**Machine Learning Documentation:**

1. **Numerical Dataset:**

**2.1) General Information about the dataset:**

**Name of the used Dataset: House Pricing**

**Name of Target : price**

**Toal Number of used Samples: 4600**

**Number of samples after removing outliers: 4253**

**Number of samples used in training/validation: 3402**

**Number of samples used in testing: 851**

**2.2) Implementation details:**

**a. At feature extraction phase:**

**The features were extracted is: 12**

**The name of features: (bedrooms, sqft\_living,**

**sqft\_lot,floors,view,condition,sqft\_above,sqft\_basement,yr\_built,**

**yr\_renovated,city,country)**

**b. cross validation is used**

**Numbers of Folds: 10**

**Ratio of training / validation : 9/1**

**c. Hyperparmeters used in knn regressor:**

**n\_neighbors=k**

**weights = 'distance'**

**These hyperparameters influence the model’s performance and should be tuned**

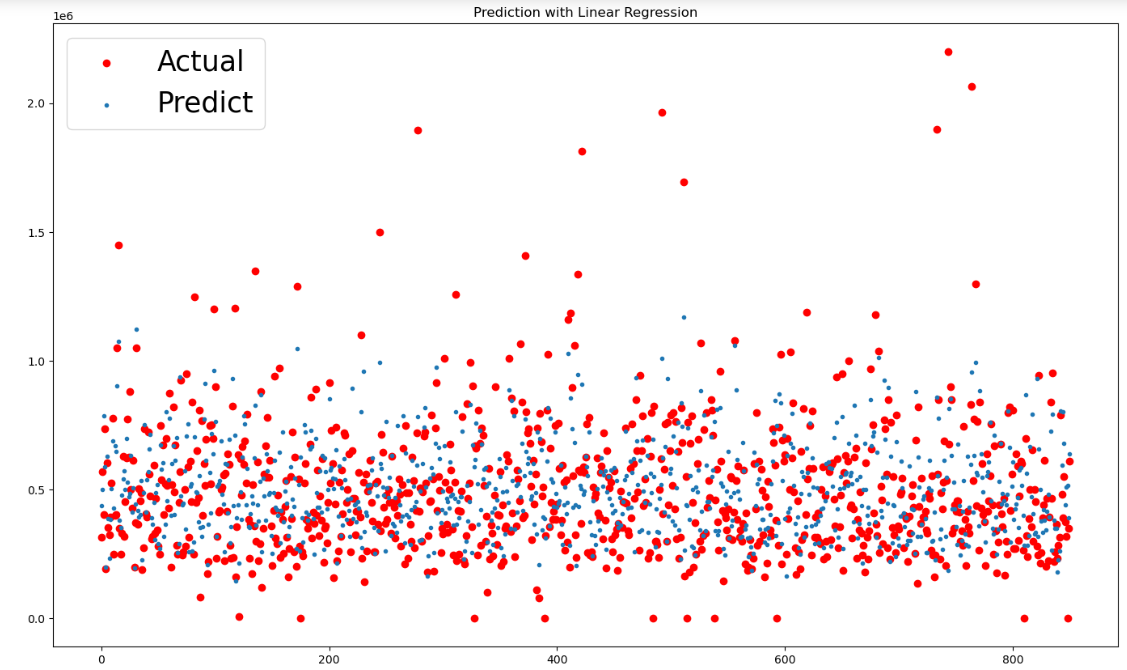
**based on the characteristics of the data.**

