***Project description document***

1. *Image dataset*
2. ***general information about dataset***

|  |  |
| --- | --- |
| *Name of dataset* | ***Eye Diseases Classification*** |
| *Number of classes* | *4* |
| *Lables* | *['glaucoma', 'cataract', 'normal', 'diabetic\_retinopathy']*  *[0,1,2,3]* |
| *Total number of samples* | *4217* |
| *Size of each image* | *32\*32* |
| *Number of training samples* | 3373 |
| *Number of validation samples* | 421 |
| *Number of testing samples* | 423 |

1. ***Implementation details:***

* ***feature extraction phase***

|  |  |
| --- | --- |
| *how many features were extracted?* | *324* |
| *their names* | *-* |
| *The dimension of resulted features* | *4217\*324* |

* ***cross-validation***

|  |  |
| --- | --- |
| *the number of fold* | *5* |
| *ratio of training/validation/testing* | 80:10:10. |

|  |  |
| --- | --- |
| ***Logistic regression*** | * **-tol (Tolerance for stopping criteria):** It determines the stopping criterion. * **-max\_iter (Maximum number of iterations):** This specifies the maximum number of iterations. * **-solver (Algorithm to use in the optimization problem):** The algorithm to use for optimization. In this case, you've chosen 'sag'. |
| ***kmeans*** | * **-n\_clusters:** The number of clusters to form, as well as the number of centroids to generate. This is a crucial hyperparameter and should be set based on the characteristics of your data and the desired number of clusters. * **-n\_init:** The number of times the algorithm will be run with different centroid seeds. distances from each point to its assigned center). * **-random\_state:** The seed used by the random number generator. |

