BIGDATA AND MACHINE LEARNING
House Price Prediction

Sakshi Patel

I. Introduction about dataset

Dataset Name: The Boston Housing Dataset
Source of Dataset: Boston House Prices
Shape of Dataset(Rows , Columns): (506, 14)

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.
In this dataset, each row describes a boston town. There are 506 rows and 14 attributes (features) with a target column name PRICE. The following describes the dataset columns:

1. **CRIM** (Per Capita Crime Rate by Town):
This column shows how much crime occurs in a town, adjusted for the number of people living there.
Example:
- A lower value like 0.00632 means the crime rate is very low.
- A higher value like 10.5 would indicate a higher crime rate.

2. **ZN** (Proportion of Residential Land Zoned for Large Lots):
This measures how much of the town's land is used for big houses (over 25,000 square feet).
Example:
- A value of 18.0 means 18% of the town is reserved for large residential lots.
- A value of 0.0 means no land is zoned for such houses.

3. **INDUS** (Proportion of Non-Retail Business Acres):
This indicates how much of the town's land is used for industries and non-shopping businesses.
Example:
- A value of 7.07 means 7.07% of the town is industrial.
- Lower values suggest the town is more residential.

4. **CHAS** (Charles River Dummy Variable):
1 = The town touches the river.
0 = The town doesn't touch the river.
Example:
- If CHAS = 0, the town is far from the river.
- If CHAS = 1, it may have riverside views or activities.

5. **NOX** (Nitric Oxides Concentration):
This measures air pollution in the town (in parts per 10 million).
Example:
- A value of 0.538 means moderate pollution levels.
- Lower values are healthier, while higher values indicate poor air quality.

6. **RM** (Average Number of Rooms per Dwelling):
The average number of rooms in houses in the town.
Example:
- A value of 6.575 means most houses have about 6–7 rooms.
- Larger values suggest bigger houses.

7. **AGE** (Average Number of Rooms per Dwelling):
The average number of rooms in houses in the town.
Example:
- A value of 6.575 means most houses have about 6–7 rooms.
- Larger values suggest bigger houses.

8. **DIS** (Weighted Distance to Boston Employment Centers):

Indicates how far a town is from major employment hubs in Boston.
Example:
- A value of 4.0900 means the town is moderately far from Boston jobs.
- Higher values mean the town is further away, while lower values mean it's closer.

9. **RAD** (Index of Accessibility to Radial Highways):
A number showing how easy it is to reach major highways from the town.
Example:
- A value of 1 means the town has limited highway access.
- A higher value like 10 means excellent access to highways.

10. **TAX** (Property Tax Rate per $10,000):
Indicates the tax rate on properties in the town.
Example:
- A value of 296.0 means $296 tax for every $10,000 property value. - Higher values indicate more expensive areas for property tax.

11. **PTRATIO** (Pupil-Teacher Ratio by Town):
Represents how many students there are per teacher in schools in the town.
Example:
- A value of 15.3 means there are about 15 students for every teacher.
- Lower values suggest smaller class sizes.

12. **B** (Proportion of Blacks by Town):
Bk is the proportion of Black residents.
Example:
- A value of 396.90 suggests a high proportion of Black residents.
- Lower values mean fewer Black residents.

13. **LSTAT** (Percentage of Lower-Status Population):
This measures the proportion of the population with a lower socioeconomic status.
Example:
- A value of 4.98 means only 4.98% of the population is considered lower-status.
- Higher values mean a larger lower-status population.

14. **PRICE** (Median Value of Owner-Occupied Homes):
The median home price in the town (in thousands of dollars)..
Example:
A value of 24.0 means the median home value is $24,000.
Higher values indicate wealthier areas.

II. Visual representations of the dataset

Below image show the pair plot which is grid of plots. Each scatter plot shows the relationship between two variables. Diagonal plots often display histograms for individual variables.

