Project Description Document

Numerical dataset

General Information on Dataset

- **Dataset Name:** **housing.csv**
- **Number of Classes:** NONE
- **Labels:** NONE
- **Total Number of Samples:** **20640**
- **Sample Size (if applicable):** [numerical]
- **Samples Used:**
  - Training: 16512
  - Validation: [none]
  - Testing: 4128

Implementation Details

- **Feature Extraction Phase:**
  - **Number of Features Extracted:** 13

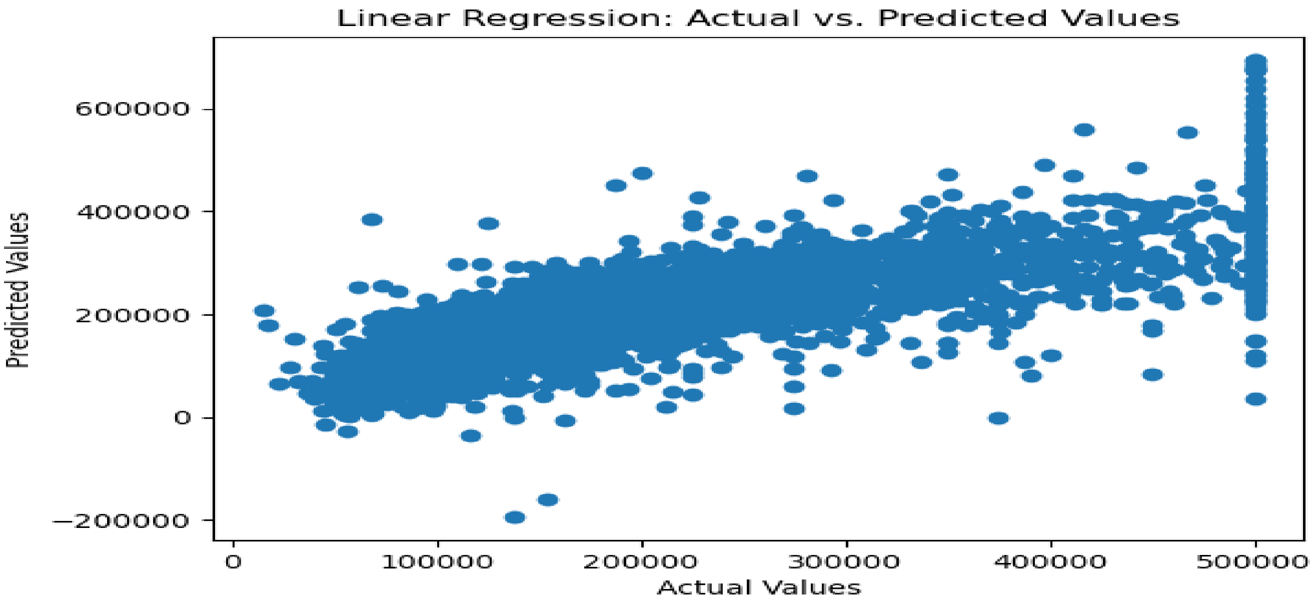
Feature Names: **['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households', 'median_income', 'ocean_proximity_<1H OCEAN', 'ocean_proximity_INLAND', 'ocean_proximity_ISLAND', 'ocean_proximity_NEAR BAY', 'ocean_proximity_NEAR OCEAN']**

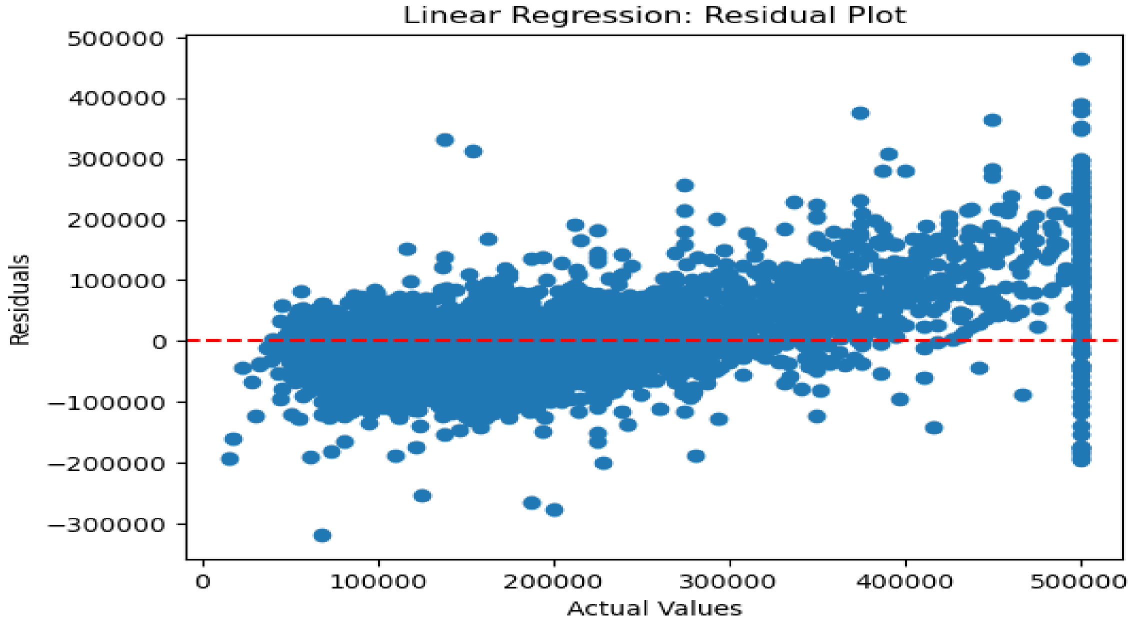
Dimension of Resulted Features: **(20640, 13)**

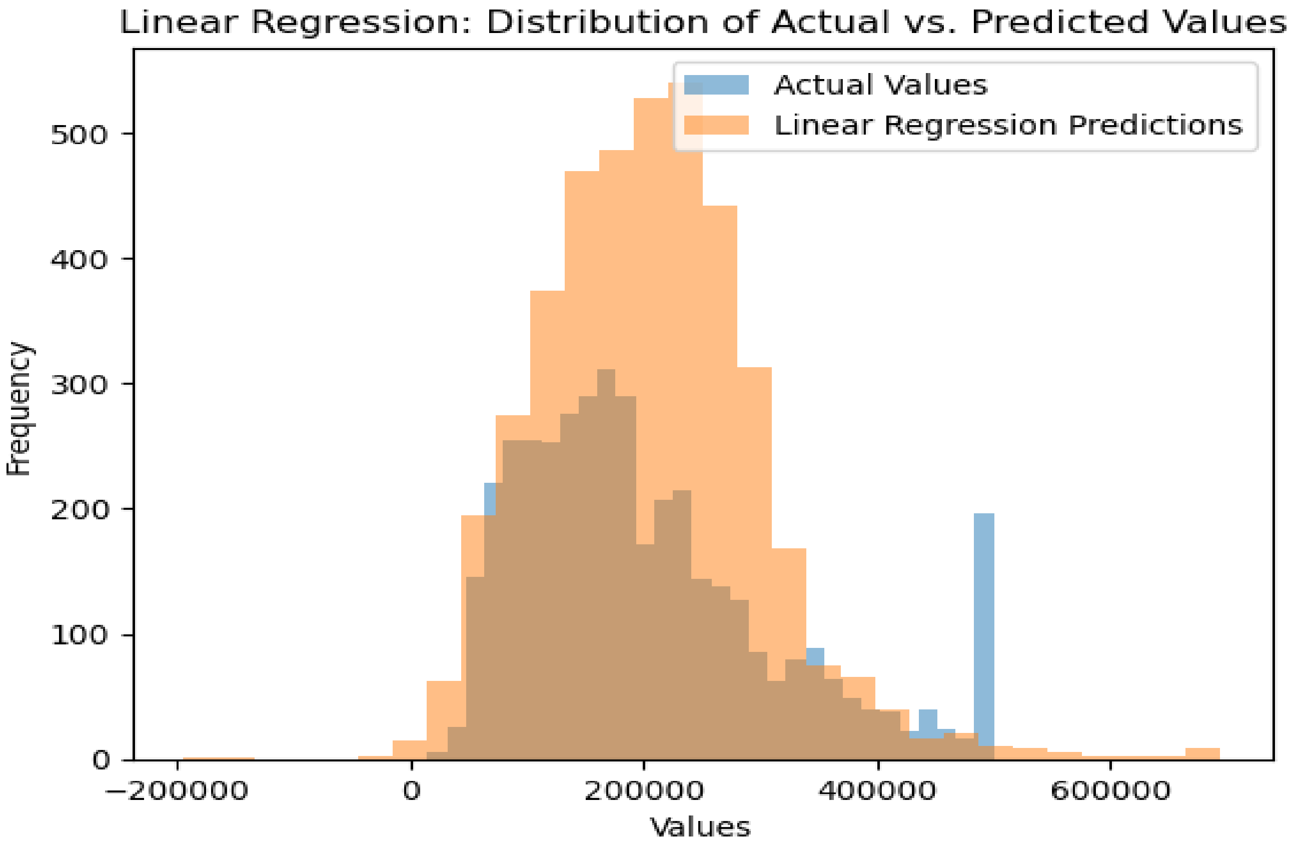
- **Cross-Validation:**
  - **Used?:** [**Yes**/No]
  - **Number of Folds:** [**5**]
  - **Training/Validation Ratio:** **4:1**
- **Hyperparameters Used:**
  - **Initial Learning Rate:** [Rate]
  - **Optimizer:** [Optimizer]
  - **Regularization:** [Type and strength]
  - **Batch Size:** [Size]
  - **Number of Epochs:** [Number]

