ML PROJECT DOCUMENTATION

Datasets used:

1st Numerical Dataset: House prices advanced regression techniques

2nd Numerical Dataset: California House price

Image Dataset: STL-10

House prices advanced regression techniques

- **Number of samples (rows)**: 1,460
- **Number of features (columns)**: 81 (including both numeric and categorical features)

SalePrice: The target variable you're trying to predict, representing the saleprice of the house.

Some columns contain missing data

1,460 samples in the training set:

- **80% for training**:1168 samples for training
- **20% for testing**:292 samples for testing

ALGORITHMS USED:

1. KNN
2. Linear Regression

EVALUATION METRICS FOR BOTH ALGORITHMS

Linear Regression	KNN
Mean Squared Error: 2964512852.810488 Root Mean Squared Error: 54447.340181229134	Mean Squared Error: 2975441399.479452 Root Mean Squared Error: 54547.60672549669
Mean absolute Error: 24441.136345746843	Mean absolute Error: 34373.29452054795
R-squared score: 0.5707244049229452	R-squared score: 0.5691418992609443

California Housing Prices

- **Number of samples (rows)**: **20,640**
- **Number of features (columns)**: **8** (including both numeric and categorical features)

Median House value: The target variable you're trying to predict.

20,640 samples in the training set:

- **80% for training**:

16,512 samples for training

- **20% for testing**:

4,128 samples for testing

ALGORITHMS USED:

1. KNN
2. Linear Regression

EVALUATION METRICS FOR BOTH ALGORITHMS

Linear Regression	KNN

`Mean Squared Error: 4718206968.301578`
`Root Mean Squared Error: 68689.20561705148`
	`Mean Squared Error: 6134568846.257751`
	`Root Mean Squared Error: 78323.4884709418`
	In [37]:

`Mean absolute Error: 49697.07016481124`	`Mean absolute Error: 50802.798449612405`
`R-squared score: 0.6381617983930403`	`R-squared score: 0.5295413334182277`

STL-10

The **STL-10** dataset is a collection of images used for machine learning and computer vision tasks. It consists of **10 classes**, which are the categories of objects present in the dataset.

1. **Label 0: Airplane**
2. **Label 1: Automobile**
3. **Label 2: Bird**
4. **Label 3: Cat**
5. **Label 4: Deer**
6. **Label 5: Dog**
7. **Label 6: Frog**
8. **Label 7: Horse**
9. **Label 8: Ship**
10. **Label 9: Truck**

- **Training Set**:
  - The training set consists of **5,000 labeled images** (from 10 classes).
  - These are split into **500 labeled images per class**.
- **Testing Set**:
  - The test set consists of **8,000 labeled images**.
  - These are used for evaluation, with **800 images per class**.

The images in the **STL-10** dataset are of size **96x96 pixels**.

ALGORITHMS USED:

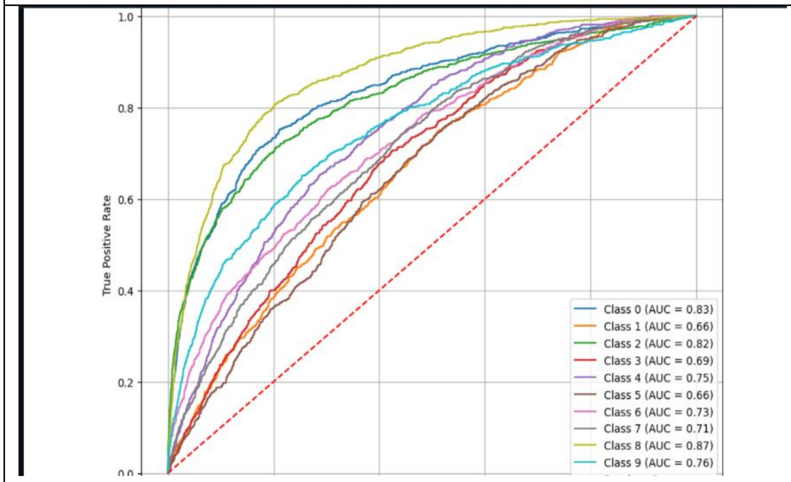
1. KNN

2. Logistic Regression

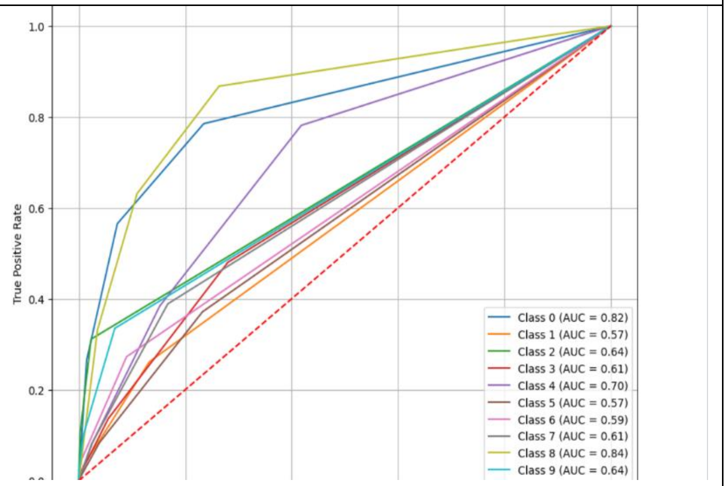
Logistic Regression	KNN
Accuracy: 0.3215	Accuracy: 0.2715
Loss value: 3.3704672	Loss value: 8.647071
Roc_auc: 0.747326840277778	Roc_auc:0.6596844444444444