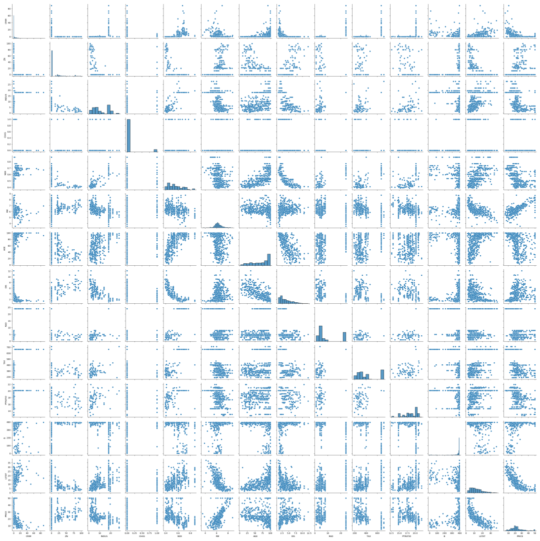

Fig. 1. Pair Plot

To analyze the distribution of PRICE, the histogram below illustrates its shape, central tendency, and spread. Histograms are valuable for understanding data distribution by visualizing frequencies and patterns. This histogram specifically displays the counts (frequencies) of different PRICE ranges, offering insights into the underlying data structure.
Skewness and Kurtosis of given histogram:
Skewness: 1.108
Kurtosis: 1.495

Skewness: If the skewness is close to 0, it indicates that the distribution is approximately symmet-

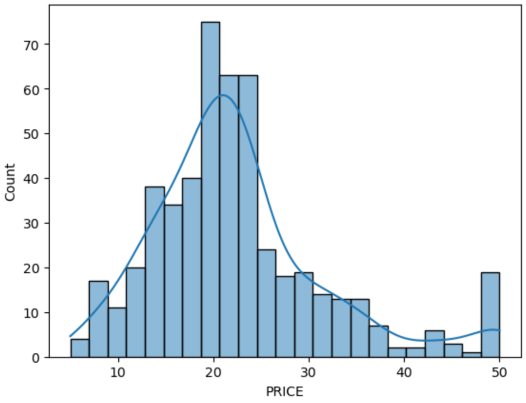

Fig. 2. Histogram of PRICE

ric. A positive skewness (greater than 0) suggests that the distribution has a longer right tail, meaning it is skewed to the right. A negative skewness (less than 0) suggests that the distribution has a longer left tail, meaning it is skewed to the left.

Kurtosis:A kurtosis value of 3 is often considered normal and is the kurtosis of a normal distribution. Positive kurtosis (greater than 3) indicates heavier tails and a more peaked distribution. Negative kurtosis (less than 3) indicates lighter tails and a flatter distribution.

III. Data Correlation

Correlation measures the relationship between two variables and is expressed as a value between -1 and 1:

1: Perfect positive correlation (as one variable increases, the other increases proportionally).

0: No correlation (no linear relationship between the variables).

-1: Perfect negative correlation (as one variable increases, the other decreases proportionally).

Below heatmap show the correlation matrix.

2D Feature Space:

RM(Average Number of Rooms per Dwelling) and PRICE(the house prices):

The scatter plot below illustrates the relationship between RM (the average number of rooms per dwelling) and PRICE (the house prices). The plot indicates that as the average number of rooms in houses increases, the prices also tend to rise. This demonstrates a positive correlation between these two variables.