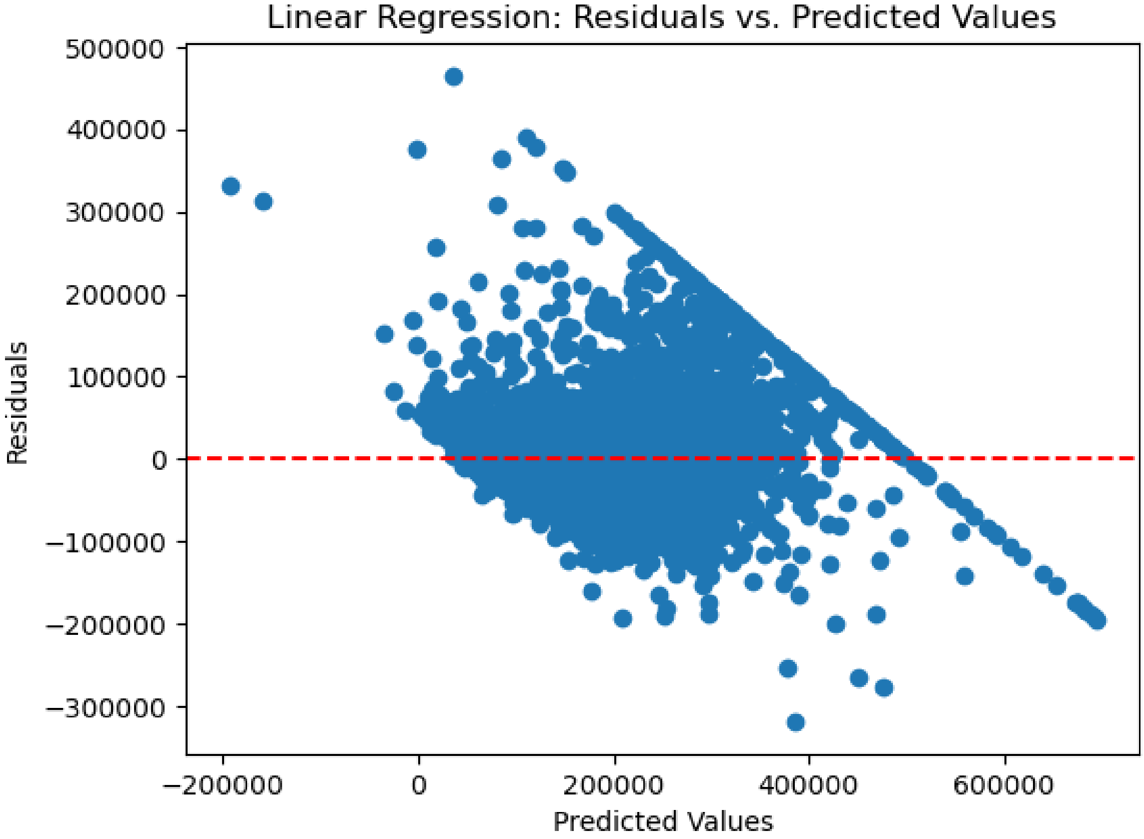
Results Details

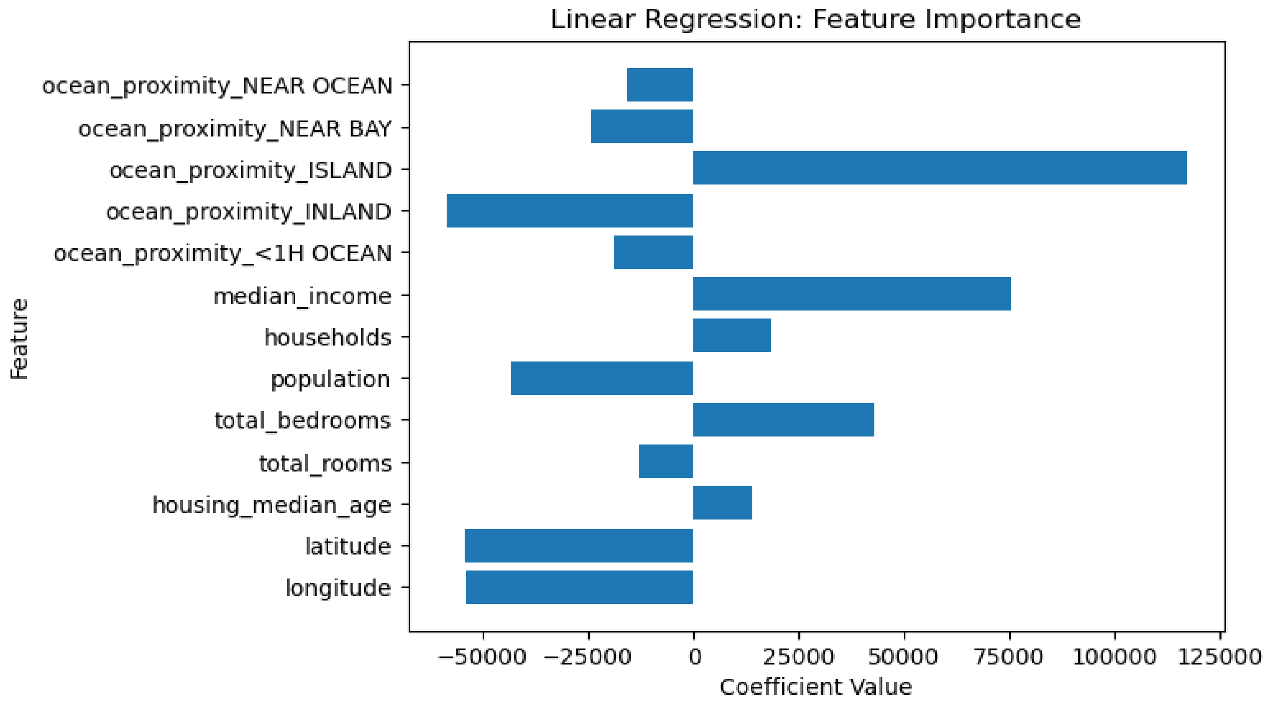
- ## Performance on Testing Data:
- **Linear Regression Score: 62.54**
- **Linear Regression MSE on Test Data: 4908476721.156613**
- **Linear Regression R-squared on Test Data: 0.6254240620553608**

