**Project Documentation: Regression and Classification Models**

**General Information**

# Numerical Dataset: California Housing Dataset

* **Dataset Name:** California Housing
* **Dataset Link:** https://www.kaggle.com/datasets/camnugent/california-housing-prices/data
* **Features:** 9 features (e.g., median income, total rooms, etc.)
* **Target Variable:** Median house value
* **Missing Data:** Total bedrooms feature had missing values, filled with the mean.
* **Total Samples:** 20640
* **Training/Testing Split:**

o **Training Samples:** 16512 o **Testing Samples:** 4128

# Image Dataset: Flower Species Recognition

* **Dataset Name:** Oxford 102 Flower Dataset
* **Dataset Link:** https://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html
* **Classes:** 5 (subset of the dataset) o Class Labels: [51, 77 , 46 , 73 , 89] I used the lables that have the most amount of images in it
* **Total Samples:**
  + **Images per Class:** o 51 -> 258 images o 77 -> 251 images o 46 -> 196 images o 73 -> 194 images o 89 -> 184 images
  + **Image Size:** 16x16 (after resizing) in knn and 128x128 in logestic
* **Training/Testing Split:**
  + Training Samples: 866
  + Testing Samples: 217

**Implementation Details**

# Regression Models on Numerical Dataset

1. **Linear Regression** 
   1. **Metrics on Testing Data:**
      1. Mean Squared Error (MSE): 5055025116.165614
      2. R² Score: 0.6142406531011786
      3. Mean Absolute Error (MAE): 51846.87784903816
2. **K-Nearest Neighbors Regressor (KNN)** 
   1. **Metrics on Testing Data:**
      1. Mean Squared Error (MSE): 3773182808.9917927ii. R² Score: 0.7120606717715767

iii. Mean Absolute Error (MAE 40879.577277131786

**Comparison Table:**

**Metric** **Linear Regression** **KNN Regressor**

Mean Squared Error 5055025116.165614 3773182808.9917927

R² Score 0.6142406531011786 0.7120606717715767

Mean Absolute Error 51846.87784903816 40879.577277131786

# Classification Models on Image Dataset

1. **Logistic Regression**

a. **Metrics on Testing Data:**

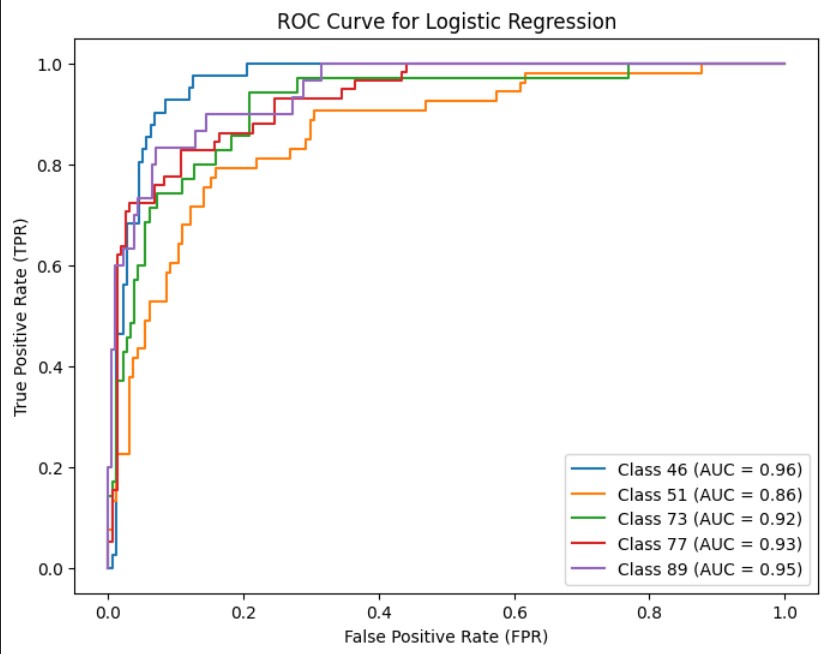
* + 1. Accuracy: 0.7419
    2. Precision: 0.7497
    3. Recall: 0.7419 iv. Loss : 1.4224

v. Confusion Matrix:

