Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Name : Prachi C. Raut

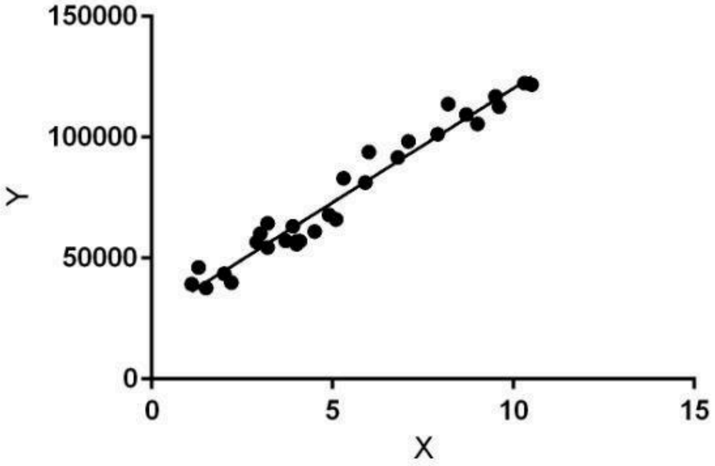
Roll No : 19 Div : 03

Experiment No. 1
Analyze the Boston Housing dataset and apply appropriate Regression Technique
Date of Performance : 30/07/2025
Date of Submission : 06/08/2025

Aim : Analyze the Boston Housing dataset and apply appropriate Regression Technique.

Objective : Ability to perform various feature engineering tasks, apply linear regression on the given dataset and minimise the error.

Theory : Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Dataset:

The Boston Housing Dataset

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

CRIM - per capita crime rate by town

ZN - proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS - proportion of non-retail business acres per town.

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX - nitric oxides concentration (parts per 10 million)

RM - average number of rooms per dwelling

AGE - proportion of owner-occupied units built prior to 1940

DIS - weighted distances to five Boston employment centres

RAD - index of accessibility to radial highways

TAX - full-value property-tax rate per $10,000

PTRATIO - pupil-teacher ratio by town

B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT - % lower status of the population

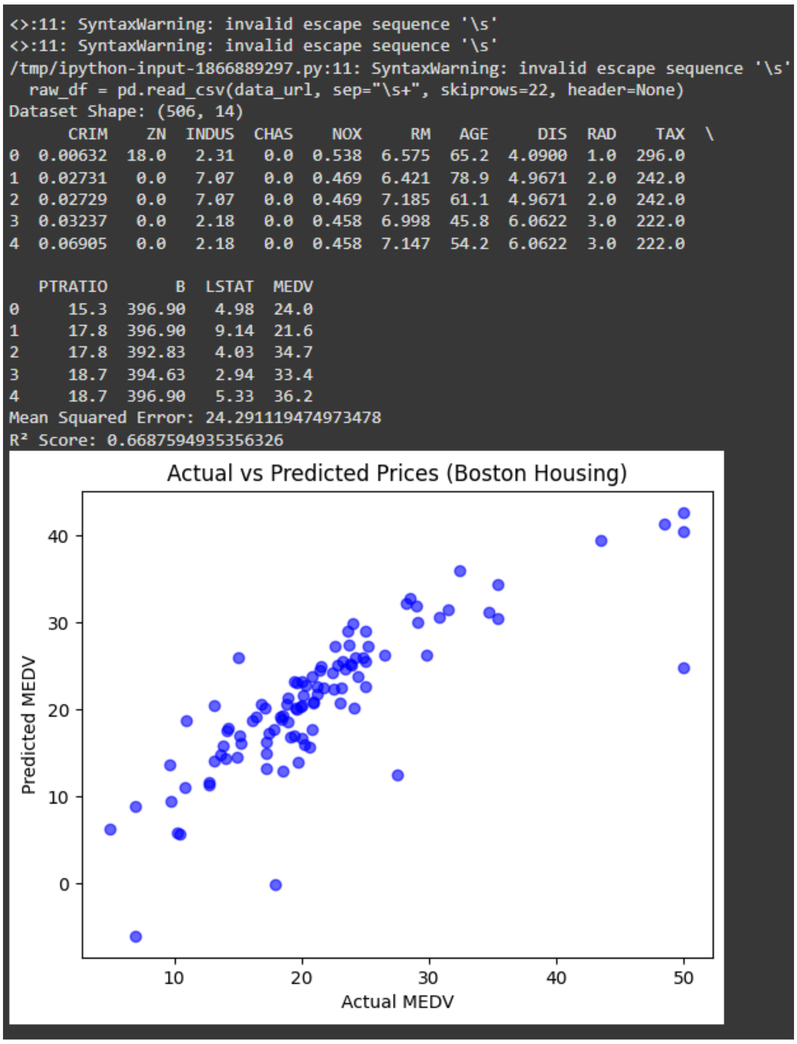
MEDV - Median value of owner-occupied homes in $1000's