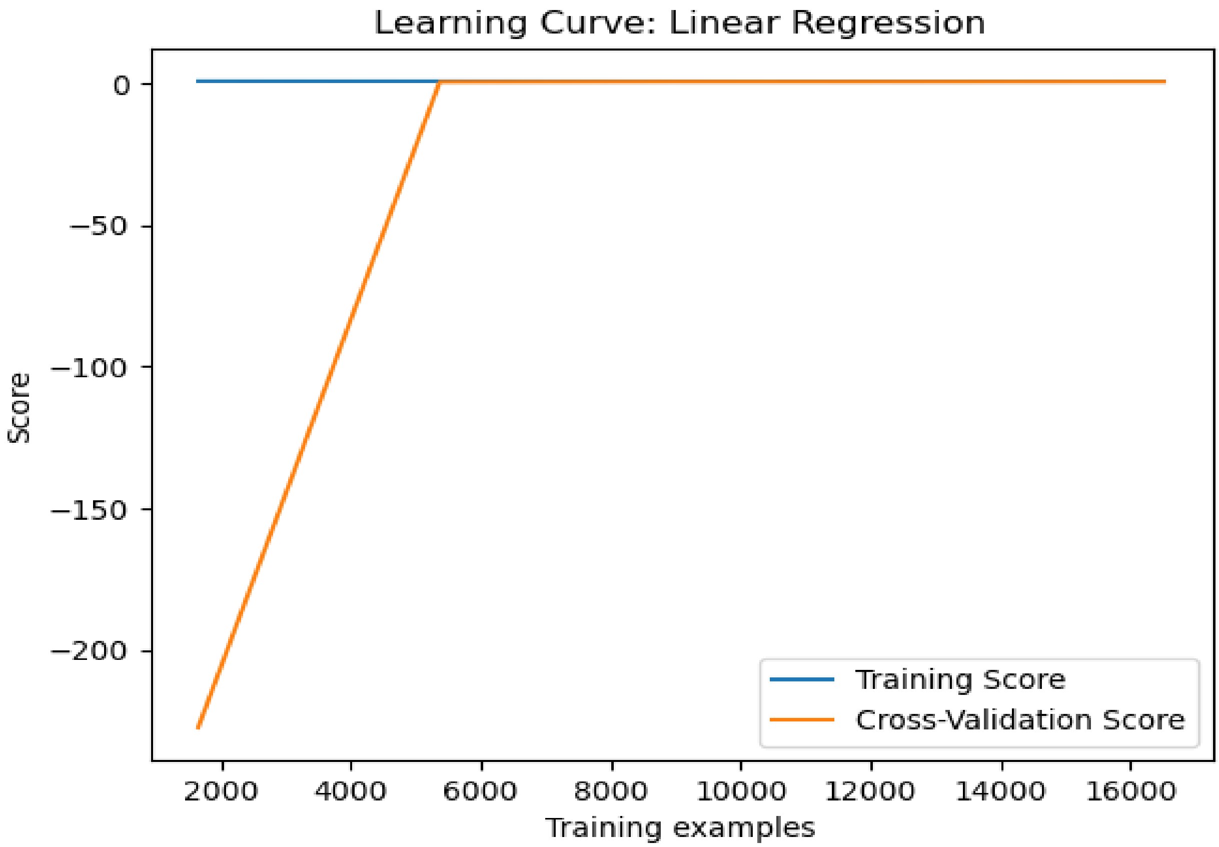


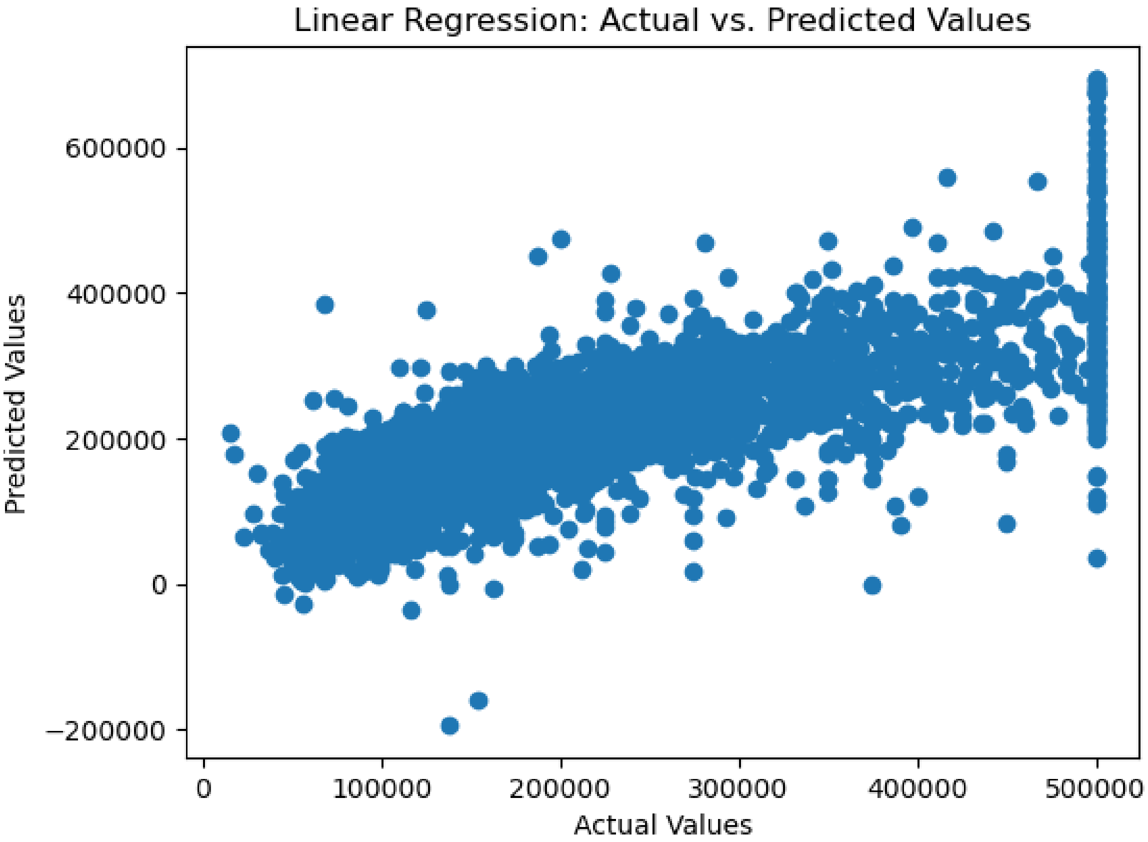



Linear Regression: Distribution of Actual vs. Predicted Values
Actual Values
Linear Regression Predictions
Frequency
0
100
200
300
400
500
−200000
0
200000
400000
600000
Values


Linear Regression: Residuals vs. Predicted Values
Residuals
500000
400000
300000
200000
100000
0
−100000
−200000
−300000
−200000
0
200000
400000
600000
Predicted Values


Linear Regression: Feature Importance
ocean_proximity_NEAR OCEAN
ocean_proximity_NEAR BAY
ocean_proximity_ISLAND
ocean_proximity_INLAND
ocean_proximity_<1H OCEAN
median_income
households
population
total_bedrooms
total_rooms
housing_median_age
latitude
longitude
Feature
−50000
−25000
0
25000
50000
75000
100000
125000
Coefficient Value


Learning Curve: Linear Regression
0
−50
−100
−150
−200
Score
2000
4000
6000
8000
10000
12000
14000
16000
Training examples
Training Score
Cross-Validation Score

