Project Documentation: Regression and Classification Models

General Information

Numerical Dataset: California Housing Dataset

- **Dataset Name:** California Housing
- **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:**
  - **Training Samples:** 16512
  - **Testing Samples:** 4128

Image Dataset: Flower Species Recognition

- **Dataset Name:** Oxford 102 Flower Dataset
- **Classes:** 5 (subset of the dataset)
  - 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:**
  - 51 -> 258 images
  - 77 -> 251 images
  - 46 -> 196 images
  - 73 -> 194 images
  - 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**
   a. **Metrics on Testing Data:**
      i. Mean Squared Error (MSE): 5055025116.165614

ii. R² Score: 0.6142406531011786

iii. Mean Absolute Error (MAE): 51846.87784903816

2. **K-Nearest Neighbors Regressor (KNN)**

   a. **Metrics on Testing Data:**

      i. Mean Squared Error (MSE): 3773182808.9917927

      ii. 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:**

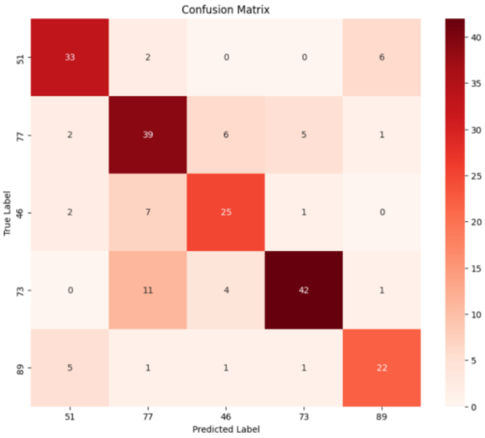
      i. Accuracy: 0.7419

      ii. Precision: 0.7497

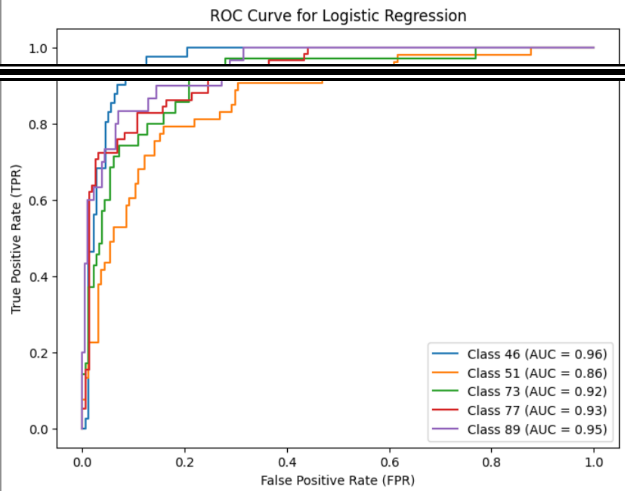
      iii. Recall: 0.7419

      iv. Loss : 1.4224

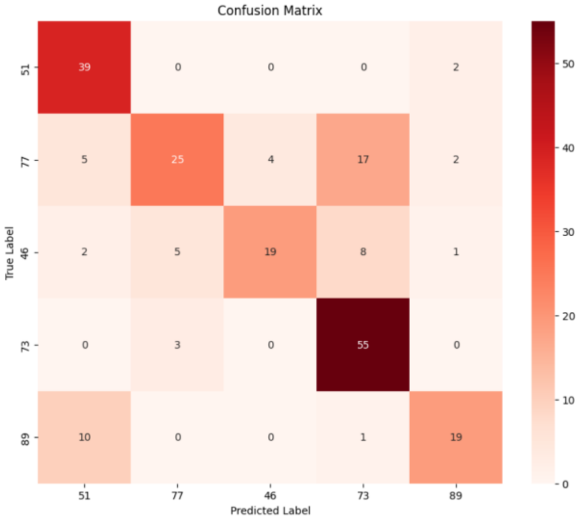
      v. Confusion Matrix:



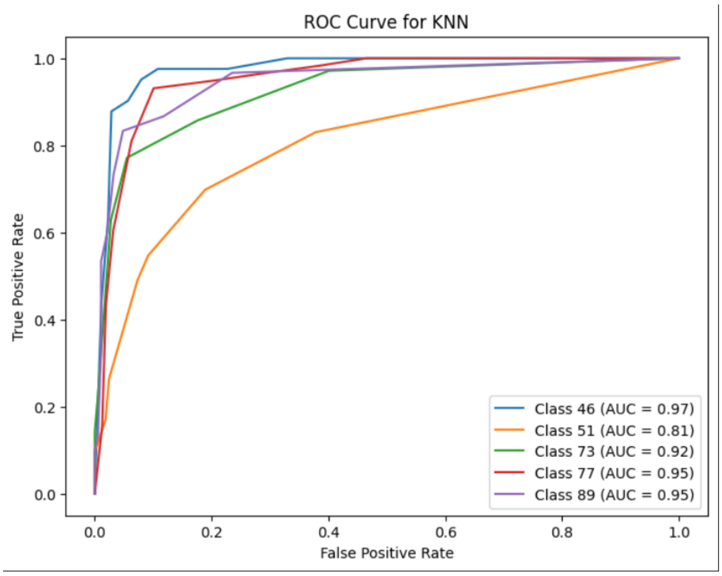
   b. **ROC Curve:**



2. **K-Nearest Neighbors Classifier (KNN)**
   a. **Metrics on Testing Data:**
      i. Accuracy: 0.7235
      ii. Precision: 0.7408
      iii. Recall: 0.7235
      iv. Loss : 2.3965
      v. Confusion Matrix:



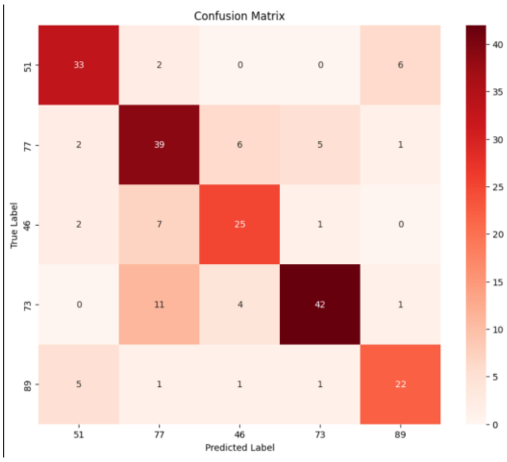
   b. **ROC Curve:**



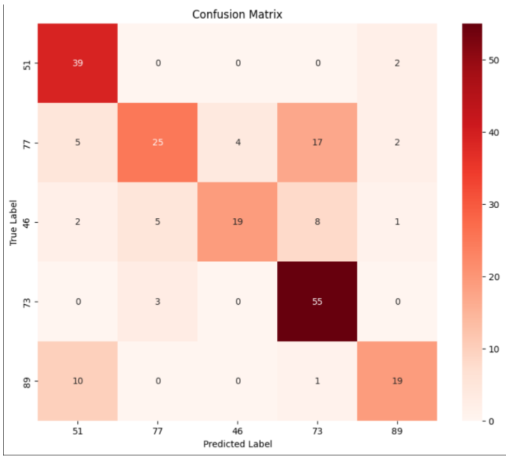
Comparison Table:

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

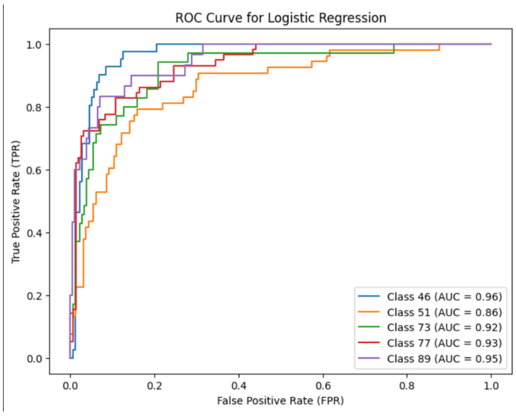


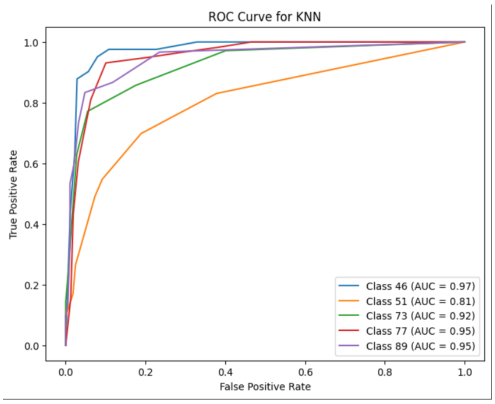
True Label \ Predicted Label	51	77	46	73	89
51	33	2	0	0	6
77	2	39	6	5	1
46	2	7	25	1	0
73	0	11	4	42	1
89	5	1	1	1	22



True Label \ Predicted Label	51	77	46	73	89
51	39	0	0	0	2
77	5	25	4	17	2
46	2	5	19	8	1
73	0	3	0	55	0
89	10	0	0	1	19

ROC Curve and AUC Values


ROC Curve for Logistic Regression
True Positive Rate (TPR)
False Positive Rate (FPR)
Class 46 (AUC = 0.96)
Class 51 (AUC = 0.86)
Class 73 (AUC = 0.92)
Class 77 (AUC = 0.93)
Class 89 (AUC = 0.95)


ROC Curve for KNN
True Positive Rate
False Positive Rate
Class 46 (AUC = 0.97)
Class 51 (AUC = 0.81)
Class 73 (AUC = 0.92)
Class 77 (AUC = 0.95)
Class 89 (AUC = 0.95)