Documentation for numeric dataset:

General information.

- Dataset Name: Melbourne Housing Prices Dataset

Dataset Size: 13,580 rows

Missing Values: There Was 13,256 Which we replaced with mean (numeric columns) and mode (categorical values)

Missing Values:

Columns	Number of Missing Values	Handle missing values
Car	62	mean
Building Area	6450	mean
Year Built	5375	mean
Council Area	1369	mode

1. Test Set (20%): 0.2×13,580=2,716

2. Training and Validation Set (Remaining 80%): 80%×13,580=10,864

Algorithms we used:

1. Linear Regression
2. KNN (K nearest Neighbor)

	Linear Regression	KNN
Mean squared error	88633106697.58	54964458955.03
R squared	68%	80%
Mean Absolute Error	227297.18	168937.94

Documentation for image dataset:

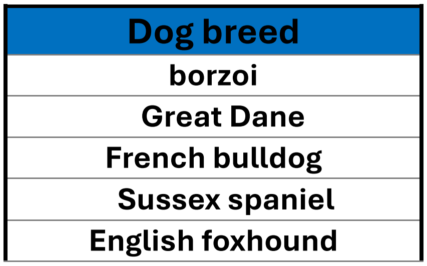
General information:

- Dataset Name: Stanford Dogs

Dataset Size: 1548 rows

Numbers of classes : 5 classes

Labels :



Dog breed
borzoi
Great Dane
French bulldog
Sussex spaniel
English foxhound

1. Test Set (30%): 0.3× 1548 =465

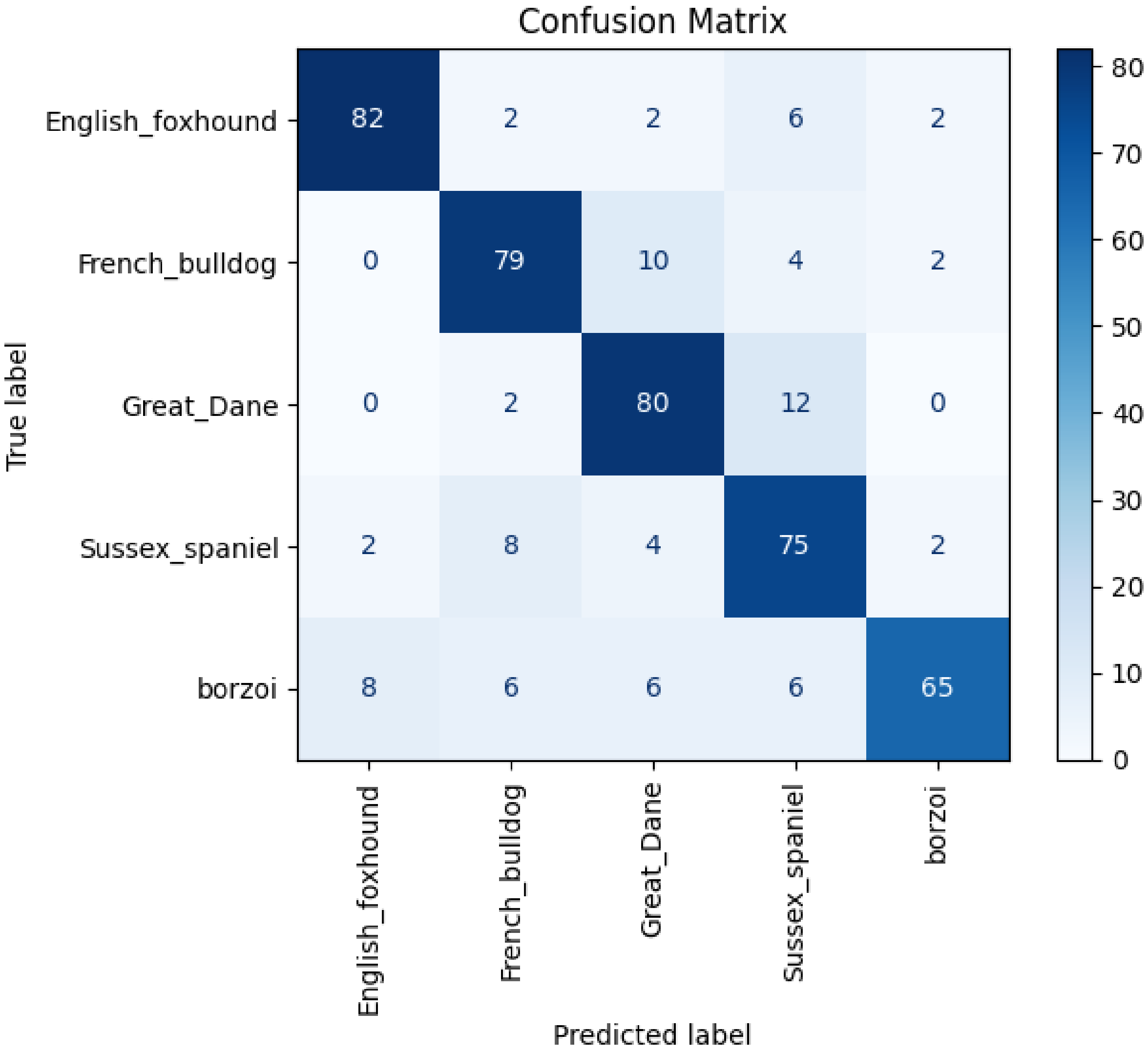
2. Training and Validation Set (Remaining 70%): 70%×1548=1038