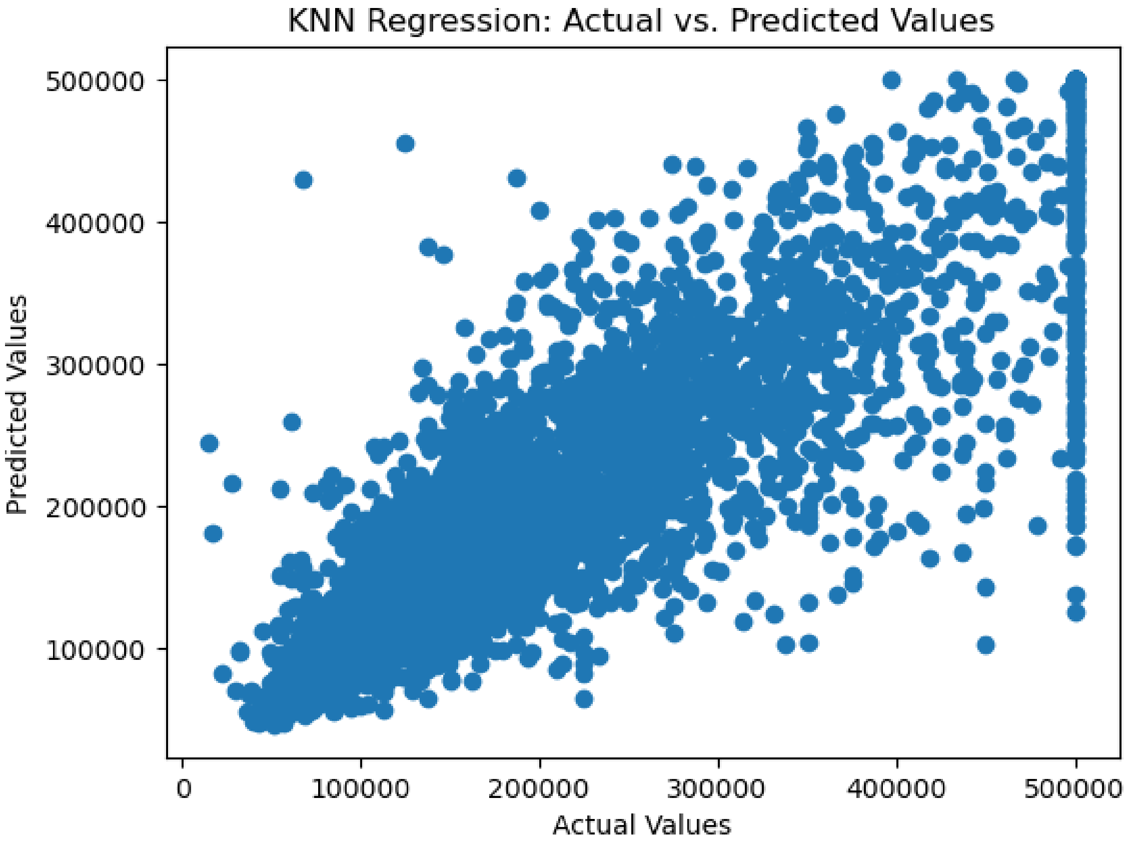

Linear Regression: Actual vs. Predicted Values
Predicted Values
600000
400000
200000
0
−200000
0
100000
200000
300000
400000
500000
Actual Values