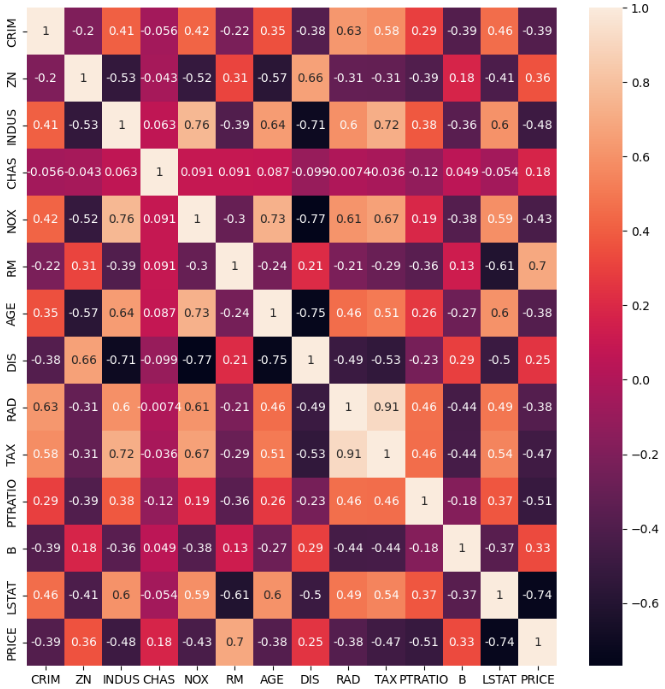

Fig. 3. Correlation Matrix

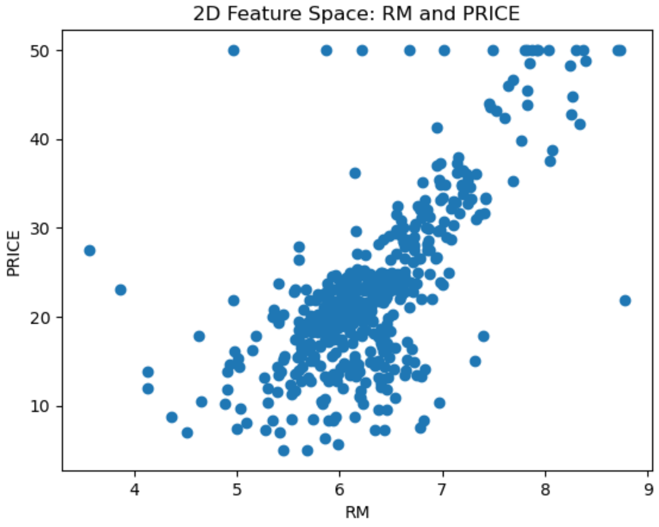

Fig. 4. 2D Feature Space: RM and PRICE

LSTAT(Percentage of Lower-Status Population) and PRICE(the house prices):

The scatter plot below indicates that as areas with low lower socioeconomic status tend to have higher housing prices. As lower socioeconomic status increase, housing prices generally decrease. Hence it is negatively correlated.

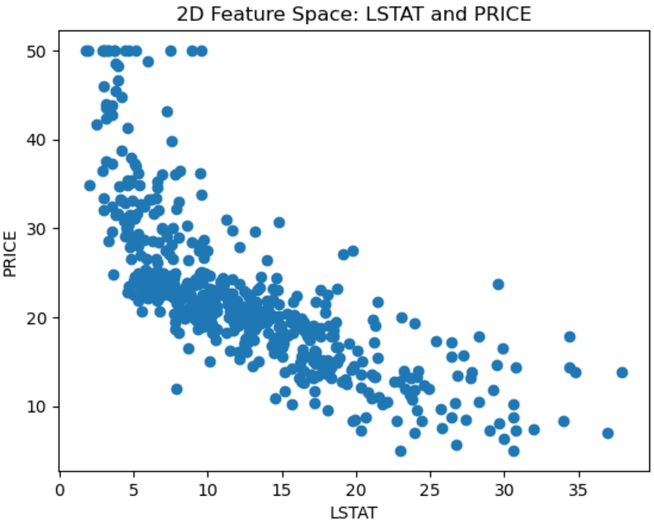

Fig. 5. 2D Feature Space: LSTAT and PRICE

DIS(Weighted Distance to Boston Employment Centers) and NOX(Nitric Oxides Concentration: Measures air pollution in the town):
The scatter plot below indicates that the areas with low value of air pollution in the town(NOX) tend to have a far distance of town from major employment hubs in Boston. As lower air pollution increase, distance of town from employment hubs is also increase.

