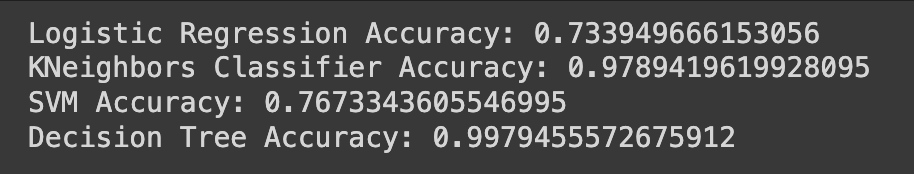
**Netflix Report**

The Netflix dataset consists of movies and tv shows and my main aim was to analyze how different factors affect the trend of tv shows and movies available on netflix. In my work I used the different factors to make graphs, first to visualize the data and then to also see how the total number of shows/movies is increasing or decreasing. At last I made prediction models to predict if these factors will be able to tell the type (tv show/movie) of something if it was available on netflix.

When I started working on the dataset I realized that there were a lot of missing values especially but not in the columns which I wanted to use. I imputed the missing values mainly using mode because these were all categorical columns. Then I made multiple graphs using different factors to just visualize and see how much these were affecting the number of tvshow/movies. I also wanted to look deeper if there was a specific weekday the movie was released on. The one thing I struggled the most was with making the double bar graph which shows the increase/decrease of tv shows/movies in different years. The main problem there was putting different years and their values in one data frame. So I just created a new dataframe all together and individually typed the values there. But i think if i would’ve coded it so it automatically transferred into a new dataframe then it would've been for future changes. After all this i turned the categorical variables into numerical ones, split the dataset into test/train sets, then trained/tested the dataset using 4 classification models: KNN classifier, SVM classifier, Decision Tree and Logistic regression. 

Explanation of the accuracy score:

Logistic regression:

This shows that the model achieved an accuracy of approx 73.4% and this percentage of the dataset was correctly predicted.

KNeighborsClassifier:

KNeighbors classifier works really well when the data has a clear separation between classes and this dataset has that, thats why it has a higher accuracy score of 97.89%. This could also be because of the factors clearly indicated if the film was a movie or a tv show.

SVM Classifier:

SVM tried to maximize the difference between the data points but their performance depends on the choice variables you are using. Here we got 76.73% as the accuracy score which means that there were overlapping data points which were not accurately predicted.

Decision Tree:

We got the highest accuracy score 99.76% for this one and I think this is because this model would’ve had many branches and internal nodes, that's why the data was narrowed down and predicted accurately.