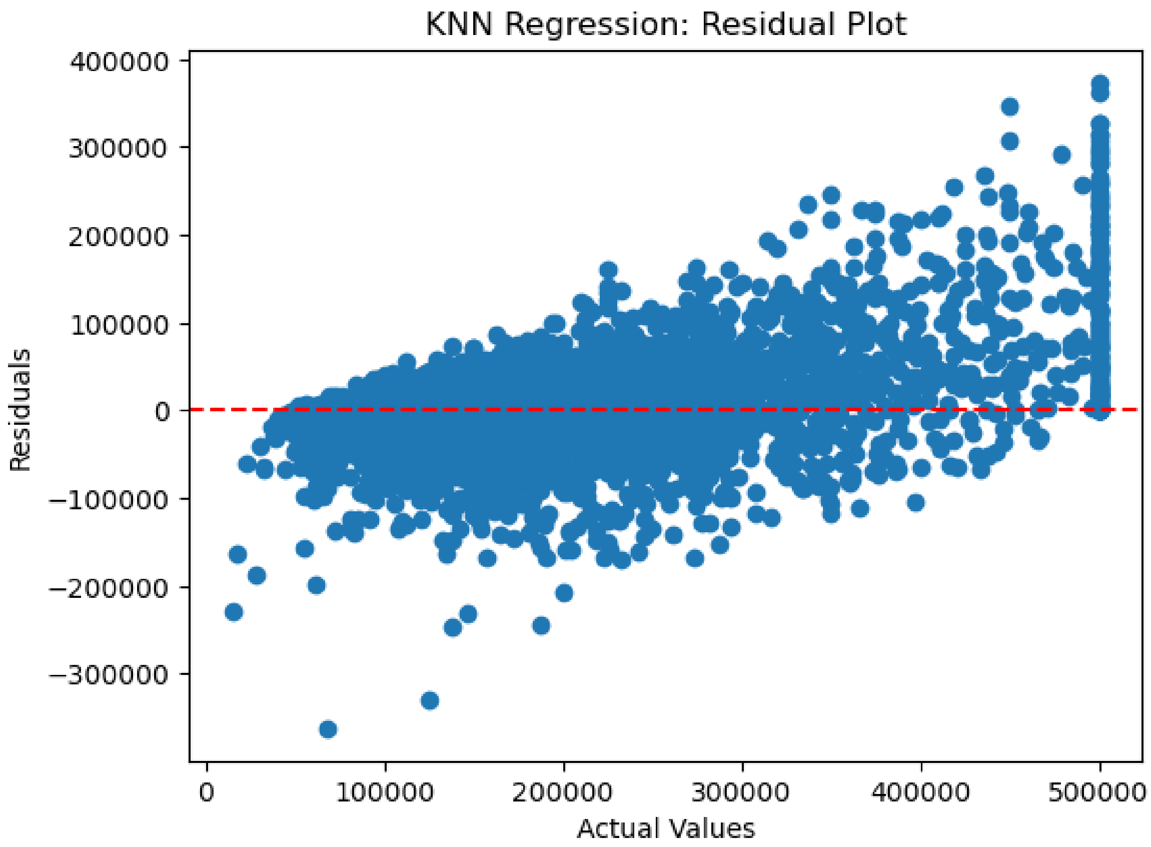
KNN

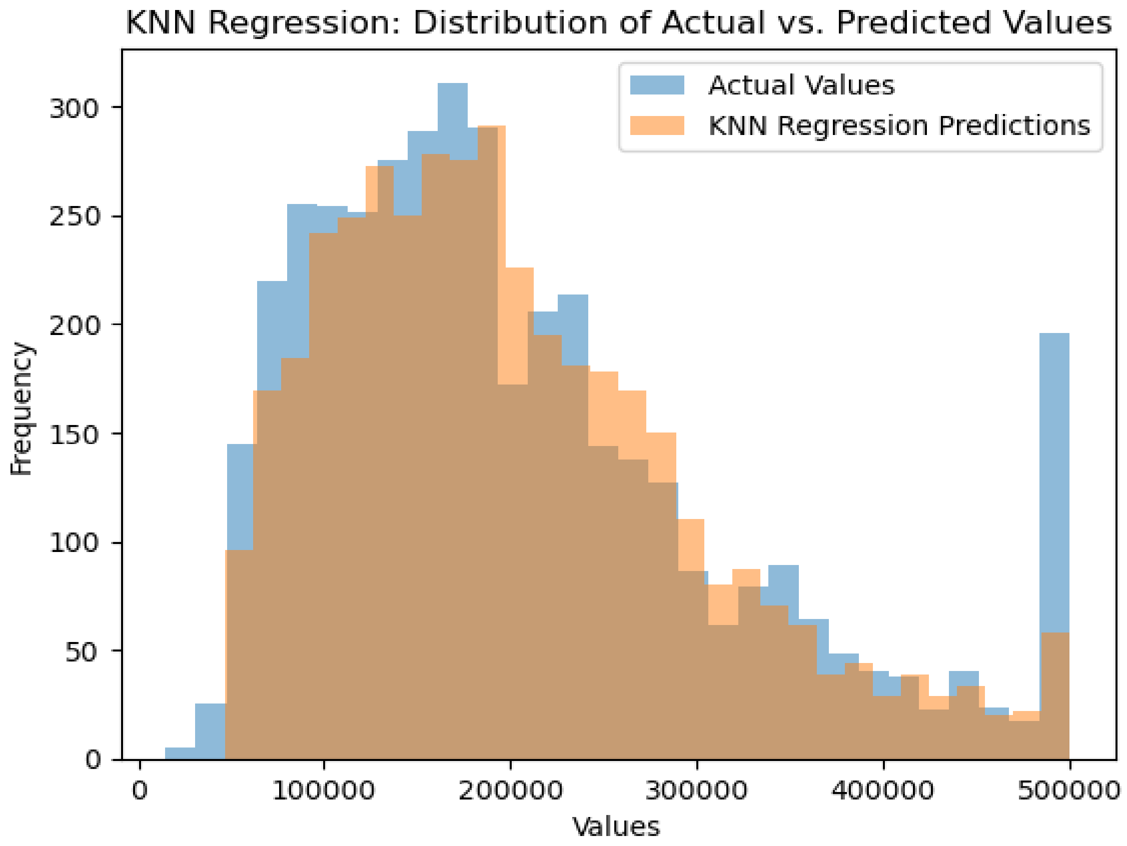
- ## Hyperparameters Used: K=5

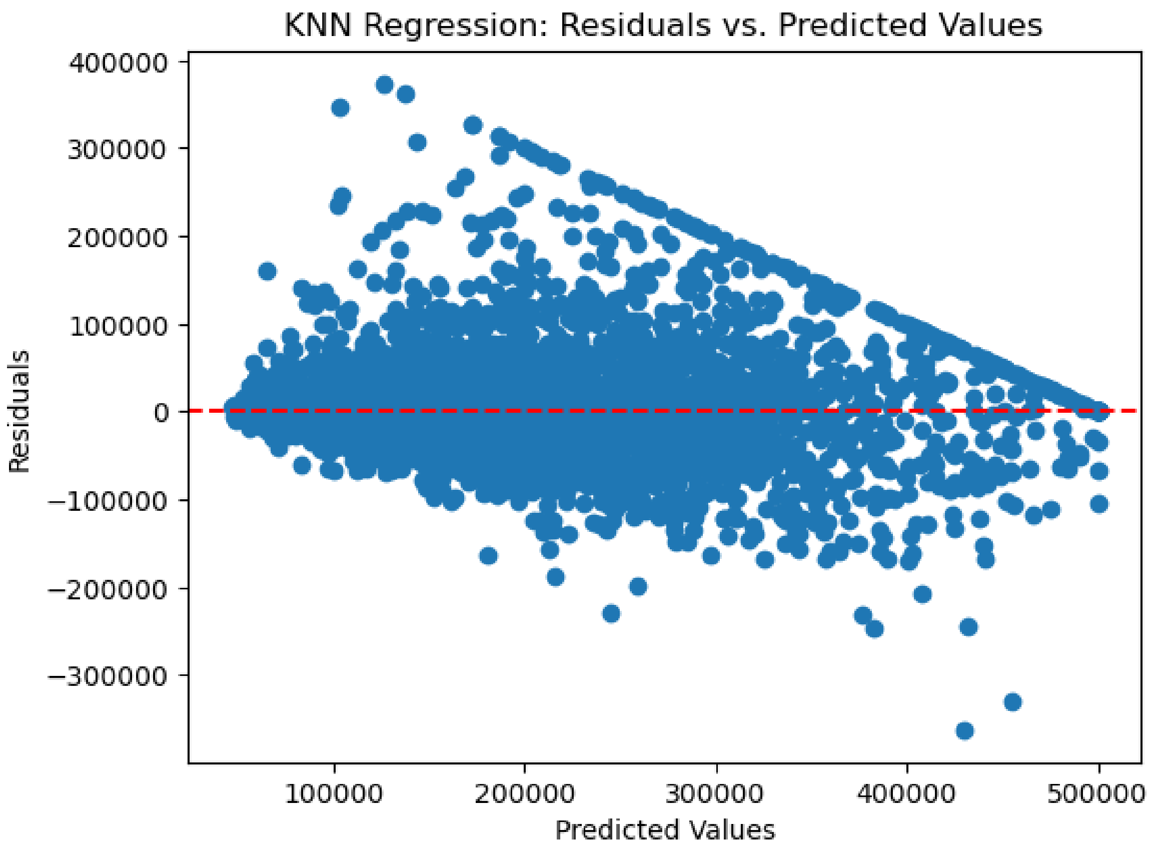
Results Details

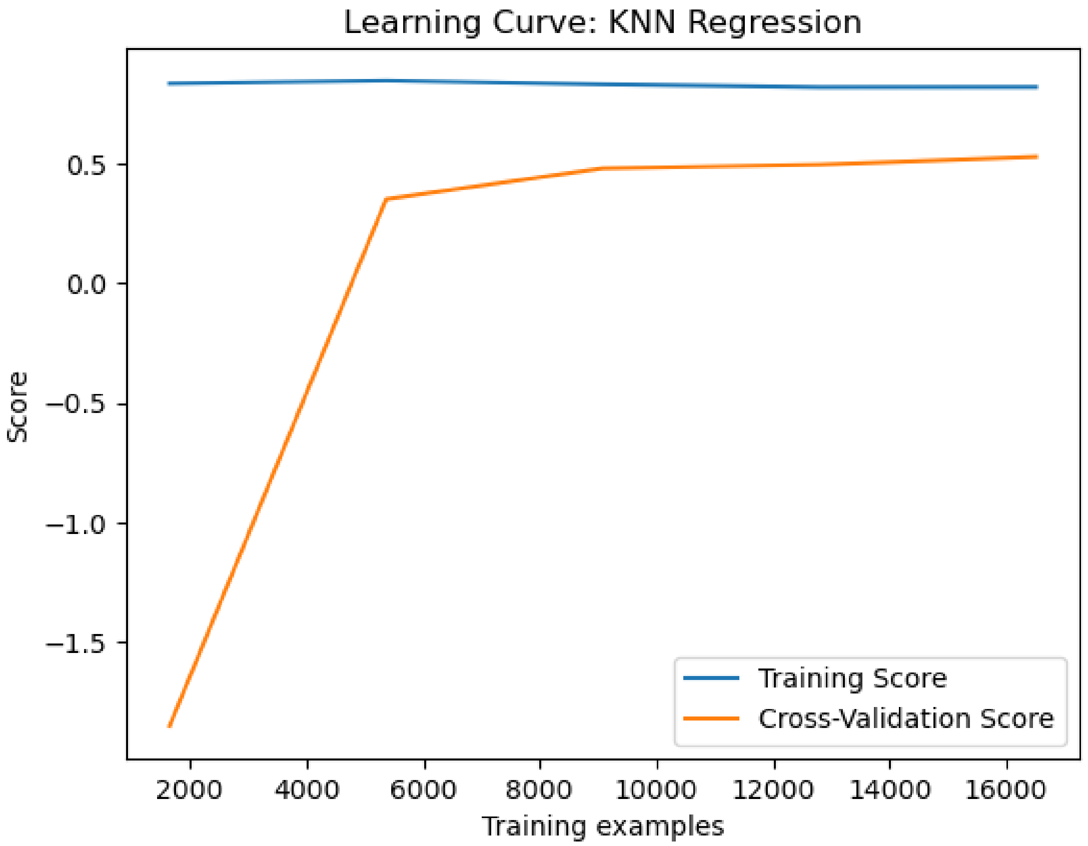
- **Performance on Testing Data:**
- **KNN Regression Score: 71.3**
- **KNN Regression MSE on Test Data: 3760982284.460552**
- **KNN Regression R-squared on Test Data: 0.7129917188518262**