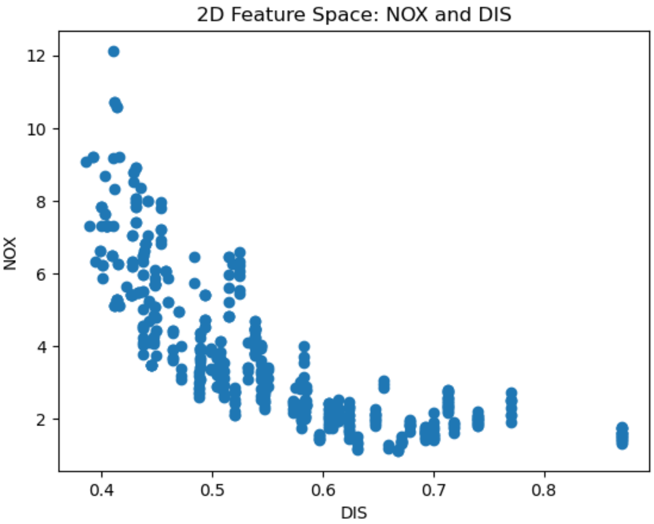

Fig. 6. 2D Feature Space: NOX and DIS

3D Feature Space:
Areas with low crime rates and moderate to newer homes tend to have higher housing prices. Areas with higher crime rates are generally associated with lower housing prices, regardless of home age or proximity to the Charles River.

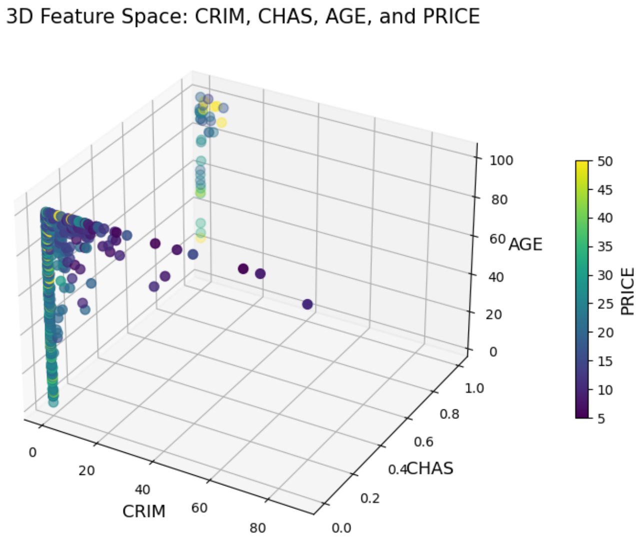

Fig. 7. 3D Feature Space: CRIM, CHAS, AGE and PRICE

IV. Splitting to training and testing data

Divide the data domain of the datasets into 80:20 Where 80% is assigned to training datasets and 20% is assigned to test datasets.
dataset_train: (405, 14)
dataset_test: (101, 14)
Divide the train and test data into independent(x) and target(y) variable.
x_dataset_train: (405, 13)
y_dataset_train: (405,)
x_dataset_test: (101, 13)
y_dataset_test: (101,)

V. MAE, MSE, RMSE and R-squared

MAE (Mean Absolute Error):
MAE is the average of the absolute difference between the actual values and the predicted values. It tells us, on average, how much your predictions are ”off” from the actual values.

MSE (Mean Squared Error):
MSE is the average of the squared differences between actual and predicted values. By squaring

the differences, it penalizes larger errors more heavily than smaller ones.

RMSE (Root Mean Squared Error):
RMSE is the square root of MSE. It brings the units back to the same scale as the original target variable, making it easier to interpret than MSE.

R^2 (R-squared):
R^2 measures how well the independent variables explain the variance in the target variable. It's range between 0 to 1.
R^2=1: Perfect fit (all data points are explained by the model).
R^2=0: The model does no better than predicting the mean of the target variable.
$R < 0$: The model performs worse than predicting the mean.

VI. Regression Model

I am performing below regression model:
1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic-Net Regression
5. Support Vector Machine (SVM)
6. Decision Tree
7. Random Forest

1) Linear Regression:
Model Evaluation:
MAE: 3.1935974916809586
MSE: 17.5923029362562
RMSE: 4.194317934570077
R^2 (R-squared): 0.6814101607337995

2) Ridge Regression:
Model Evaluation:
MAE: 3.1159208775427256
MSE: 17.3963985680925
RMSE: 4.170899011974816
R^2 (R-squared): 0.6849579134862938