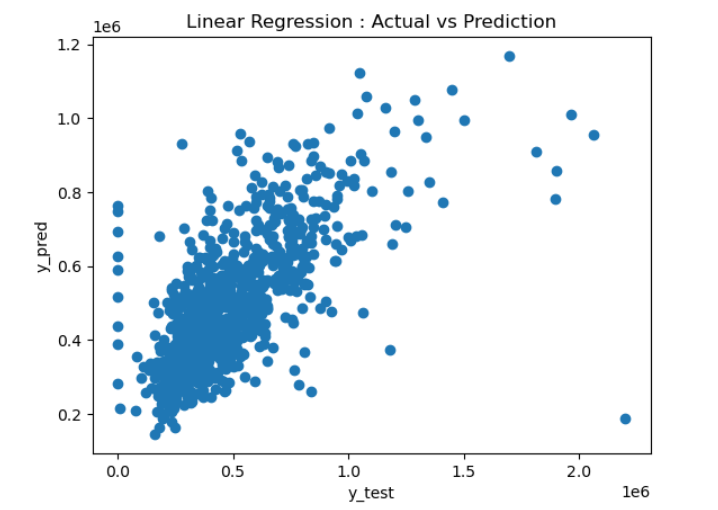
**2.3) Results Details:**

**Linear Regression:**

**The RMSE value is 39366602465.76078**

**The R square value is 0.44738168058928346**

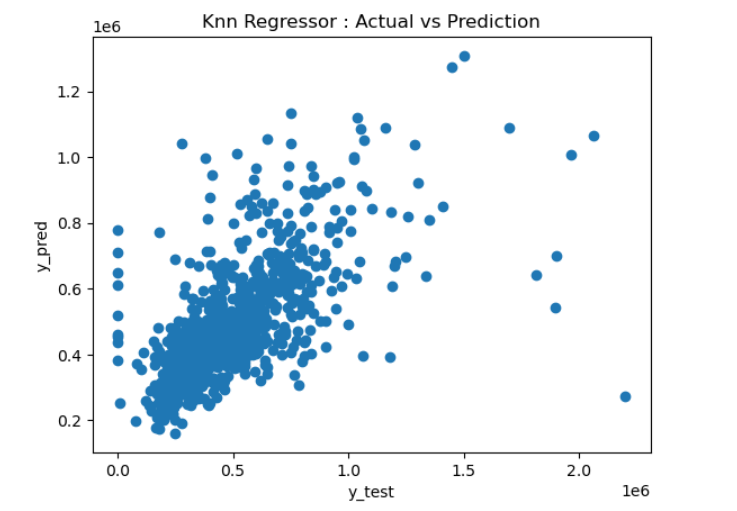
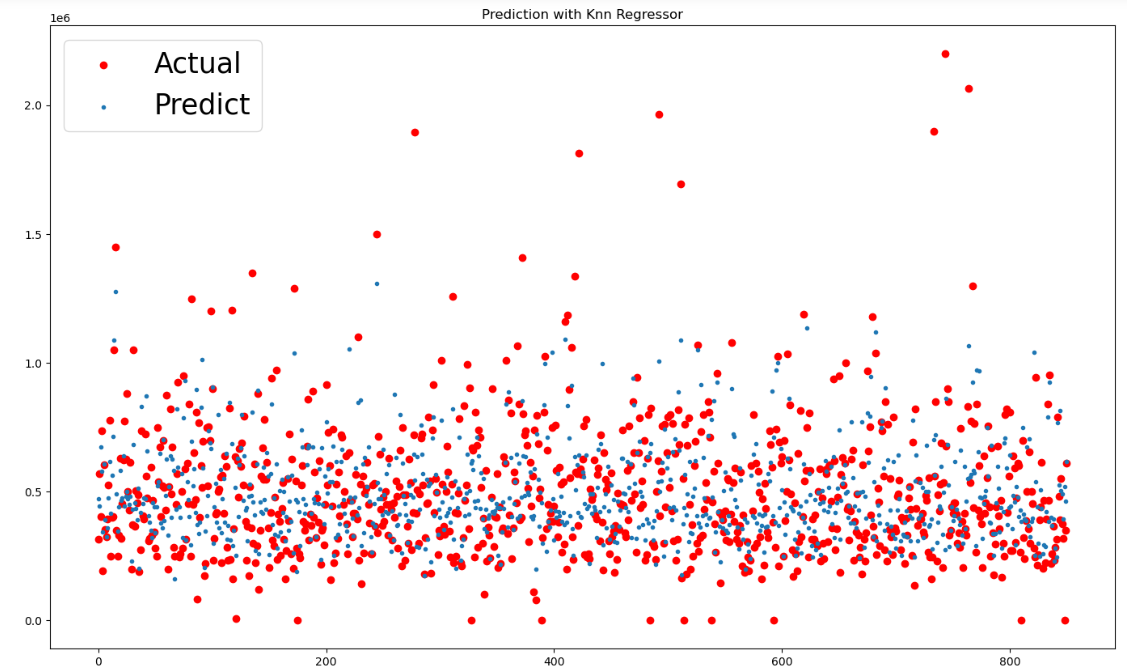
****