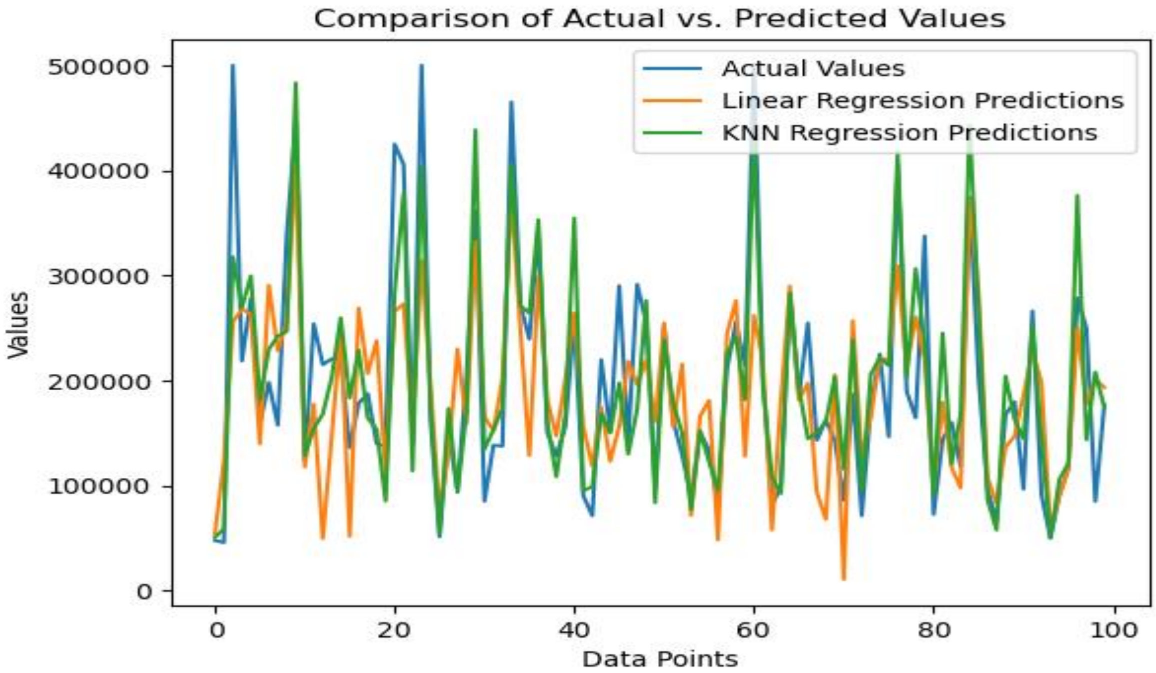


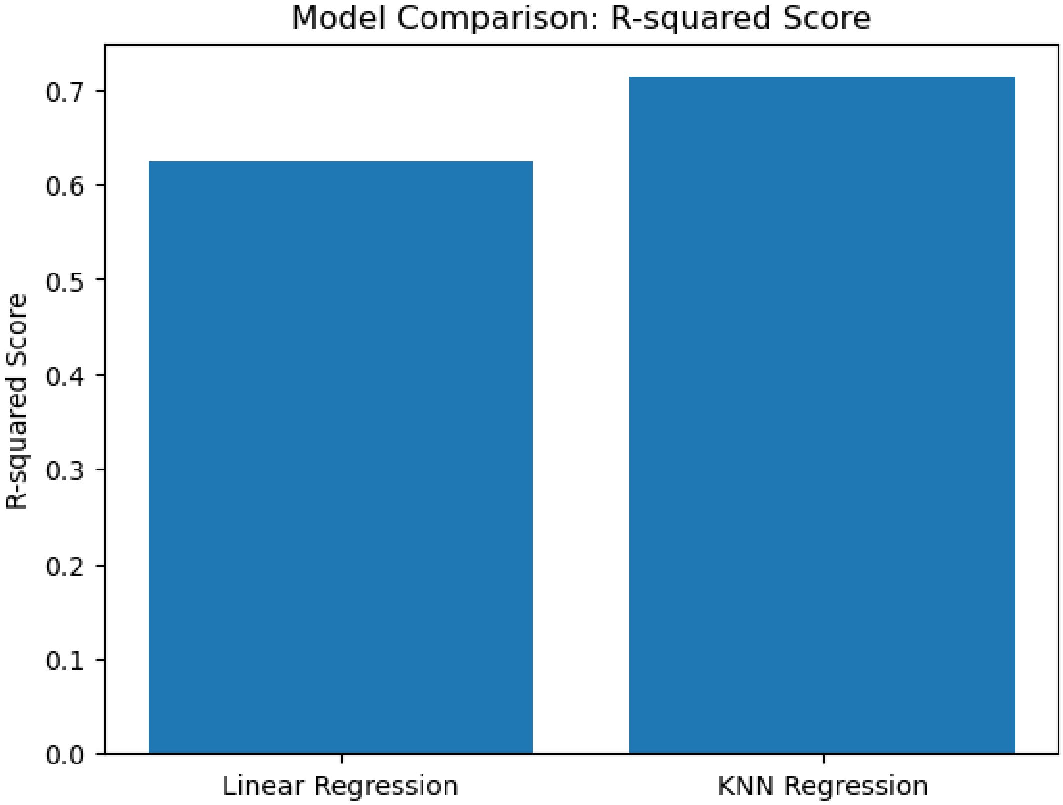

KNN Regression: Residual Plot
Residuals
400000
300000
200000
100000
0
−100000
−200000
−300000
0
100000
200000
300000
400000
500000
Actual Values


KNN Regression: Distribution of Actual vs. Predicted Values
Actual Values
KNN Regression Predictions
Frequency
300
250
200
150
100
50
0
0
100000
200000
300000
400000
500000
Values