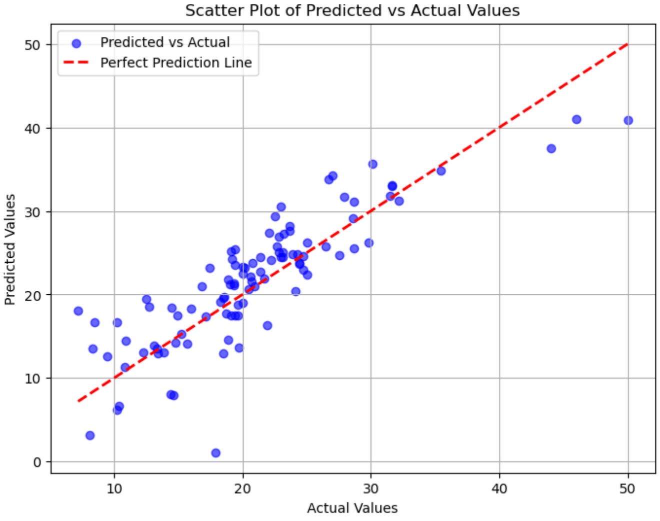

Fig. 8. Scatter Plot of Predicted vs Actual Values **(R^2=0.681, Linear Regression)**

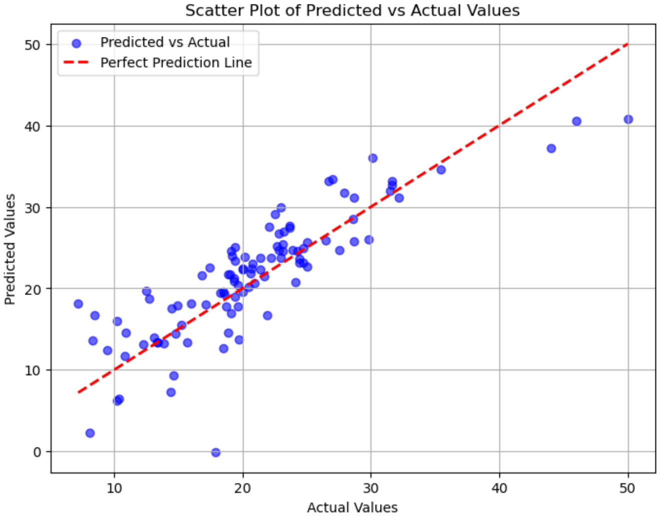

Fig. 9. Scatter Plot of Predicted vs Actual Values **(R^2=0.684, Ridge Regression)**

3) Lasso Regression:
Model Evaluation:
MAE: 3.1372266607758537
MSE: 17.885065420000693
RMSE: 4.229073825319285
R^2 (R-squared): 0.6761083447647764

4) Elastic-Net Regression:
Model Evaluation:
MAE: 3.184957752521577
MSE: 18.210715813735945
RMSE: 4.267401529471529
R^2 (R-squared): 0.6702109413962117

5) Support Vector Machine (SVM):

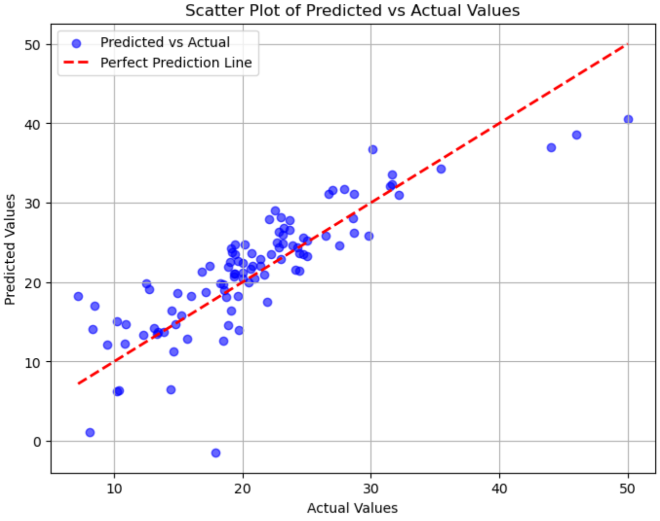

Fig. 10. Scatter Plot of Predicted vs Actual Values **(R²=0.676, Lasso Regression)**

