviendas se estima en 42.00 unidades.

Respecto al coeficiente R^2 que en este caso es 0.17110122287639162, indica que aproximadamente 17.1% de la variabilidad en la variable MEDV puede explicarse por la variable NX utilizando este modelo de regresión lineal simple. Esto sugiere que la concentración de óxidos nítricos tiene una influencia limitada en la predicción del valor medio de las viviendas, ya que la mayoría de la variabilidad en MEDV no puede ser explicada por esta variable en el contexto de este modelo lineal simple.

Para concluir, como se puede ver, los tres coeficientes R^2 encontrados demuestran que no existe un gran ajuste en los modelos de regresión lineal simple con las variables seleccionadas. Por lo tanto, no se puede aceptar los modelos de regresión lineal simple para predecir la variable objetivo. Por consiguiente, se procede a realizar una regresión lineal multiple.

- **Regresión lineal múltiple**

Teniendo en cuenta que $\mathbf{X_1}$ se refiere a la variable LSTAT, $\mathbf{X_2}$ se refiere a la variable RM, $\mathbf{X_3}$ se refiere a la variable NX y **Y** se refiere a la variable objetivo MEDV. Tras el entrenamiento del modelo de regresión lineal múltiple y predicción de resultados, se encontró que la ecuación obtenida es:

$$y = -1.00 - 0.63x_1 + 5.29x_2 - 3.25x_3$$

Esto significa que para cada unidad de cambio en las variables predictoras (X_1, X_2, X_3), se espera un cambio correspondiente en la variable de respuesta (y). Por ejemplo, un aumento de una unidad en x_1 se asocia con un descenso de 0.63 unidades en y, manteniendo constantes las otras variables.

Con respecto a la precisión del modelo, como se puede observar, el coeficiente de determinación R^2 es de aproximadamente 0.604 esto indica que alrededor del 60.4% de la variabilidad en la variable objetivo (MEDV) puede ser explicada por las variables predictoras seleccionadas (LSTAT, RM, NX) utilizando este modelo de regresión lineal múltiple. Un valor de R^2 de 0.604 sugiere un ajuste moderado del modelo, lo que significa que las variables predictoras explican una proporción considerable de la variabilidad en la variable de respuesta.

- **Regresión Polinómica**

En este análisis, se evalúan modelos de regresión polinómica de grado 2, 3 y 4 para explicar la variabilidad en MEDV. Se utiliza el coeficiente de determinación (R^2) como medida de la capacidad explicativa de cada modelo.

Regresión polinómica de grado 2

Para la regresión polinómica de grado 2, el modelo muestra un coeficiente de determinación (R^2) de aproximadamente 0.748. Esto significa que este modelo polinómico cuadrático explica casi el 74.8% de la variabilidad en la variable dependiente (MEDV).

Regresión polinómica de grado 3

El modelo polinómico de grado 3 tiene un R^2 de aproximadamente 0.685. Esto indica que alrededor del 68.5% de la variabilidad en MEDV puede ser explicada por este modelo. Aunque todavía es un ajuste razonable, notamos una disminución en la capacidad del modelo para explicar la variabilidad con respecto al modelo de grado 2.

Regresión polinómica de grado 4

Por último, el modelo de grado 4 exhibe un R^2 de aproximadamente 0.667. Esto sugiere que cerca del 66.7% de la variabilidad en la variable de interés puede ser explicada por este modelo. Al igual que con el modelo de grado 3, hay una disminución en la capacidad explicativa en comparación con el modelo de grado 2.

Conclusiones

El análisis de regresión realizado sobre el "Boston Housing Dataset" permitió explorar la relación entre el valor medio de las viviendas (MEDV) y diversas variables predictivas como el porcentaje de población de bajos ingresos (LSTAT), el número medio de habitaciones por vivienda (RM) y la concentración de óxidos nítricos (NX). A través de la aplicación de diferentes modelos de regresión, se buscó predecir el valor medio de las viviendas basándose en estas variables.

Los modelos de regresión lineal simple mostraron relaciones significativas entre MEDV y cada una de las variables predictoras por separado, evidenciando una relación inversa con LSTAT y NX y una relación directa con RM. Sin embargo, los coeficientes de determinación (R^2) de estos modelos indicaron un ajuste limitado, lo que sugiere que ninguna de estas variables por sí sola puede explicar de manera completa la variabilidad en MEDV.

La implementación de un modelo de regresión lineal múltiple, que incluye las variables LSTAT, RM y NX simultáneamente, mejoró la capacidad predictiva con un R^2 de aproximadamente 60.4%. Este resultado destaca la importancia de considerar múltiples factores al analizar el valor de las viviendas, aunque todavía una proporción significativa de la variabilidad en MEDV quedó sin explicar.

El análisis de regresión polinómica ofreció una mayor capacidad explicativa, especialmente con el modelo de grado 2, que alcanzó un R^2 de aproximadamente 74.8%. A medida que

aumenta el grado del polinomio, se observa una disminución en la capacidad explicativa de los modelos en comparación con el modelo de grado 2. Esto sugiere que el modelo cuadrático es el más efectivo para este conjunto de datos, sugiriendo que las relaciones entre las variables predictoras y la variable dependiente pueden ser no lineales. Sin embargo, es importante tener en cuenta que un polinomio de grado excesivamente alto se ajustará demasiado a los datos de entrenamiento (overfitting), perjudicando la capacidad del modelo para generalizar a nuevos datos.

En conclusión, este análisis resalta la complejidad de predecir el valor de las viviendas en los suburbios de Boston, evidenciando la necesidad de modelos que integren múltiples variables y consideren la posibilidad de relaciones no lineales entre ellas. El modelo de regresión polinómica de grado 2 emergió como el más prometedor para tales fines, ofreciendo una mayor capacidad explicativa en comparación con los modelos de regresión lineal simple y múltiple. Sin embargo, aún existe espacio para explorar modelos más complejos o técnicas de análisis adicionales que puedan mejorar la precisión de las predicciones y proporcionar insights más profundos sobre los factores que influencian el valor de las viviendas en esta área.