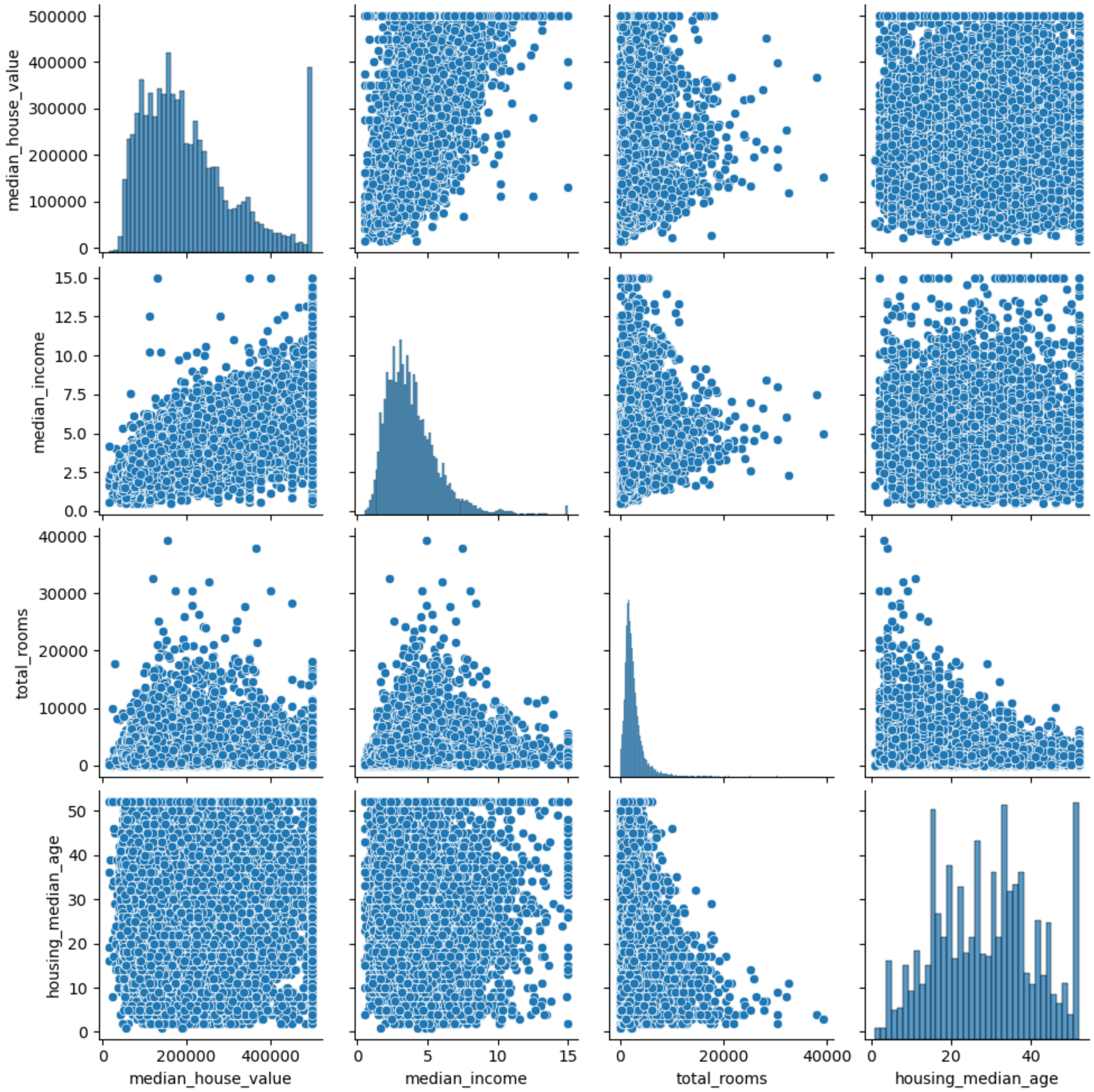

KNN Regression: Residuals vs. Predicted Values
Residuals
400000
300000
200000
100000
0
−100000
−200000
−300000
100000
200000
300000
400000
500000
Predicted Values


Learning Curve: KNN Regression
Score
0.5
0.0
−0.5
−1.0
−1.5
2000
4000
6000
8000
10000
12000
14000
16000
Training examples
Training Score
Cross-Validation Score


Comparison of Actual vs. Predicted Values
Actual Values
Linear Regression Predictions
KNN Regression Predictions
Values
500000
400000
300000
200000
100000
0
0
20
40
60
80
100
Data Points


Model Comparison: R-squared Score
R-squared Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Linear Regression
KNN Regression

Pair Plot for Selected Features



Logistic Regression Model

Image dataset

General Information on Dataset

- **Dataset Name: dataset.csv**
- **Number of Classes: 3**
- **Labels: [0 2   3]**
- **Total Number of Samples:** 35887
- **Sample Size (if applicable): (48,48)**
- **Samples Used:**
  - **Training: 2400**
  - **Validation: 1800**
  - **Testing: 300**

Implementation Details

- **Feature Extraction Phase:**
- **Number of Features Extracted:**

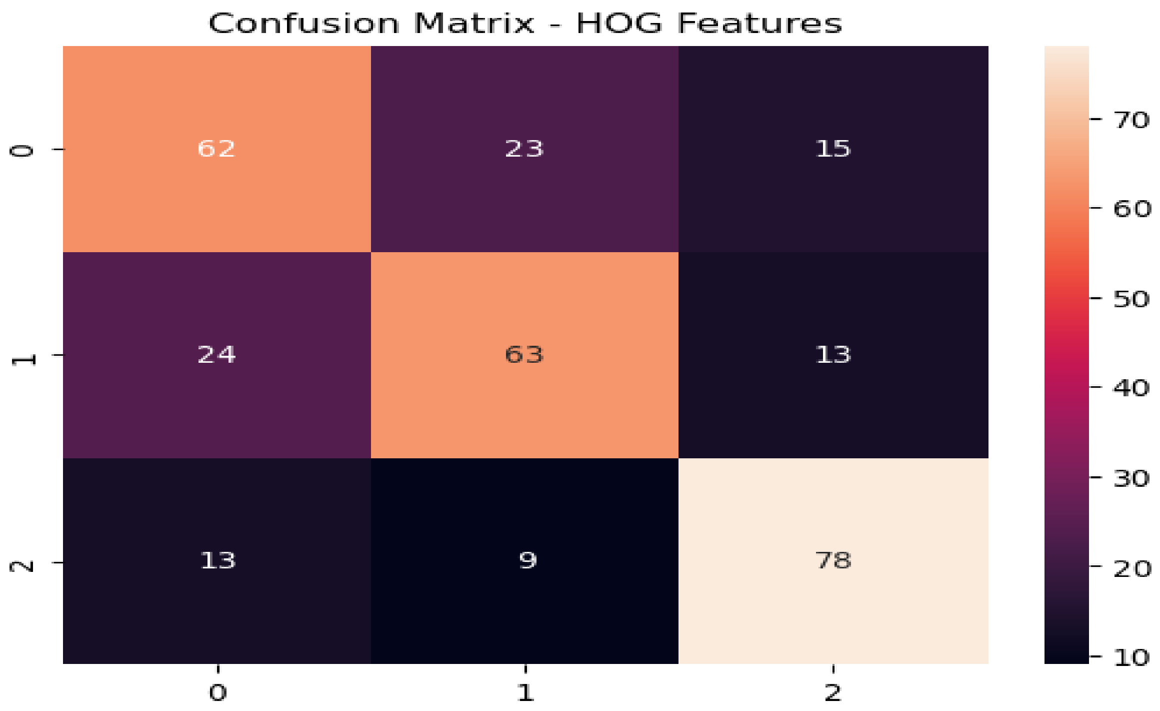
Number of features extracted per image: 800

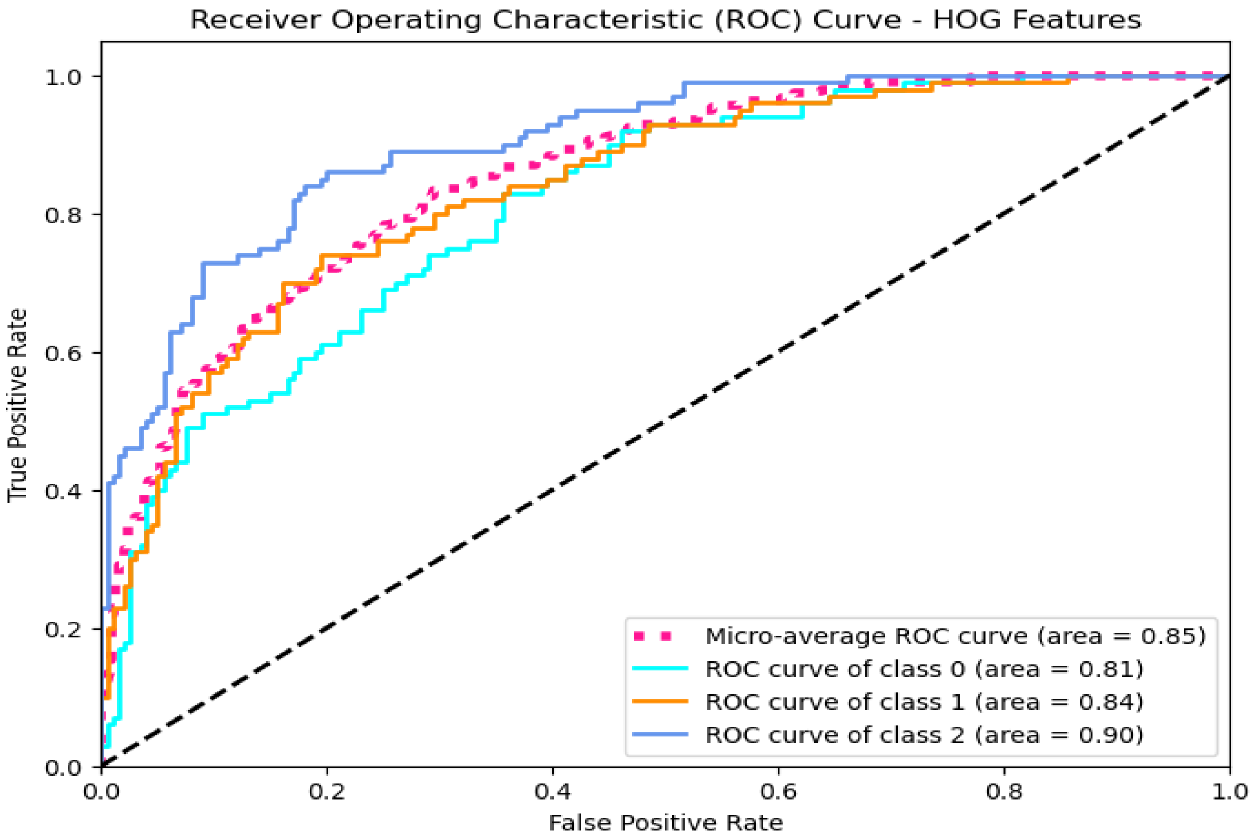
  - **Feature Names: HoG**
  - **Dimension of Resulted Features: (800,)**
- **Cross-Validation:**
  - **Used?:** [Yes/[✓ **No**]]
  - **Number of Folds:** [Number]
  - **Training/Validation Ratio:** [Ratio]
  -