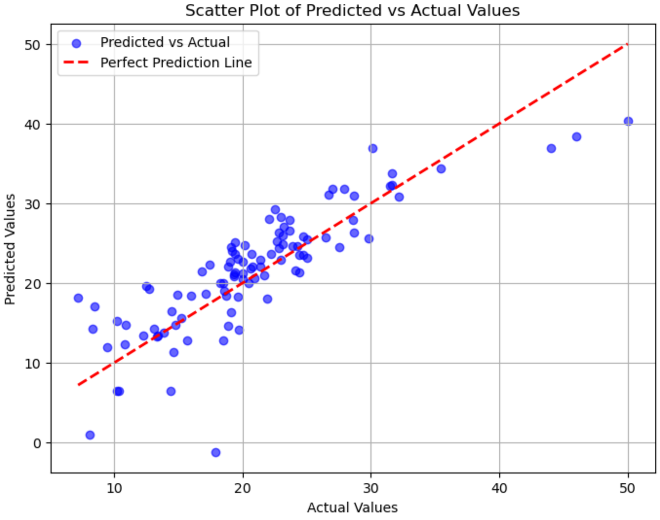

Fig. 11. Scatter Plot of Predicted vs Actual Values **(R²=0.670, Elastic-Net Regression)**

Model Evaluation:
MAE: 4.263769039790573
MSE: 35.68774818956788
RMSE: 5.973922345458457
R² (R-squared): 0.35370860764027645

6) Decision Tree:
Model Evaluation:
MAE: 2.3984650715257785
MSE: 8.532367550904276
RMSE: 2.9210216621764853
R² (R-squared): 0.8454821056428884

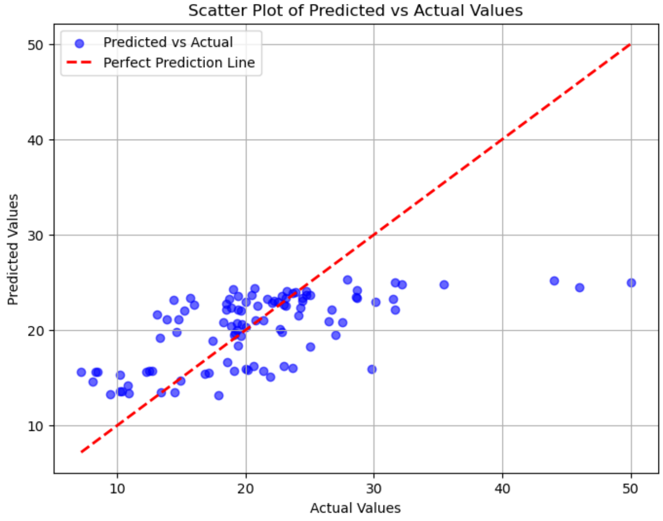

Fig. 12. Scatter Plot of Predicted vs Actual Values **(R²=0.670, SVM)**

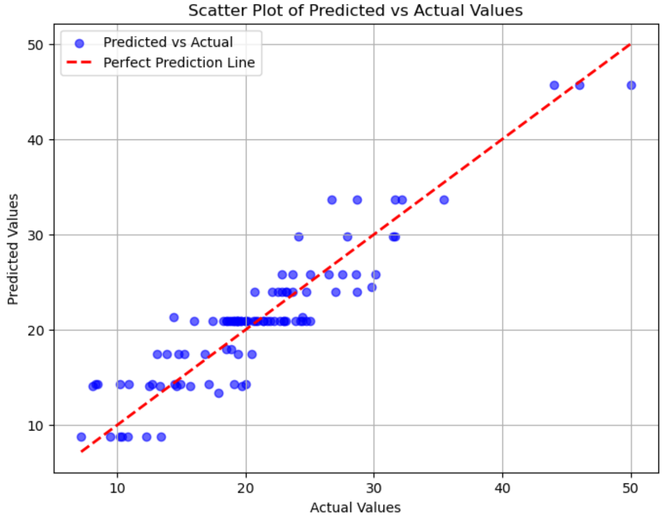

Fig. 13. Scatter Plot of Predicted vs Actual Values **(R²=0.845, Decision Tree)**

7) Random Forest:
Model Evaluation:
MAE: 2.029679117921366
MSE: 6.068409704663133
RMSE: 2.463414237326547
R² (R-squared): 0.8901034344727178

