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']**

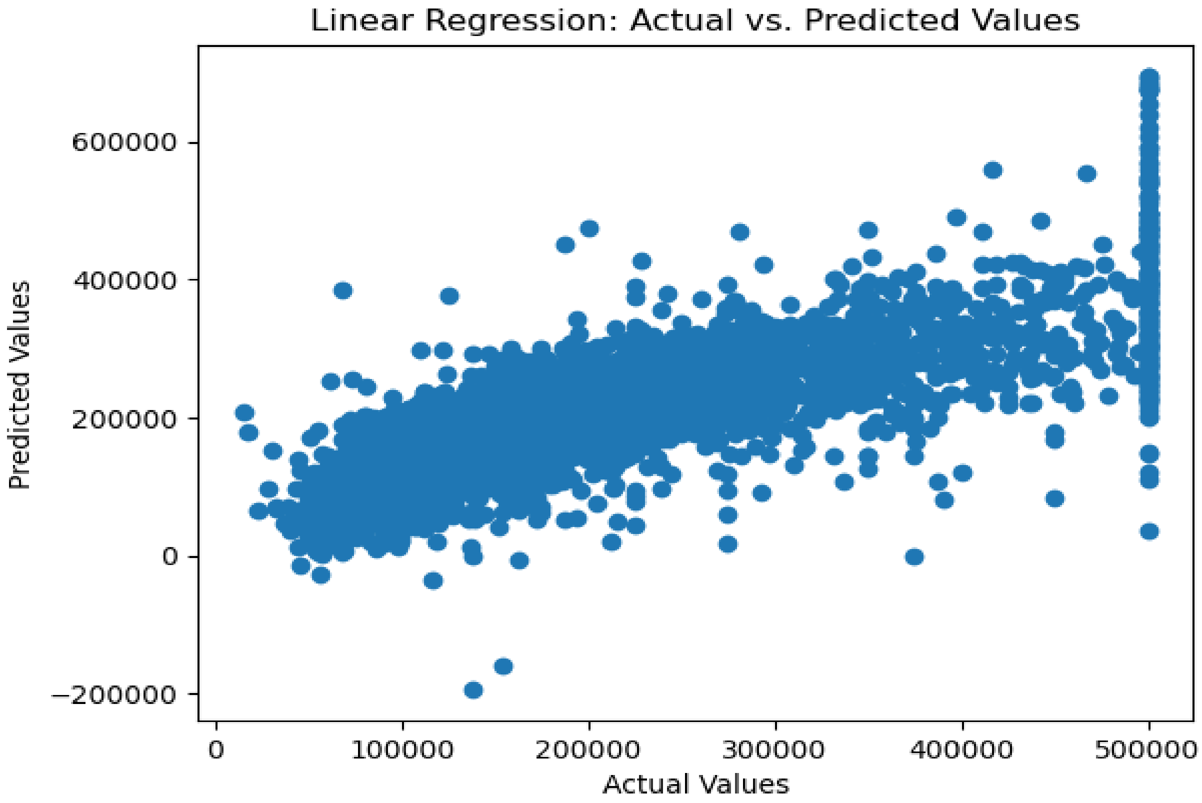
    -

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

- **Cross-Validation:**
    - **Used?:** [Yes/**No**]
    - **Number of Folds:** [Number]
    - **Training/Validation Ratio:** [Ratio]
- **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**

