Title of the project – Predicting Piracy Rates of Movies using ML Algorithms

For this project, we will be using Movie Dataset from Pirated Sites data from Kaggle. This data provides information on various movies, including their IMDb rating, appropriate age group, director, downloads, industry, language, posted date, release date, run time, storyline, title, views, and writer. These columns can be used to gain an insight into the factors that contribute to a movie's success, and to build predictive models which will forecast the success of new movie. In this project, we will be performing different data science analyses which will include data cleaning, correlation analysis, and predicting modeling, with the aim of answering the questions:

1) What factors influence the number of downloads (Also meaning mostly pirated) a movie receives? How much of an impact does the industry, language, or release date have on the number of downloads?
2) Does the IMDb rate of any movie affect the download number of the movie? What is the highest pirated movie together and from each industry?
3) Which industry is mostly affected by piracy?

For data science tasks we could build a predictive model that predicts the number of views for a movie based on its various features which will include regression, decision tree or neural networks. Then we could use classification analysis to build a model that predicts the likelihood of a movie being pirated based on its IMDb rating. Here we could train the model on a dataset of movies with known piracy rates and IMDb ratings, and then use it to predict the piracy rates of new movies. We could also build separate models for different industries to identify the highest pirated movie in each industry. Before performing any of the data science tasks data cleaning and preparation is the important part where we will be handling the missing data and dealing with duplicates. After that we will be performing EDA to get a better understanding of the data and identify any patterns or trends which could involve creating visualization, heat maps. This process will help to explore the relationship between various features and the target variable. Correlation is also an important part of the analysis where we could figure out the features that are most strongly correlated with the target variable. Visualization is an important part of this analysis and to incorporate business intelligence aspect we will be using Tableau to make interactive dashboards of our findings from the data science task that we will perform. We could create a scatter plot showing the relation between IMDb and number of views.

There is some missing value and repetition in this data so dealing with that will be one of the first thing we will need to do to get a better output. Language could be one of the factors which could potentially cause biases in the output when analyzing this dataset. We could use visualization to see different languages and see how their IMDb ratings differs from one another.

Weekly Timeline

Week 1 – First Proposal

Week 2 – Second Proposal and GitHub repository

Week 3 – EDA and data cleaning

Week 4, 5,6 – Data Science Task/ ML learning / Visualization

Week 7 – PowerPoint Presentation

Week 8 – Conclusion