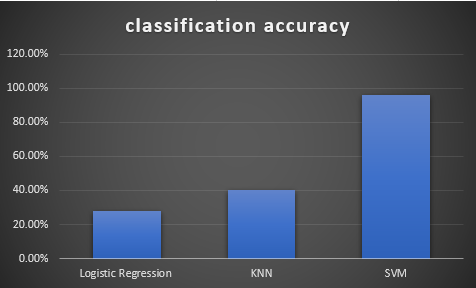
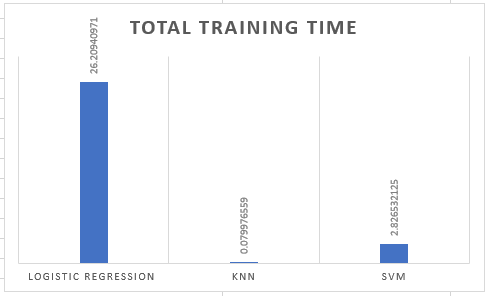
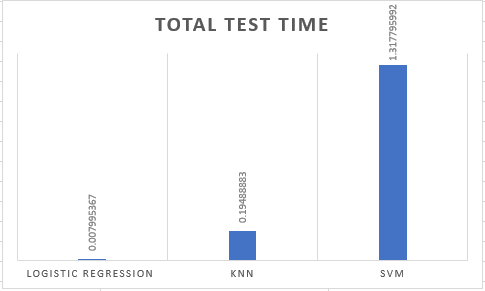
Milestone 2 Report

* Summarize the classification accuracy, total training time, and total test time using three bar graphs.

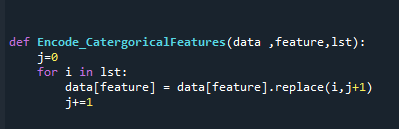






* we change some data preprocessing techniques to make data suitable for classification
* Encode Categorical features

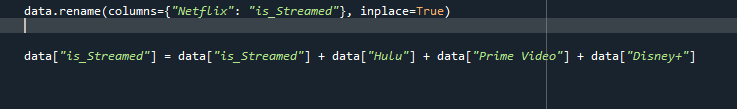
We didn’t use one hot encoding again to avoid curse of dimensionality!



* Impute missing data

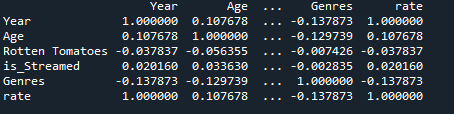


* Aggregate some features (Netflix-Hulu-Prime Video-Disney+)

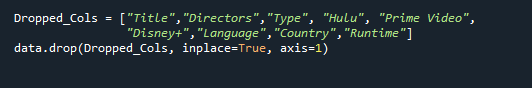


Then, make Feature Selection and drop very uncorrelated features

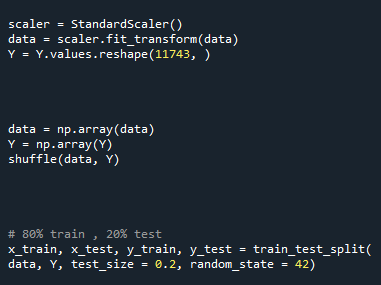
We got the correlation between features



We Selected the features by trial and error, In supervised learning, the trial and error method is mainly used, where we test to give us a known answer, We first enter values and compare them with the correct answer, so they would improve the results gradually until they give a precise answer.

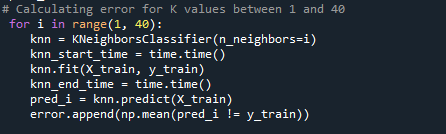


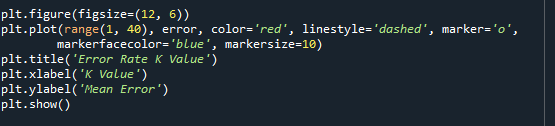
* At Last we Normalize and split data (80% training & 20% test)

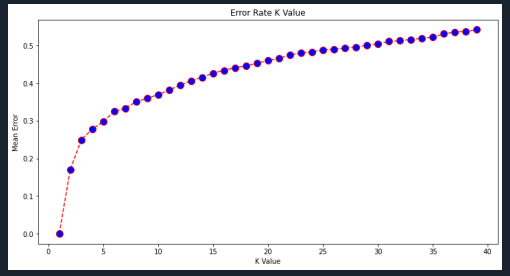


* For hyperparameter tuning we choose
* n\_neighbors in KNeighborsClassifier

Calculating error for K values between 1 and 40 and plot this relation







We notice here increasing of K value Cause increasing in Mean Error!

The best K will be in range of 1 and 5 maybe 3

* C ,the regularized parameter in SVM

Make C=1000

SVC(C = 1000, kernel='rbf' ,gamma=0.1)

Get : Training accuracy 100.0 %

Make C=0.1

SVC(C = 0.1, kernel='rbf' ,gamma=0.1)

Get : Training accuracy is 11.0 %

Make C=1 and kernel linear

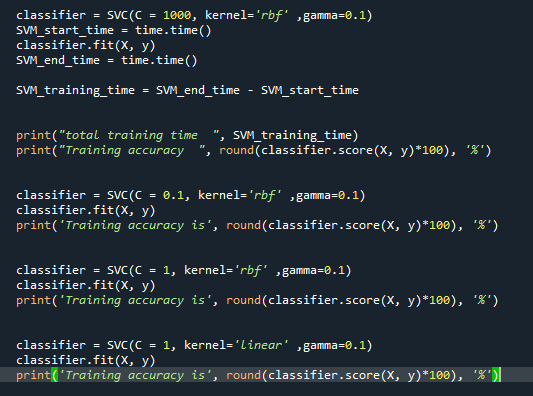
SVC(C = 1, kernel='linear' ,gamma=0.1)

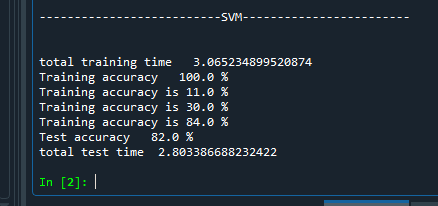
Get: Training accuracy is 84.0 %

Make C=1

SVC(C = 1, kernel='rbf' ,gamma=0.1)

Get: Training accuracy is 30.0 %





* Conclusion

We noticed that SVM have the test smallest error and the highest Accuracy as SVM is useful to solve any complex problem with a suitable Kernel,and it is efficient algorithm ,

it has a regularization parameter so causes generalization.

Knn is not very bad and easy to understand but Knn needs homogenous features, Logistic Regression gives the worst accuracy as it relies on entire data and it's difficult to capture complex relationships using logistic regression.