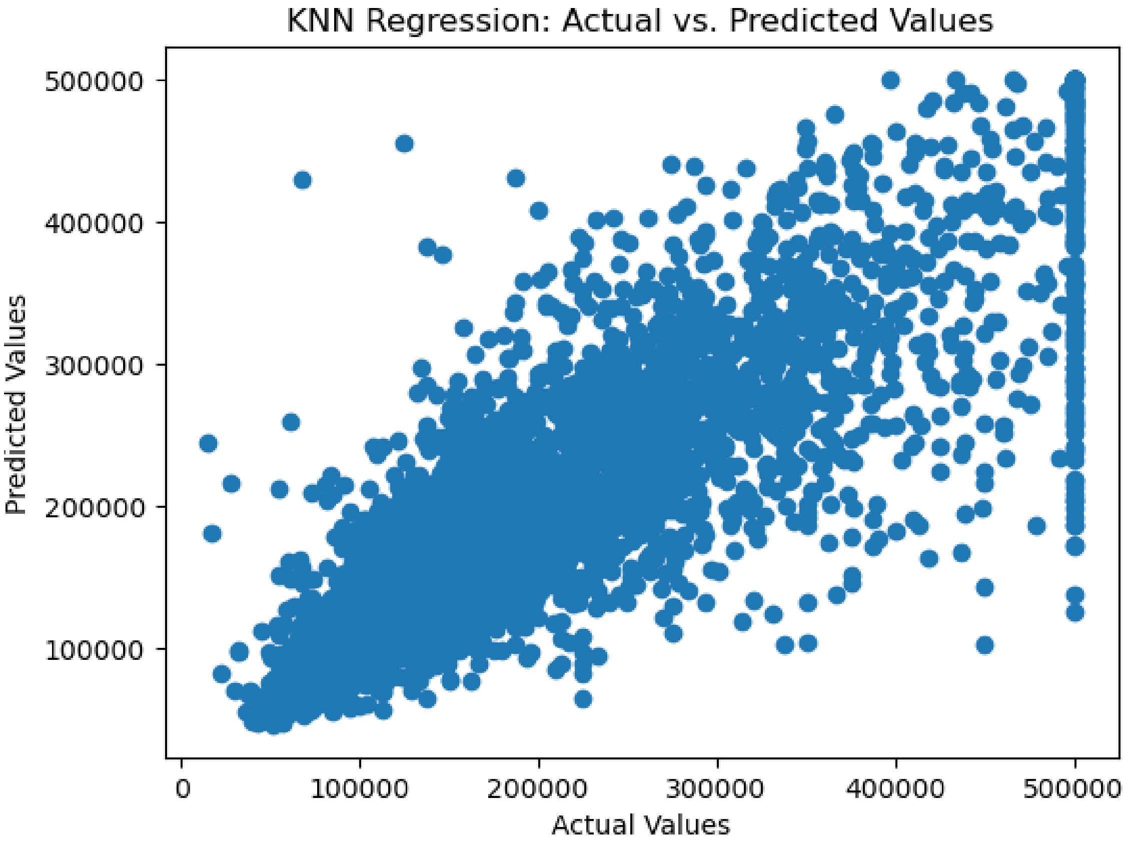


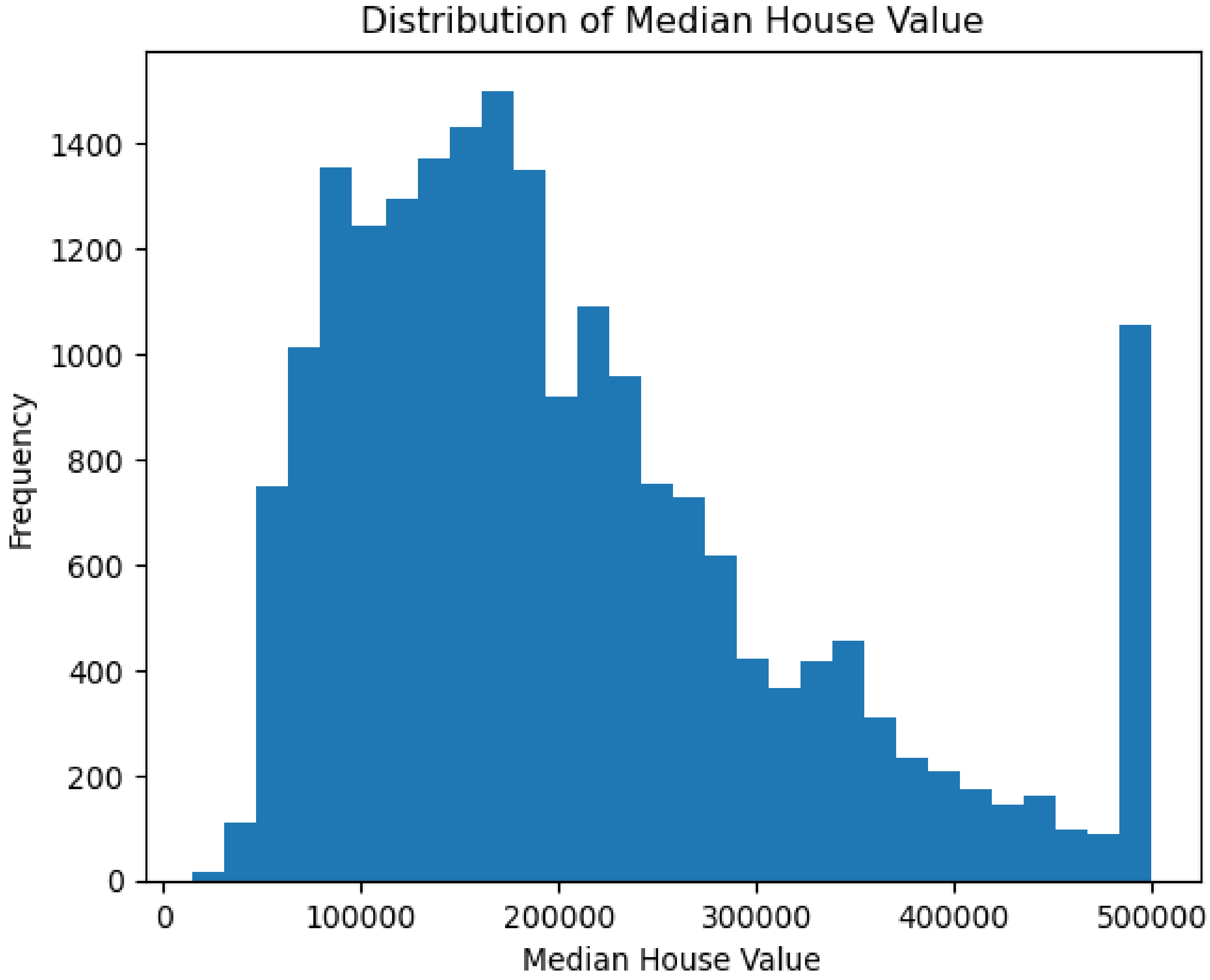
KNN

- ## 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**




Distribution of Median House Value
Frequency
0
200
400
600
800
1000
1200
1400
0
100000
200000
300000
400000
500000
Median House Value