A diagram of a confusion matrix

Description automatically generated

b. **ROC Curve:**

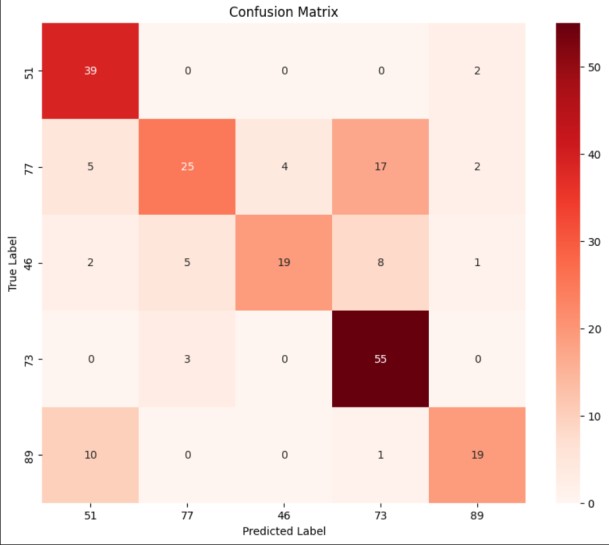


1. **K-Nearest Neighbors Classifier (KNN)**

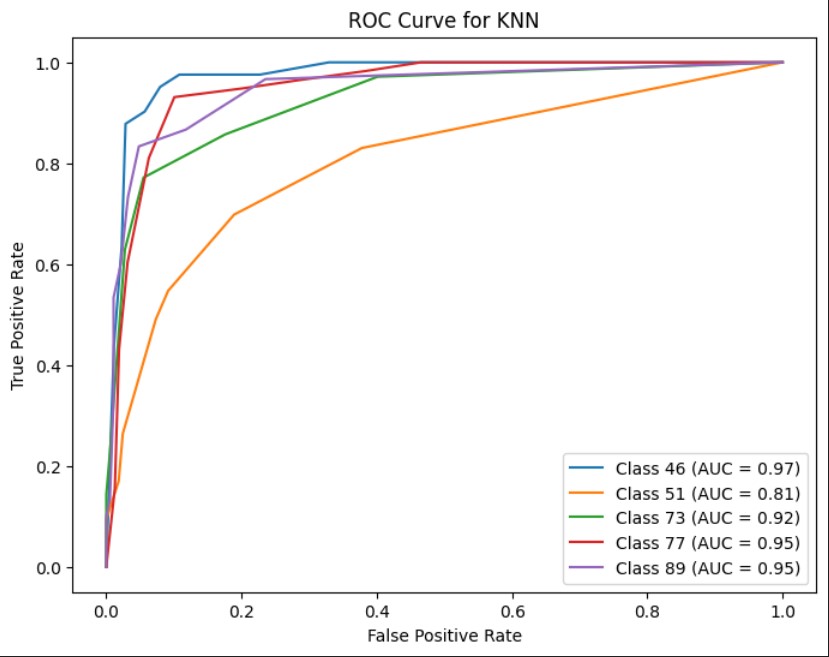
a. **Metrics on Testing Data:**

* + 1. Accuracy: 0.7235
    2. Precision: 0.7408
    3. Recall: 0.7235 iv. Loss : 2.3965

v. Confusion Matrix:



b. **ROC Curve:**



**Comparison Table:**

**Metric** **Logistic Regression**  **KNN Classifier**

Accuracy 0.7419 0.7235

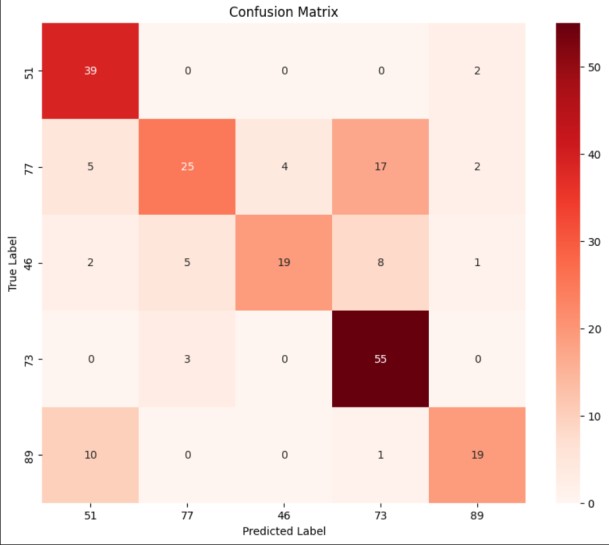
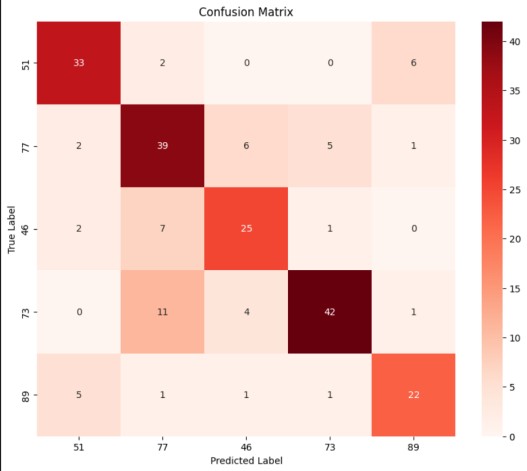
Precision 0.7497 0.7408

Recall 0.7419 0.7235

Loss 1.4224 2.3965

Average AUC 0.9216 0.9150

Confusion matrix



ROC Curve and AUC Values