* ***Hyperparameters***

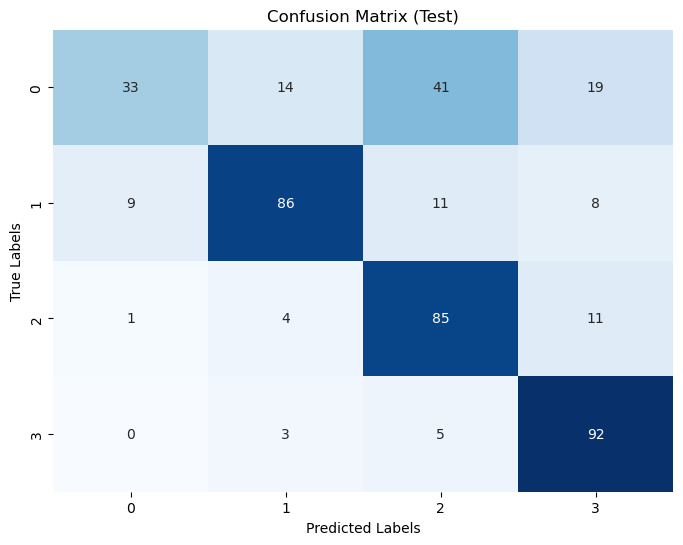
***Accuracy for logistic regression***

******

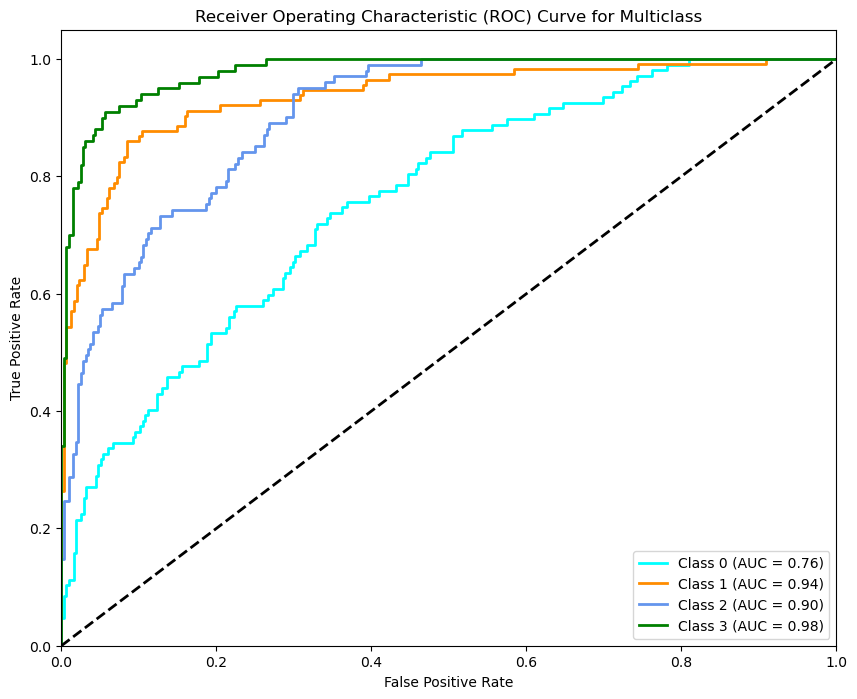
***Accuracy for Kmeans***

******

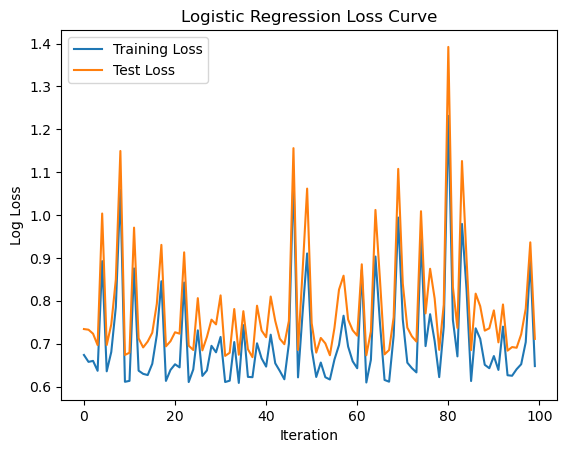
***Confusion Matrix for logistic regression***

**

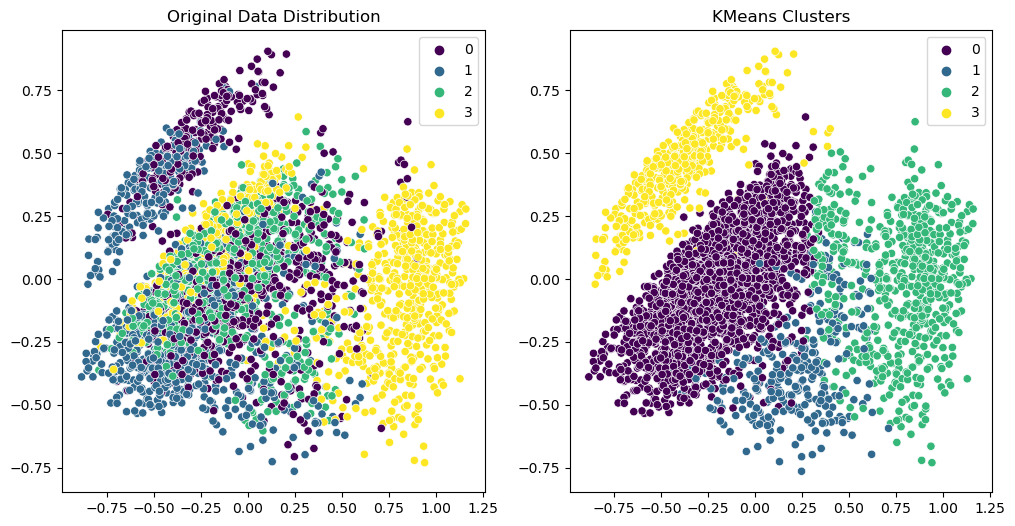
***ROC for linear regression***

******

***Loss curve for logistic regression***

**

***The difference between original data distribution and kmeans clusters***

******

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

***2.Numerical dataset***

***a) general information about dataset***

|  |  |
| --- | --- |
| *Name of dataset* | **Predicting house price** |
| *Total number of samples* | *21613* |
| *Number of training samples* | 17290 |
| *Number of testing samples* | 4323 |

**b)*Implementation details:***

* ***Feature***

|  |  |
| --- | --- |
| *how many features?* | *18* |
| *their names* | [Price, Bedrooms, Bathrooms, Sqft\_living, Sqft\_lot, Floors, Waterfront, View, Grade, Sqft\_above, Sqft\_basement, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15] |

* ***Hyperparameters***

|  |  |
| --- | --- |
| *KNN* | *we used ‘grid search’ and ‘best params’ to find the best hyperparameters*  *-* *Best Hyperparameters: {'metric': 'manhattan', 'n\_neighbors': 9, 'weights': 'distance'}* |
| *Linear regression* | [θ0,…………………..,θ18**]** |

