Inteligencia artificial avanzada para la ciencia de datos I

(Gpo 101)

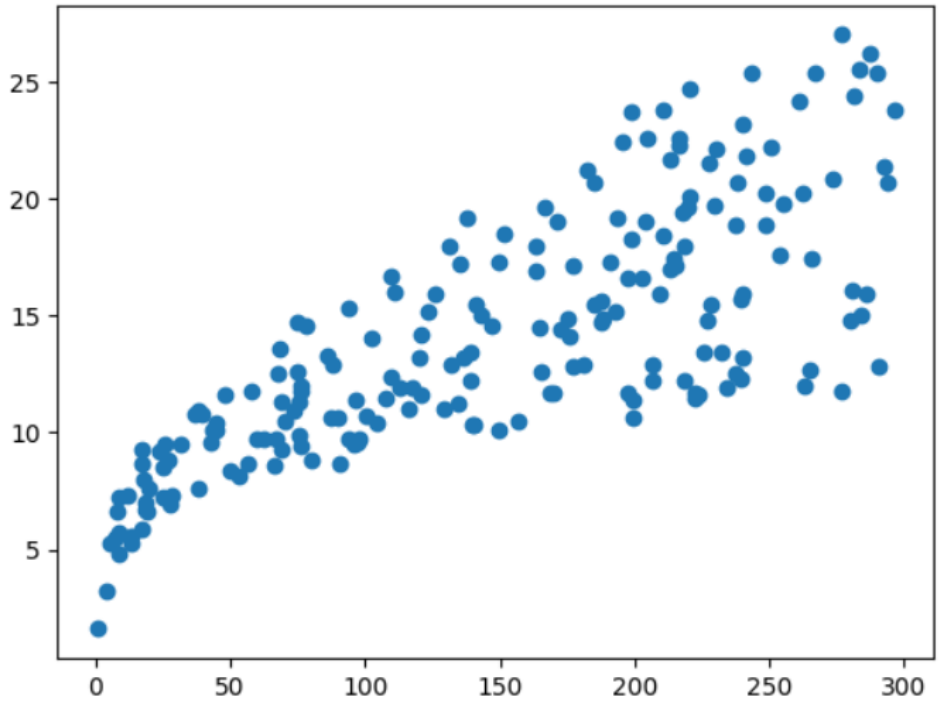
Portafolio de Análisis

Josemaría Robledo Lara

A01612376

12/09/2023

El dataset utilizado contiene la información del número de televisores y la cantidad de ventas de éstos. Al ser ambos valores numéricos y al no existir outliers dentro de los datos, no fue necesario aplicar la técnica de “Data Cleaning”. El objetivo es conocer la tendencia de las ventas y generar un modelo de predicción fiable.



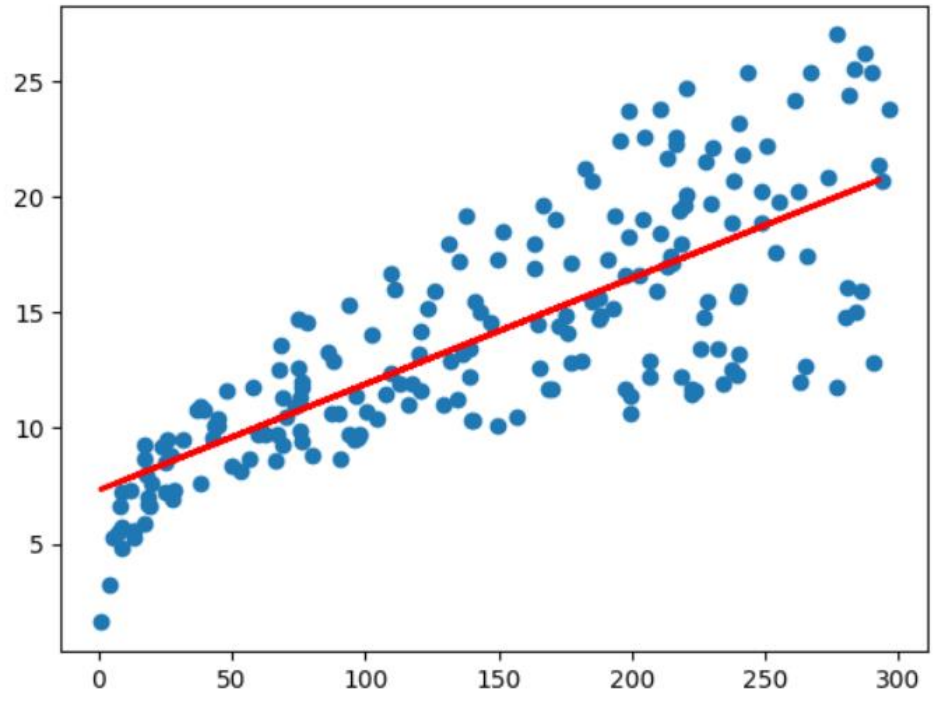
En el gráfico de dispersión se demuestra una tendencia positiva, esto debido a que van en aumento las ventas por la cantidad de televisores.

Se dividen los datos en dos conjuntos: X_train y y_train (conjunto de entrenamiento) y X_test y y_test (conjunto de prueba). Esto es fundamental para evaluar la capacidad predictiva del modelo, ya que se entrena en un conjunto de datos y se evalúa en otro para comprobar su capacidad para generalizar patrones en datos desconocidos. Igualmente, ayuda a evitar el sobreajuste.

