📊 Analytical Report

House Price Prediction using Regression Models

---

1. Introduction

Regression is a supervised learning technique used to predict continuous numerical values. In this assignment, multiple regression models were implemented to predict house prices using the Boston Housing dataset. The objective was to compare different regression techniques and evaluate their performance using standard evaluation metrics.

The models implemented were:

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression

2. Dataset Description

The Boston Housing dataset contains 506 observations with 13 independent features and 1 target variable (MEDV), which represents the median value of owner-occupied homes.

Some important features include:

- CRIM – Crime rate
- RM – Average number of rooms
- NOX – Nitric oxide concentration

- LSTAT – Percentage of lower status population

The target variable:

- MEDV – Median house price (in $1000s)

Some missing values were found in the dataset and were handled using mean imputation.

---

3. Data Preprocessing

The following preprocessing steps were performed:

1. Missing values were handled using SimpleImputer (mean strategy).
2. The dataset was split into training (80%) and testing (20%) sets.
3. Feature scaling was applied using StandardScaler to ensure better performance of regularized models.
4. Polynomial features of degree 2 were generated for Polynomial Regression.

Proper preprocessing ensured that data leakage was avoided.

4. Model Evaluation Metrics

The following evaluation metrics were used:

- MAE (Mean Absolute Error): Measures average absolute error.
- MSE (Mean Squared Error): Penalizes larger errors.
-

R^2 Score: Indicates how well the model explains variance in the target variable.

---

5. Results and Comparison

Model	MAE	MSE	R^2
Linear	3.15	25.00	0.6591
Ridge	3.15	25.01	0.6590
Lasso	3.22	26.12	0.6438
Polynomial	2.68	17.42	0.7625

Observations:

- Polynomial Regression achieved the highest R^2 score (0.7625), indicating better ability to capture non-linear relationships.
- Linear and Ridge Regression performed almost identically, suggesting minimal overfitting in the base model.
- Lasso Regression slightly underperformed due to coefficient shrinkage.
- Polynomial features significantly improved predictive performance.

---

6. Model Stability Discussion

- Ridge Regression reduces coefficient magnitude but did not significantly improve performance compared to Linear Regression.
- Lasso Regression performs feature selection by shrinking some coefficients to zero, which

may reduce performance slightly.

- Polynomial Regression increases model complexity and captures non-linear patterns, leading to better performance.

- However, polynomial models may risk overfitting if the degree is too high.

---

7. Conclusion

Among all models, Polynomial Regression performed the best with the highest R^2 score and lowest error metrics. This indicates that house prices have non-linear relationships with certain features.

Regularization techniques such as Ridge and Lasso help improve model stability but did not outperform Polynomial Regression in this case.

Overall, this study demonstrates the importance of model comparison and proper preprocessing in supervised learning tasks.