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:**

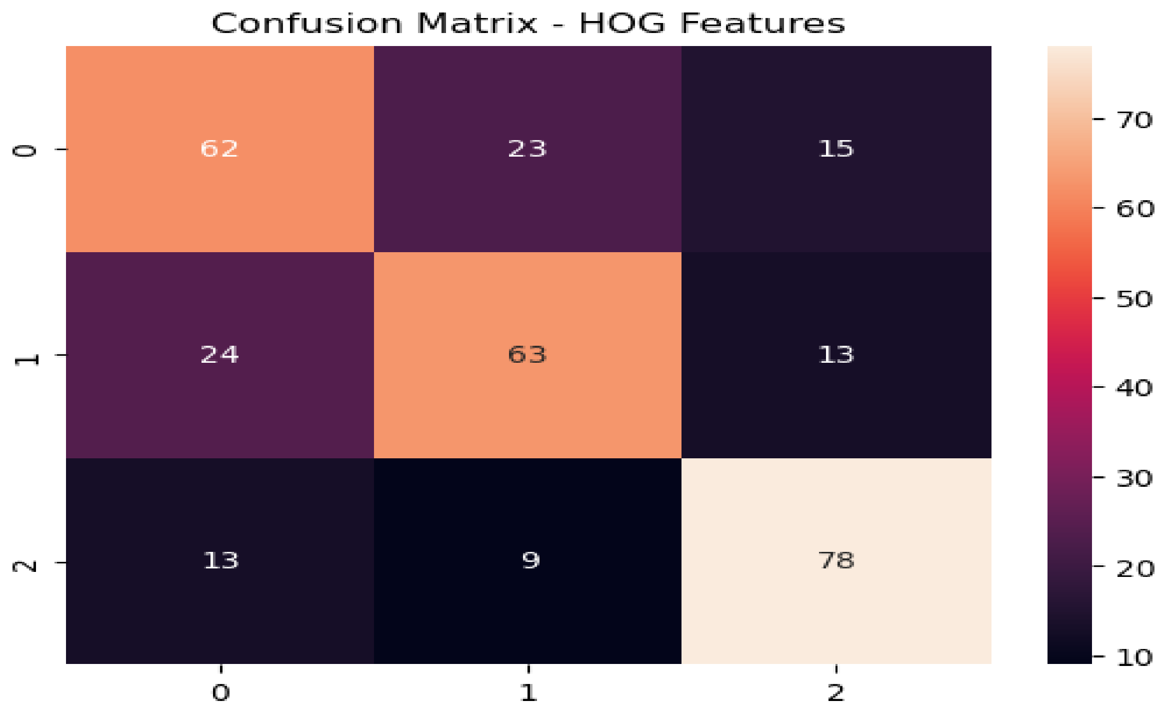
Number of features extracted per image: 800

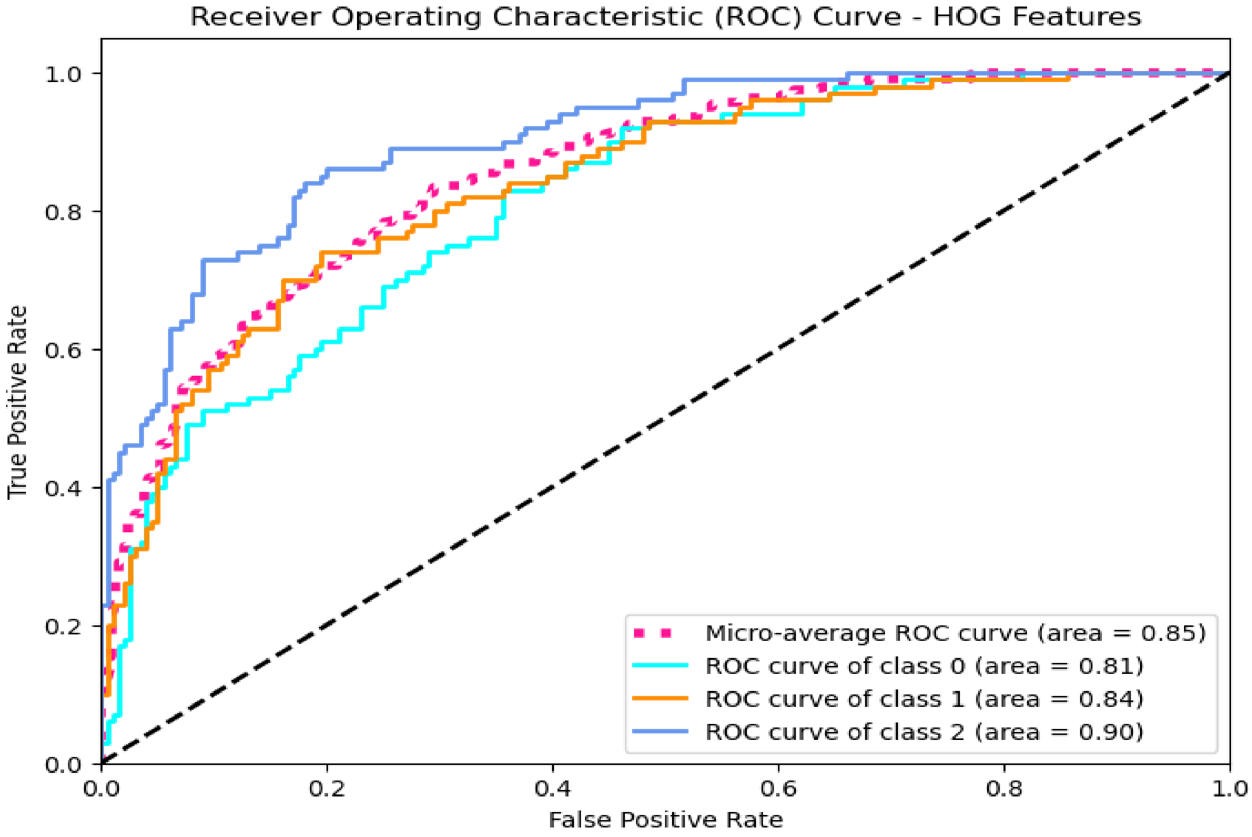
  - **Feature Names: HoG**
  - **Dimension of Resulted Features: (800,)**
- **Cross-Validation:**
  - **Used?:** [Yes/**No**]
  - **Number of Folds:** [Number]
  - **Training/Validation Ratio:** [Ratio]