Algorithms we used:

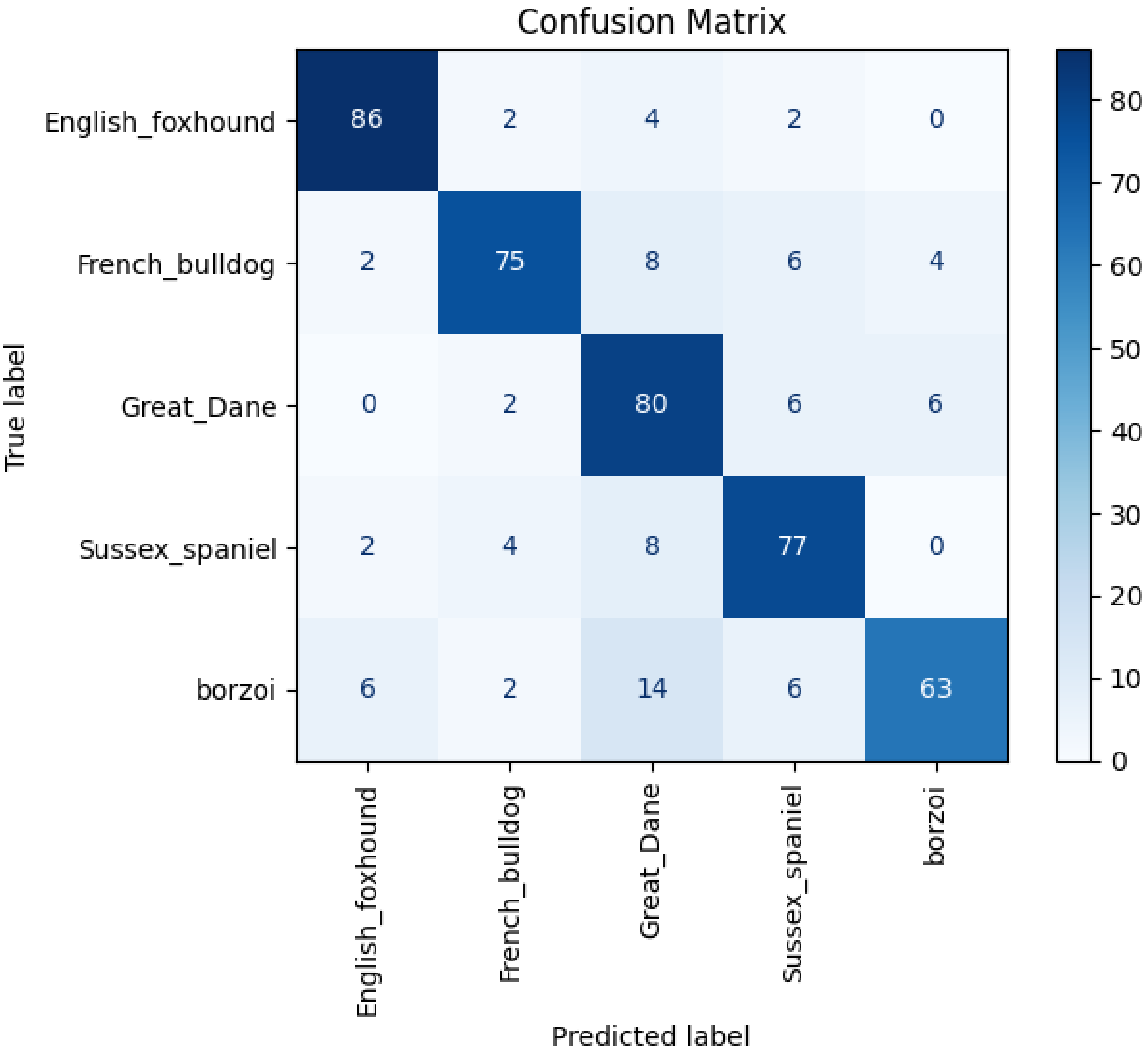
Logistic Regression	Classes	precision	recall	f1-score
	0	0.89	0.87	0.88
	1	0.81	0.83	0.82
	2	0.78	0.85	0.82
	3	0.73	0.82	0.77
	4	0.92	0.71	0.80