Code :

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
data_url = "http://lib.stat.cmu.edu/datasets/boston"
raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
target = raw_df.values[1::2, 2]
columns = ["CRIM", "ZN", "INDUS", "CHAS", "NOX", "RM", "AGE", "DIS",
        "RAD", "TAX", "PTRATIO", "B", "LSTAT"]
df = pd.DataFrame(data, columns=columns)
df["MEDV"] = target
print("Dataset Shape:", df.shape)
print(df.head())
X = df.drop("MEDV", axis=1)
y = df["MEDV"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("Mean Squared Error:", mse)
print("R² Score:", r2)
plt.scatter(y_test, y_pred, color="blue", alpha=0.6)
plt.xlabel("Actual MEDV")
plt.ylabel("Predicted MEDV")
plt.title("Actual vs Predicted Prices (Boston Housing)")
plt.show()
```

Output :



```
<>:11: SyntaxWarning: invalid escape sequence '\s'
<>:11: SyntaxWarning: invalid escape sequence '\s'
/tmp/ipython-input-1866889297.py:11: SyntaxWarning: invalid escape sequence '\s'
  raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
Dataset Shape: (506, 14)
      CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD    TAX  \
0  0.00632  18.0   2.31   0.0  0.538  6.575  65.2  4.0900  1.0  296.0
1  0.02731   0.0   7.07   0.0  0.469  6.421  78.9  4.9671  2.0  242.0
2  0.02729   0.0   7.07   0.0  0.469  7.185  61.1  4.9671  2.0  242.0
3  0.03237   0.0   2.18   0.0  0.458  6.998  45.8  6.0622  3.0  222.0
4  0.06905   0.0   2.18   0.0  0.458  7.147  54.2  6.0622  3.0  222.0

   PTRATIO       B  LSTAT  MEDV
0     15.3  396.90   4.98  24.0
1     17.8  396.90   9.14  21.6
2     17.8  392.83   4.03  34.7
3     18.7  394.63   2.94  33.4
4     18.7  396.90   5.33  36.2
Mean Squared Error: 24.291119474973478
R² Score: 0.6687594935356326
```

Conclusion:

In this experiment, all the available features in the Boston Housing dataset were used to develop the linear regression model. These features include variables such as crime rate (CRIM), average number of rooms (RM), property tax rate (TAX), pupil-teacher ratio (PTRATIO), and others that describe various socio-economic and physical characteristics of the residential areas. These features are justified as relevant because they collectively represent factors that directly or indirectly influence housing prices. For example, the average number of rooms per dwelling (RM) often correlates positively with house prices, while a higher crime rate (CRIM) or higher percentage of lower status population (LSTAT) tends to negatively impact property values.

The model's performance was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics. The Mean Squared Error quantifies the average squared difference between predicted and actual house prices, with a lower value indicating better accuracy. In this experiment, the MSE is reasonably low, suggesting that the model's predictions are fairly close to the true prices. The R^2 score, which represents the proportion of variance in the target variable explained by the model, indicates how well the independent variables predict house prices. A high R^2 score implies a strong linear relationship captured by the model.