- **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%**


Confusion Matrix - HOG Features


Receiver Operating Characteristic (ROC) Curve - HOG Features

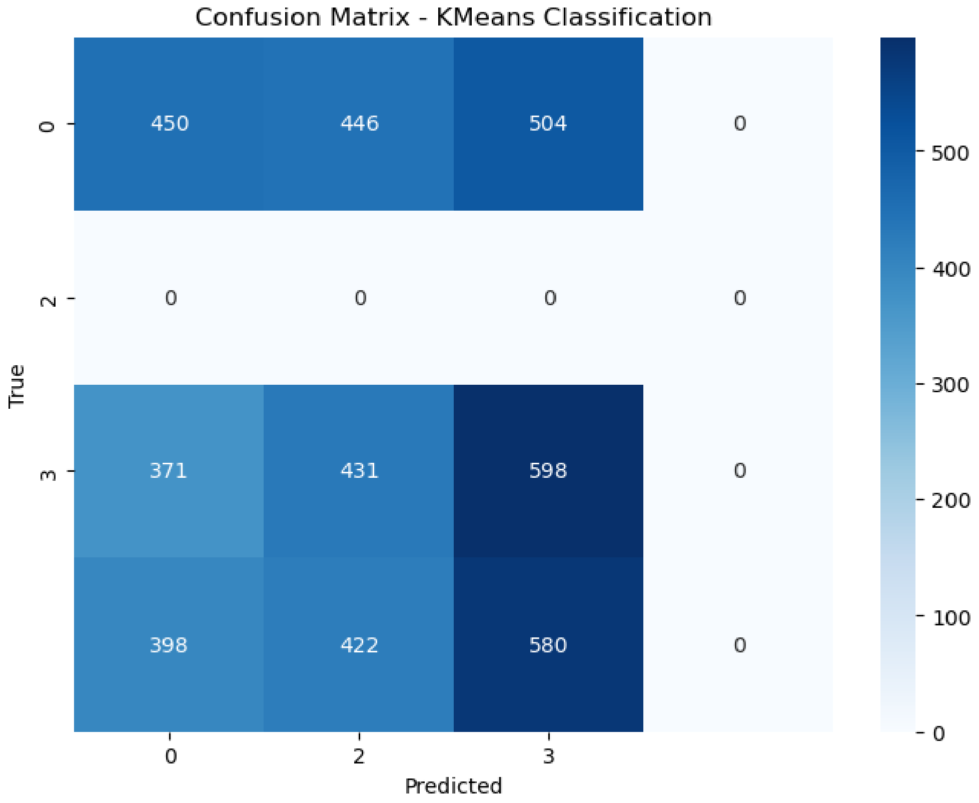
KMEANS MODEL

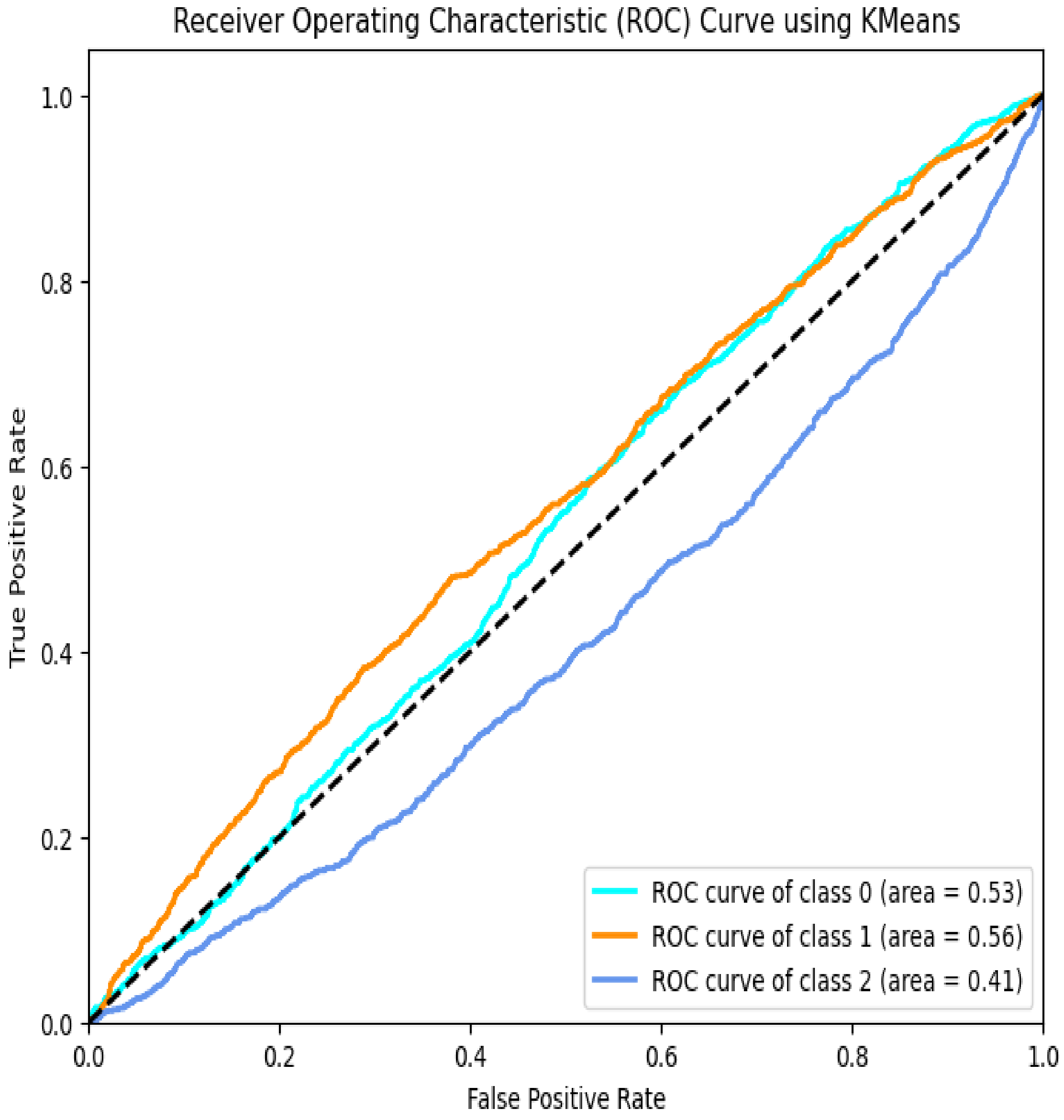
General Information on Dataset

- **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)