- **Hyperparameters Used:**
- **solver='lbfgs': The optimization algorithm to use. The 'lbfgs' solver is used for small datasets and is the default choice for Logistic Regression in scikit-learn.**
- **penalty='l2': The norm used in the penalization. 'l2' refers to the L2 regularization.**
- **C=1.0: Inverse of regularization strength; smaller values specify stronger regularization.**
- **max_iter=5000: Maximum number of iterations taken for the solver to converge.**

Results Details

- **Performance on Testing Data:**
- **Accuracy of Logistic Regression on HOG Features: 68%**





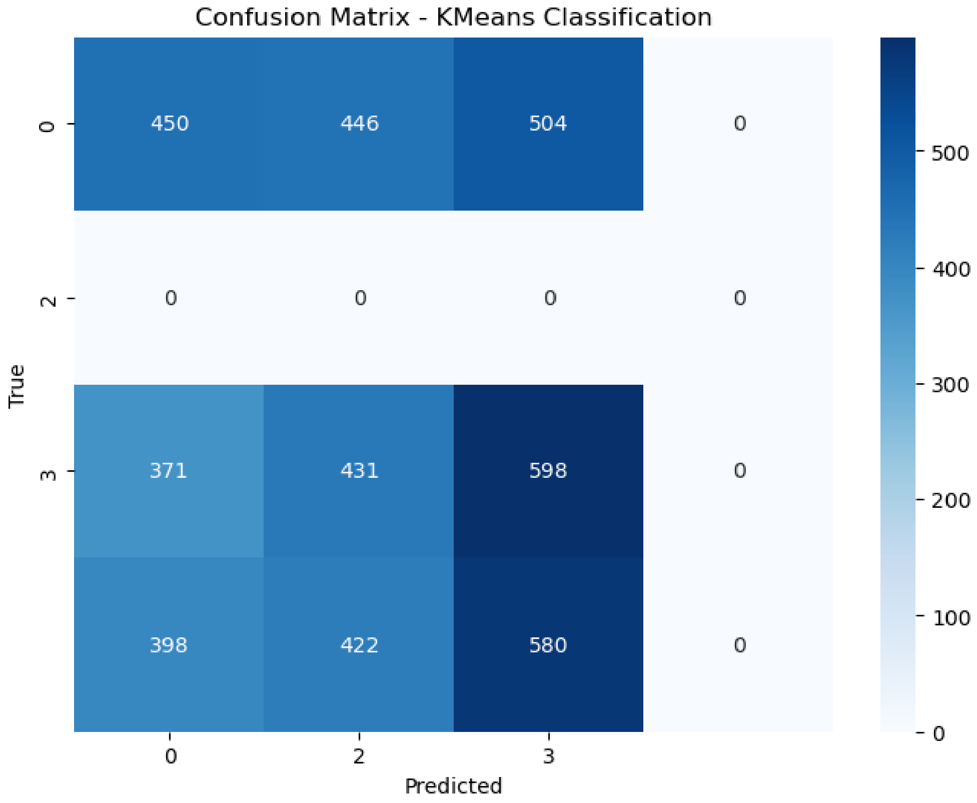
KMEANS MODEL

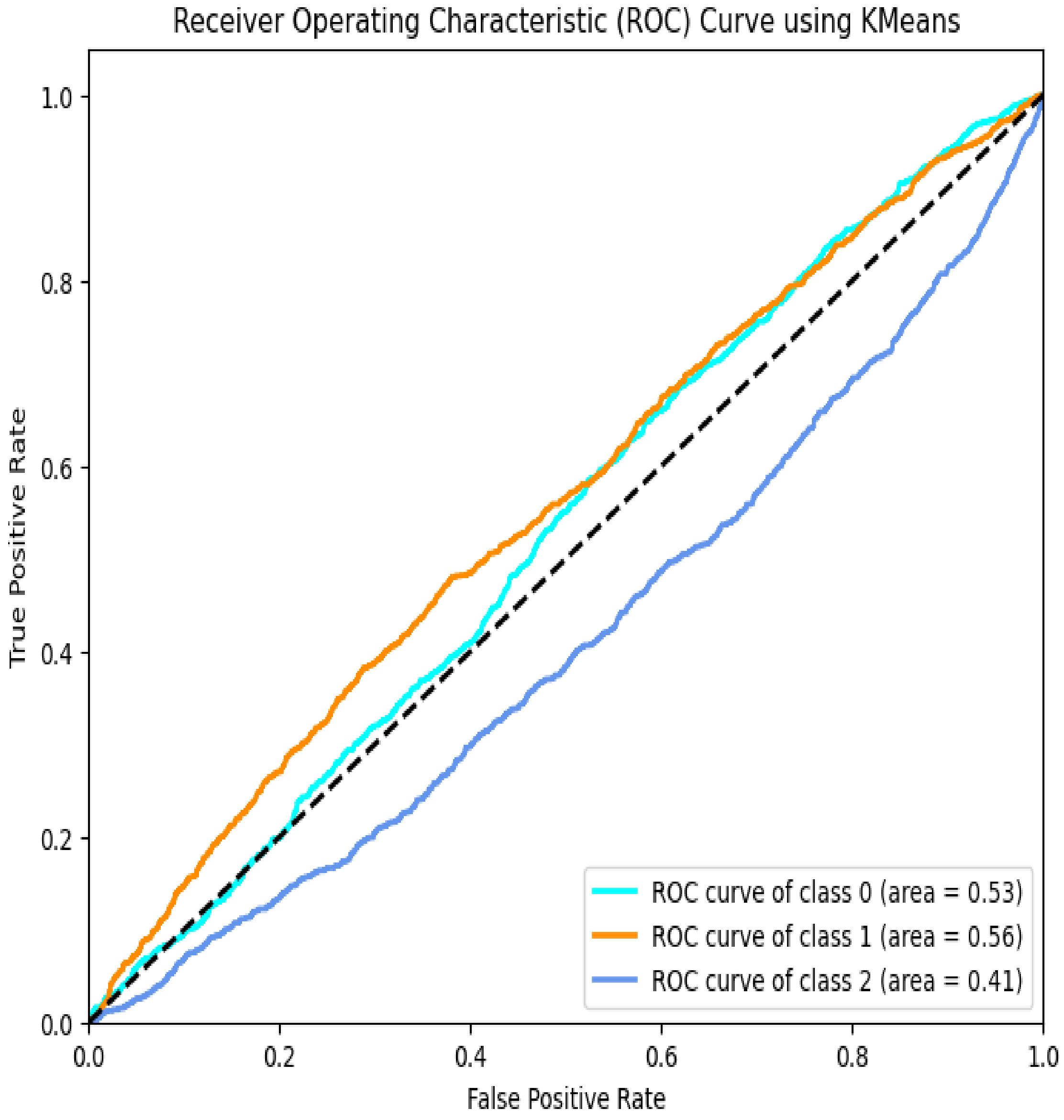
General Information on Dataset

- **Hyperparameters Used: k=3**

Results Details

- **Performance on Testing Data: 25%**




Receiver Operating Characteristic (ROC) Curve using KMeans
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.53)
ROC curve of class 1 (area = 0.56)
ROC curve of class 2 (area = 0.41)