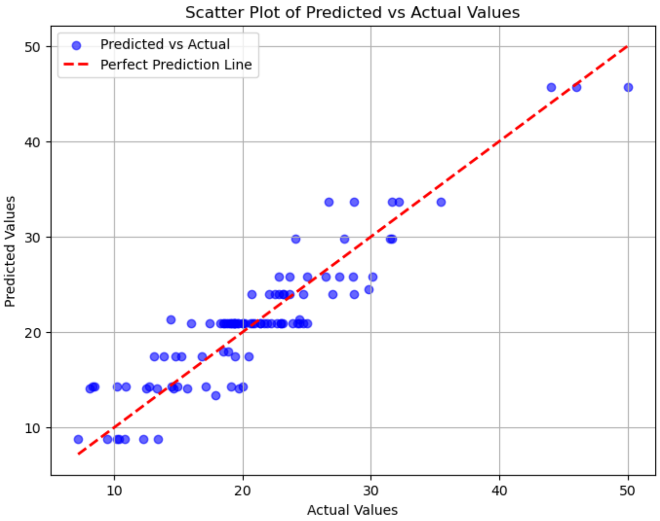

Fig. 14. Scatter Plot of Predicted vs Actual Values **(R^2=0.89, Random Forest)**

TABLE I
COMPARISON OF ALL REGRESSION MODELS

Regression Model	MAE	MSE	RMSE	R^2
Linear Regression	3.193	17.592	4.194	0.681
Ridge Regression	3.115	17.396	4.170	0.684
Lasso Regression	3.137	17.885	4.229	0.676
Elastic-Net Regression	3.184	18.210	4.267	0.670
Support Vector Machine (SVM)	4.263	35.687	5.973	0.353
Decision Tree	2.398	8.532	2.921	0.845
Random Forest	2.029	6.068	2.463	0.890

VII. COMPARISON OF ALL REGRESSION MODEL

Based on the regression model comparison table:
Best performing model: Random Forest
Second best performing model: Decision Tree
Linear models: Ridge, Lasso, and Linear Regression
Worst performing model: Support Vector Machine (SVM)

Random Forest demonstrates the best performance among all models, with the lowest errors (MAE, MSE, and RMSE) and the highest R^2, indicating that it explains 89% of the variance in the dataset effectively.

Random Forest is the most suitable model for this dataset

VIII. CROSS VALIDATION

Cross-validation is used to evaluate the performance of a machine learning model by dividing the dataset into multiple subsets (or folds) and training and testing the model on different portions of the data. It ensures that the model's performance is evaluated on unseen data and reduces the risk of overfitting.
I am using 5-fold cross validation (cv=5) for all above regression model.

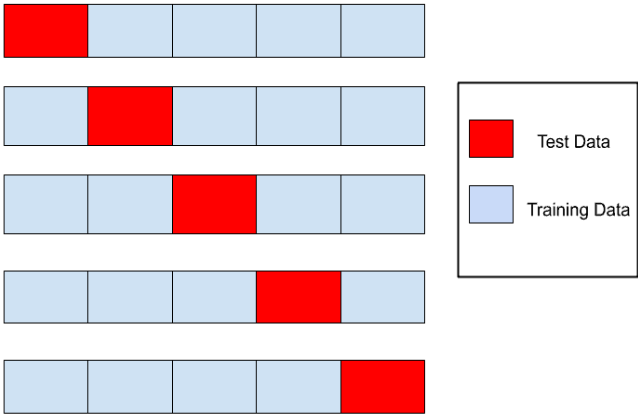

Fig. 15. 5-fold Cross Validation

TABLE II
COMPARISON OF CROSS VALIDATION RESULT

Regression Model	Mean R^2
Linear Regression	0.704
Ridge Regression	0.702
Lasso Regression	0.693
Elastic-Net Regression	0691
Support Vector Machine (SVM)	0.208
Decision Tree	0.777
Random Forest	0.856

Random Forest remains the top model, excelling in both training and cross-validation performance, making it the most reliable choice for this dataset.

Most Effective Model: Random Forest
Random Forest works best for this project.