KNN	Classes	precision	recall	f1-score
	0	0.90	0.91	0.91
	1	0.88	0.79	0.83
	2	0.70	0.85	0.77
	3	0.79	0.85	0.82
	4	0.86	0.69	0.77

Logistic Confusion Matrix:

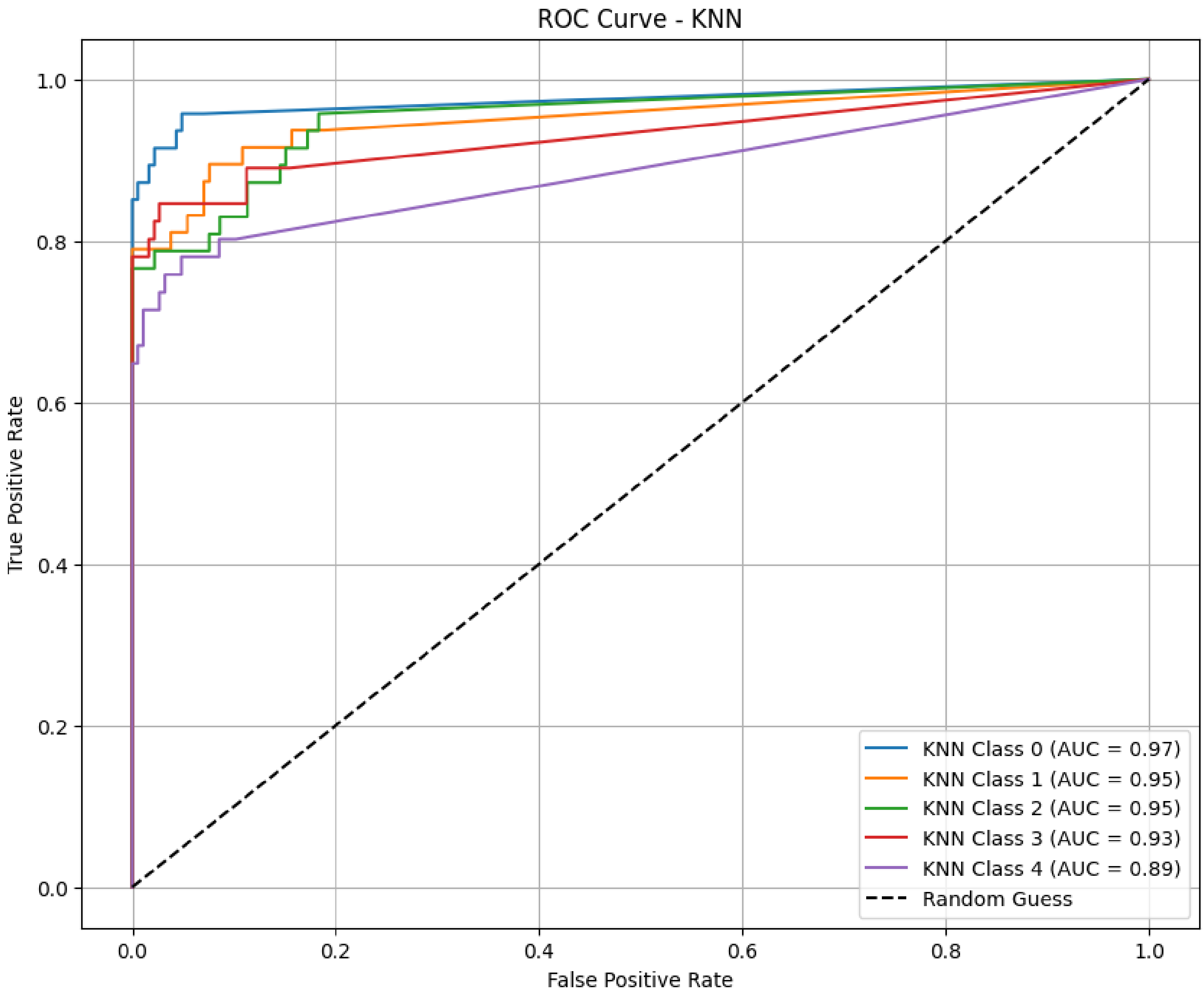

Confusion Matrix
True label
English_foxhound
French_bulldog
Great_Dane
Sussex_spaniel
borzoi
82
2
2
6
2
0
79
10
4
2
0
2
80
12
0
2
8
4
75
2
8
6
6
6
65
Predicted label
English_foxhound
French_bulldog
Great_Dane
Sussex_spaniel
borzoi
80
70
60
50
40
30
20
10
0

KNN Confusion Matrix:



True label \ Predicted label	English_foxhound	French_bulldog	Great_Dane	Sussex_spaniel	borzoi
English_foxhound	86	2	4	2	0
French_bulldog	2	75	8	6	4
Great_Dane	0	2	80	6	6
Sussex_spaniel	2	4	8	77	0
borzoi	6	2	14	6	63

ROC_AUC_ Curve for KNN



Logistic Regression and KNN for Image Dataset:

We worked with data from different dog breeds, selecting 5 classes to focus on. The images were processed by cropping them to 350x250 pixels and converting them to grayscale. We then normalized the data by dividing by 255, resulting in values between [0,1]. The dataset was split into 70% for training and 30% for testing. Through grid search, we tested various values for k, weights, and metrics. We chose k = 7, weight = distance, and the metric as Manhattan. For logistic regression, we set 1000 iterations to achieve solid predictions. Both KNN and logistic regression achieved an accuracy close to 82%.

Linear Regression and KNN for Numeric Dataset:

We utilized the Melbourne housing dataset, where we had features and pricing for properties. For linear regression, we used a straightforward model and split the data into 80% for training and 20% for testing. For KNN, we employed cross-validation to find the best metrics. We settled on k = 9 and the Manhattan distance metric. This approach allowed us to select the most accurate model for predicting house values.