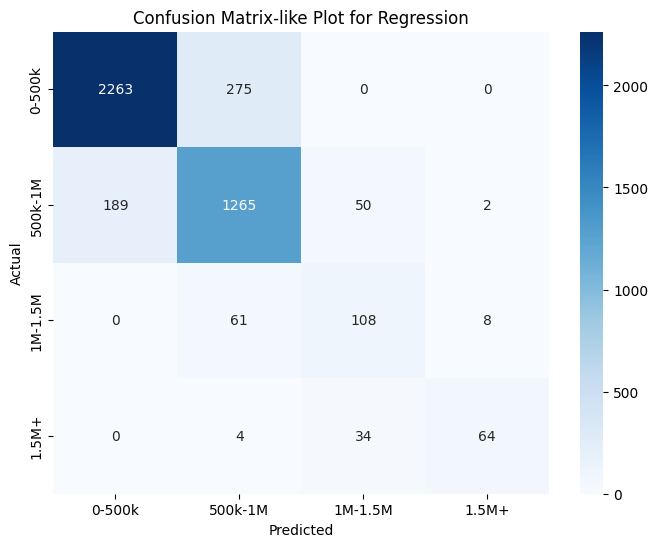
***Accuracy for linear regression***

******

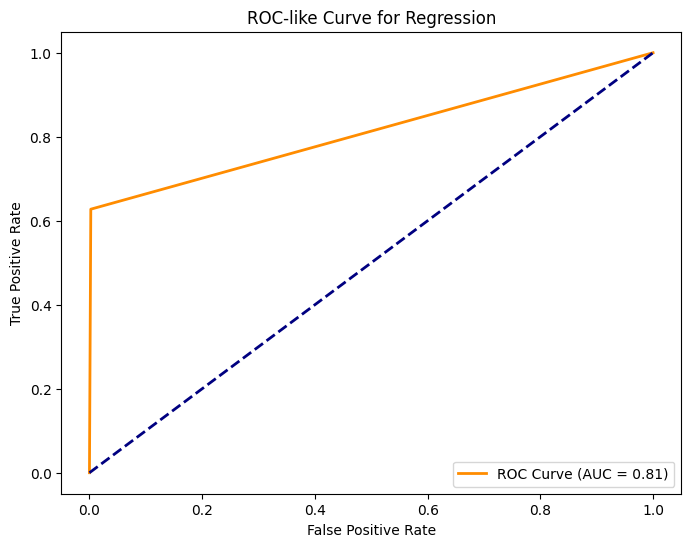
***Accuracy for KNN :***

******

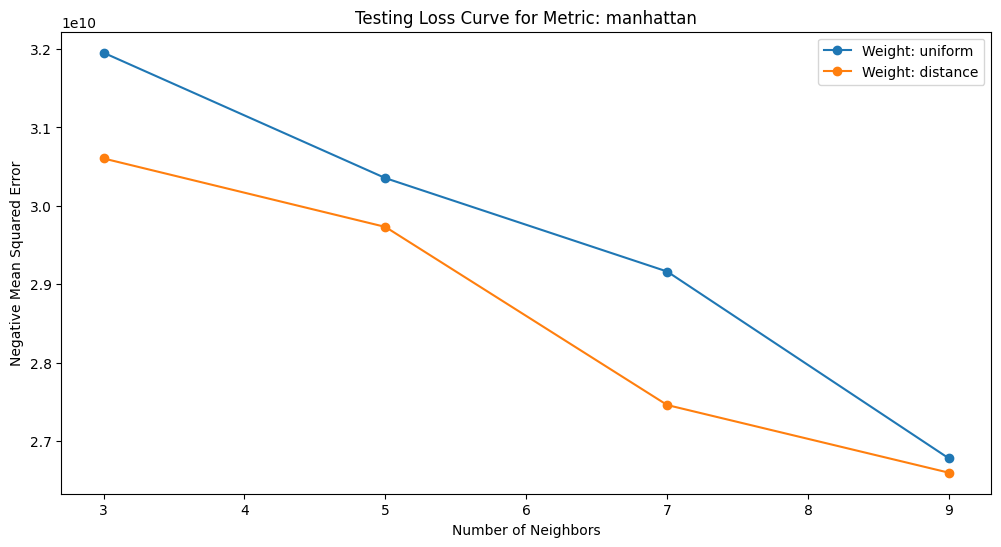
***Confusion Matrix for KNN***

******

***ROC Curve for KNN***

******

***Loss curve for KNN***

******