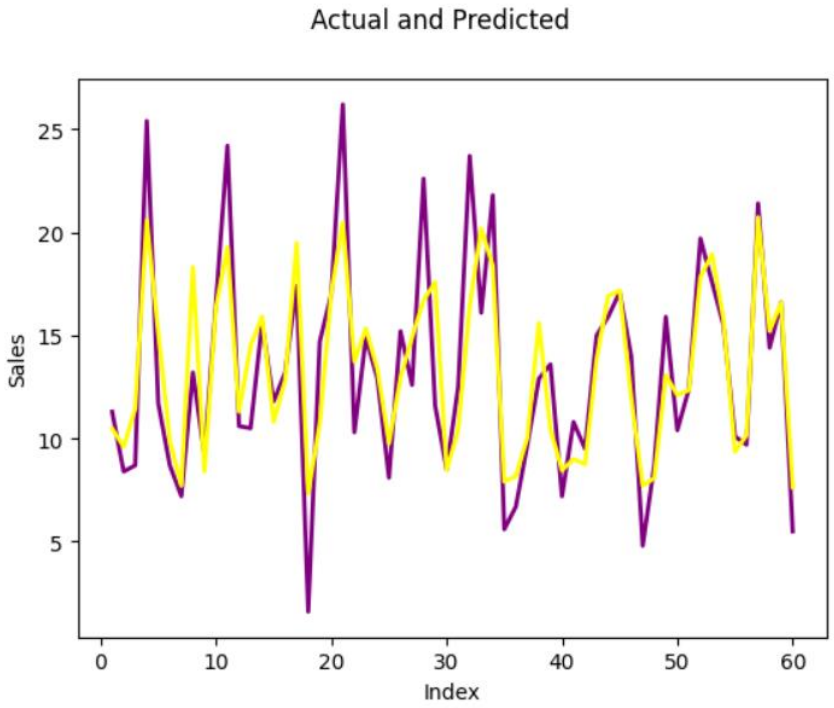
El modelo utilizado para este caso es “Linear Regression”, debido a su facilidad de interpretación, su eficiencia computacional y su bajo riesgo

de overfitting. Además, servirá a futuro como una línea base para comparar con modelos más complejos, lo que demostrará si realmente es de valor apuntar por complejidades mayores.

Se calcula el coeficiente de la regresión, porque cada coeficiente está asociado con una característica específica y representa la contribución relativa de esa característica a la predicción de la variable objetivo. Un coeficiente positivo significa que un aumento en esa característica aumenta la variable objetivo, mientras que un coeficiente negativo indica lo contrario. Esta información es crucial para interpretar el modelo y tomar decisiones basadas en los datos.



Se genera una gráfica de comparación entre los datos reales y las predicciones generadas por el modelo.



Se aprecia que los datos actuales y predichos no son completamente iguales. Para calcular qué tan lejos se está del original, se usa el cálculo del “Error Cuadrático Medio” (MSE) y el coeficiente de determinación (R^2). Esto debido al uso de Linear Regression, herramientas típicas para calcular la calidad de ajuste del modelo.

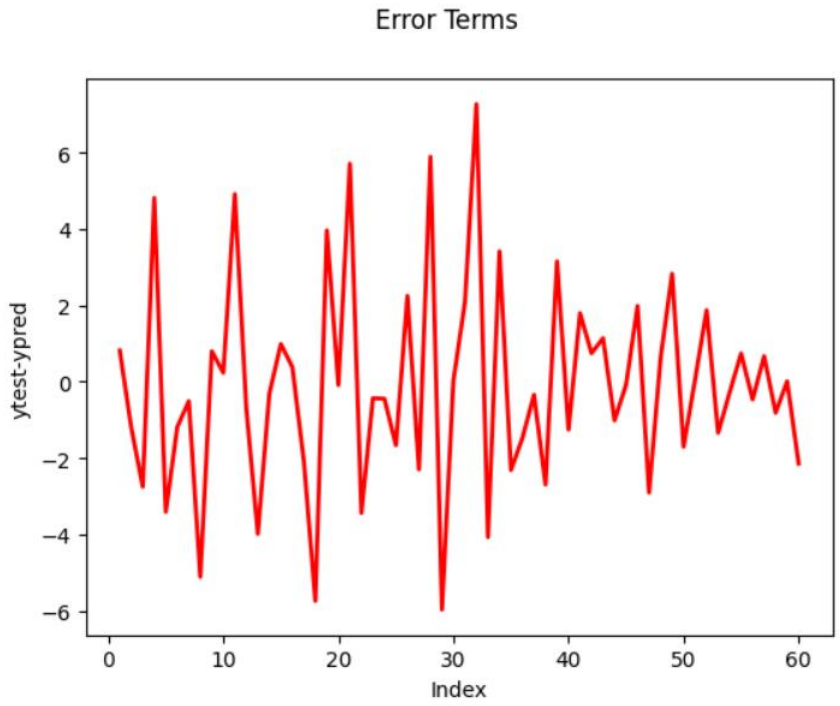


Gráfico de los términos de error entre las ventas reales y las predicciones

El valor del coeficiente de determinación (en porcentaje) es de un 72.5%, este porcentaje indica la tasa de éxito del modelo. Interpretándose como un desempeño del tipo undefitting, debido a la complejidad del modelo, no se es capaz de hacer una predicción completamente fiable de las ventas de los televisores. Sin embargo, se concluye que la tendencia es un aumento dentro de las ventas por la cantidad de televisores.