# 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<br>Root Mean Squared Error: 54447.340181229134 | Mean Squared Error: 2975441399.479452<br>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<br>Root Mean Squared Error: 68689.205617<br>05148 | Mean Squared Error: 6134568846.257751<br>Root Mean Squared Error: 78323.488470<br>9418<br>In [37]: |
| Mean absolute Error: 49697.0701648112<br>4 | Mean absolute Error: 50802.7984496124<br>05 |
| 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: Bird**
3. **Label 2: car**
4. **Label 3: Cat**
5. **Label 4: Deer**
6. **Label 5: Dog**
7. **Label 6: horse**
8. **Label 7: frog**
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**.

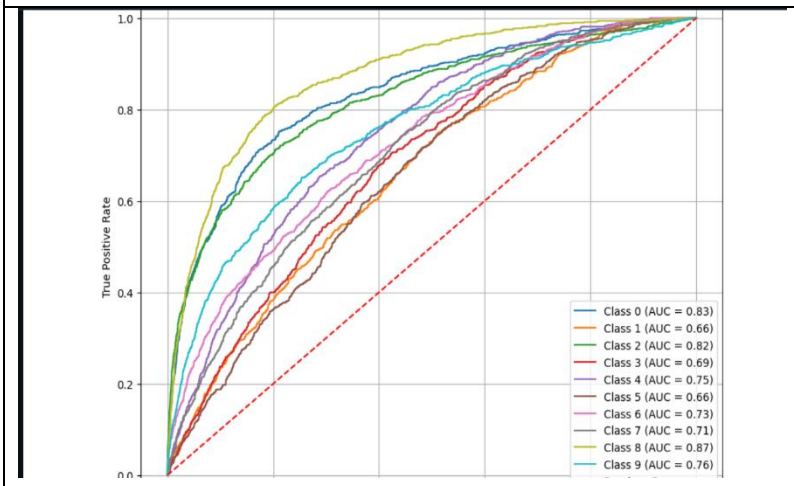
# ALGORITHMS USED:

1. KNN
2. Logistic Regression

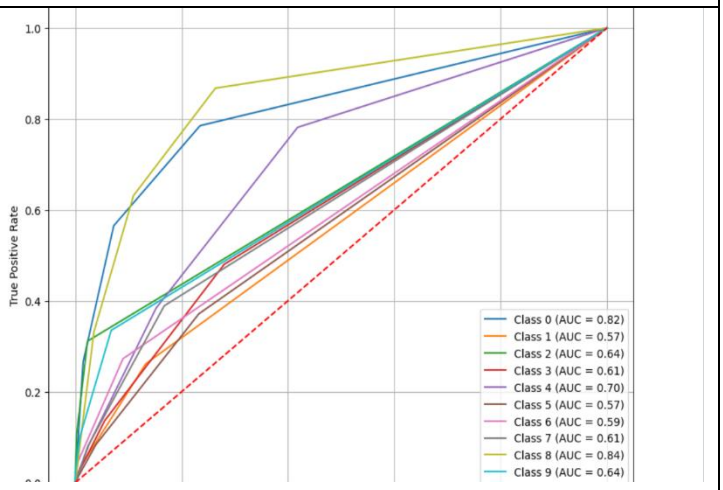
| Logistic Regression | KNN |
|---|---|
| Accuracy: 0.3215 | Accuracy: 0.2715 |
| Loss value: 3.3704672 | Loss value: 8.647071 |
| Roc_auc: 0.747326840277778 | Roc_auc:0.6596844444444444 |

# Evaluation Metrics

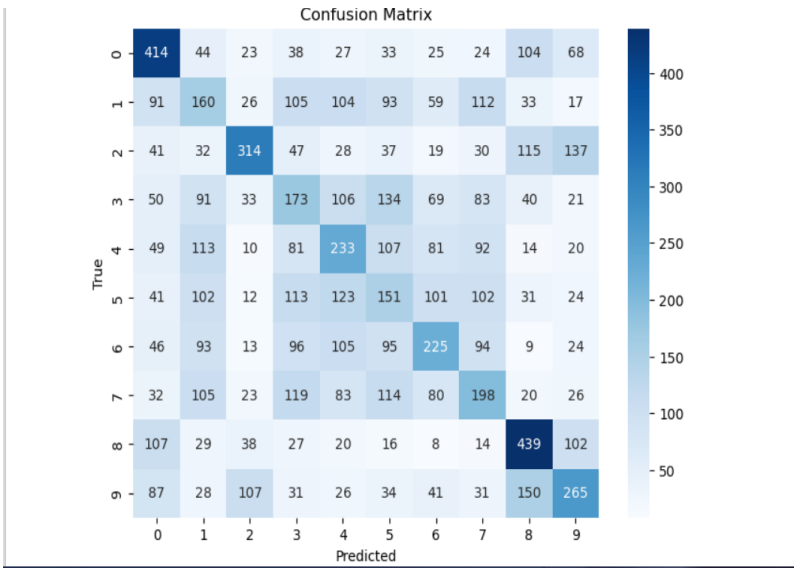
## ROC CURVE LOGISTIC




## ROC CURVE KNN




## CONFUSSION MATRIX LOGISTIC




## CONFUSSION MATRIX KNN