****

**Knn Regressor:**

**The RMSE value is 40205880582.642654**

**The R square value is 0.4379306581817678**

****

|  |
| --- |

**2. Image Dataset:**

**2.1) General Information about the dataset:**

**Name of the used Dataset:**  PlantVillage Dataset

**Number of used classes:**  3

**Their Labels:**

1. Pepper\_\_bell\_\_\_Bacterial\_spot

2. Potato\_\_\_Early\_blight

3. Potato\_\_\_Late\_blight

**Toal Number of used Samples:** 2997

**Images Dimensions**: 256\*256

**Averages size of Images:** .2MB

**Number of training samples:**  2397

**Number of testing samples:** 600

**2.2) Implementation details:**

**Average number of Features:** 366

**Average dimension Features:**  128.0

**Hyperameters used in Logistic Regression:**

multi\_class='multinomial'

Means multiple classes not binary as default

solver='lbfgs'

They are optimization algorithms used to find the weights that minimize the loss function and the choice dependes on the size of a dataset and the type of regulralization

max\_iter=2000

In the fitting step this is the number of iterations to reach to convergence (control the convergence) or most smallest loss function which is zero

**Hypermeters used in K-means:**

n\_clusters= 3

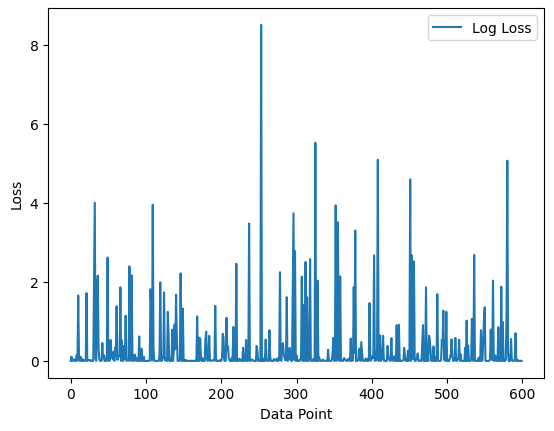
Represents the number of centroids (clusters) that the kmeans will produce

max\_iter=1000

Sets the number of maximum iterations for each initialization of the k-means algorithm

**2.3)Results Details:**

**Logistic Regression:**

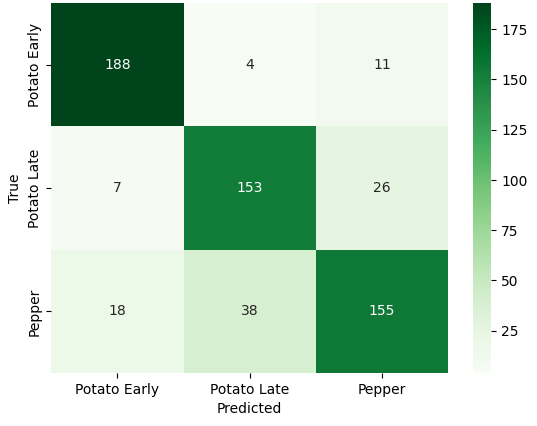
loss curve  


accuracey: 83%

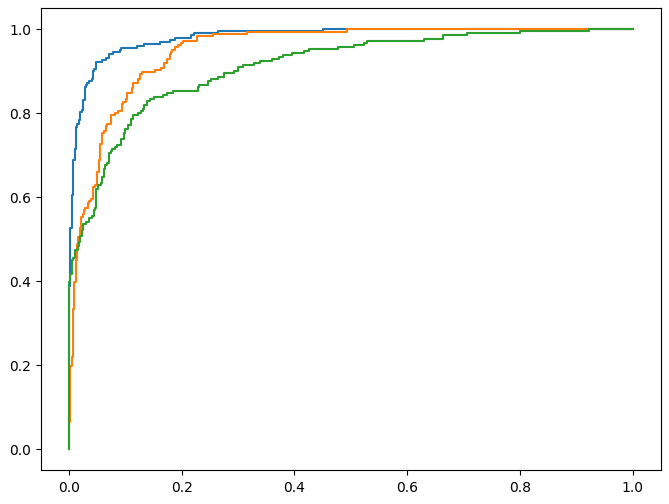
Confusion Matrix:

Rows represent the actual classes.

Columns represent the predicted classes



ROC Curve:

representation of the performance of a binary classification model across different thresholds. It helps you to understand the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various classification thresholds

**K-Means:**

Accuracy: 43%

**Note**: kmeans doesn’t have confusion matrix, roc curve, or loss curve; because it’s unsupervised. Getting those would require mapping the cluster labels to actual labels.