Evaluation Metrics

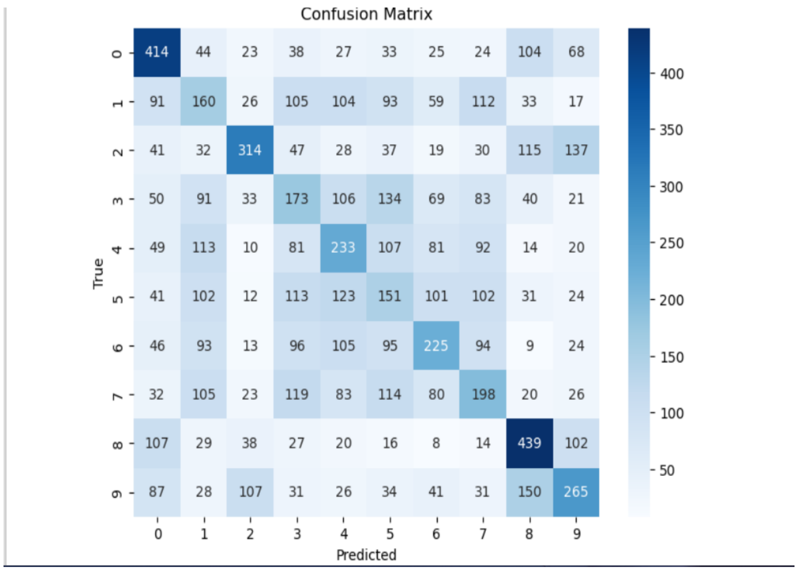
ROC CURVE LOGISTIC



ROC CURVE KNN

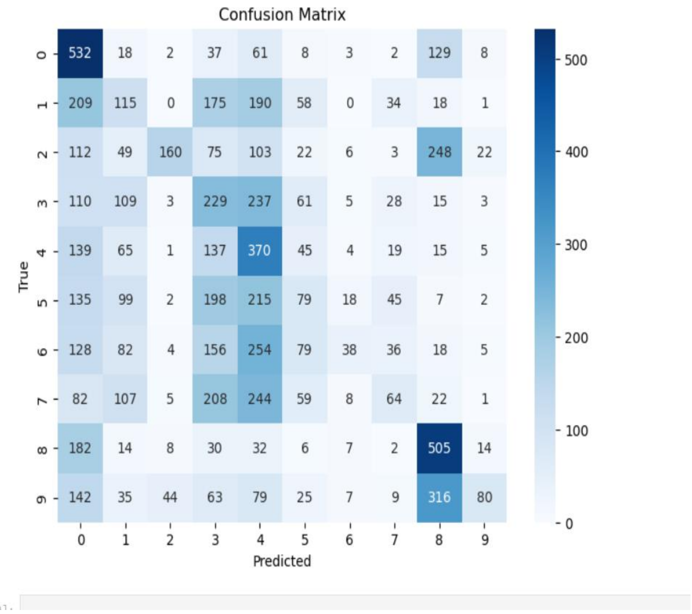


CONFUSSION MATRIX LOGISTIC


Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	414	44	23	38	27	33	25	24	104	68
1	91	160	26	105	104	93	59	112	33	17
2	41	32	314	47	28	37	19	30	115	137
3	50	91	33	173	106	134	69	83	40	21
4	49	113	10	81	233	107	81	92	14	20
5	41	102	12	113	123	151	101	102	31	24
6	46	93	13	96	105	95	225	94	9	24
7	32	105	23	119	83	114	80	198	20	26
8	107	29	38	27	20	16	8	14	439	102
9	87	28	107	31	26	34	41	31	150	265

CONFUSSION MATRIX KNN


Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	532	18	2	37	61	8	3	2	129	8
1	209	115	0	175	190	58	0	34	18	1
2	112	49	160	75	103	22	6	3	248	22
3	110	109	3	229	237	61	5	28	15	3
4	139	65	1	137	370	45	4	19	15	5
5	135	99	2	198	215	79	18	45	7	2
6	128	82	4	156	254	79	38	36	18	5
7	82	107	5	208	244	59	8	64	22	1
8	182	14	8	30	32	6	7	2	505	14
9	142	35	44	63	79	25	7	9	316	80