**Note: There is a ppt explaining the starting part of the project, you might have a look at it though I am explaining everything here.**

**Project:-**

This is a machine Learnig based Python project. The aim of the project was to develop model(s) to predict the success of Bollywood Movies, their IMDb Rating to be specific. Hence,in a gist I used algorithms like Regression,SVMs,Neural Networks to predict the rating.

Features Used:-

Actor rating

Director rating

Genre

Budget

Youtube Trailer views

Runtime

All these features were available to us before the movie release. How, these features affect the rating was seen graphically (in graphs folder).

**Dataset**:-

Our database included movies from 2011 to 2015. Only those movies were included which had a budject above 5 crores in order to reduce the variation of movies. Also, since the data extraction APIs (omdbapi.com,cinemalytics.com) did not have complete data, our complete dataset included only 150 movies spanning over a period of 5 years. Data from these two sources was available in the JSON format and from there it was exracted and compiled into a single database using Python scripts(data.tsv).Youtube trailer data was extracted by downloading the html pages and extracting the views using Python scripts.

**Feature formation:-**

Actor rating: The rating of various Bollywood actors was downloaded from cinemalytics.com (critics rating). The highest rating of an actor in a movie was set as its Actor rating.

Director rating: The rating of various Bollywood directors was downloaded from cinemalytics.com (critics rating) and IMDb.com. The rating of the director in a movie was set as its Director rating.

Genre: The genres were labeled from 1 to 18 and they were analysed to see their effect on the IMDb rating. After some modifications the genres from 10 to 18 showed higher IMDb rating than the rest.

Budget: The budget of the movie divided by 10,00,00,000 was the feature I used as budget.

Runtime: The runtime feature of a movie was taken as the no. of minutes.

Youtube: The no. of views of the movie trailer was directly used as the feature. Earlier it was thought of taking the log(base 10) of the views but it was rejected as it reduced the precision.

**Regression**: It was already tested how the features performed individually against the IMDb rating( python library used 'scipy' ). Finally, a model was proposed including only Actor rating,Director rating and Genre as features. It was a mathematical model combining the individual models of the features. All the files are there in the 'Regression' folder, python file for testing the model (calculating errors), model.txt file for explaning the result.

**FuzzyLogic**: In this algorithm also I used only 3 features Actor rating,Director rating and Genre. The Membership functions and rules were decided and the model was trained using python library 'skfuzzy'. Through trial and error the rules and Membership functions were modified to get better results. Have a look in the 'FuzzyLogic' folder. The FuzzyLogic.txt file shows errors for all data. The python file was used for the training and error analysis.

**SVR (Support Vector Regression)**: All the features were used for this algorithm. The traning data was fed as input and the model was trained(fit into the data). Then this model was tested using the test data and the results were analysed. Python library used 'svm' in 'sklearn'. Have a look at the 'SVM' folder.

**Neural Networks**: The most widely used machine learning algorithm. This algorithm gave us the best results.The traning data was fed as input and the model was trained. Then this model was tested using the test data and the results were analysed. The Python library used was pybrain. Have a look in the 'NeuralNetwork' folder.

Note:- You can run the python scripts and see the results.

Conclusion:-

Have a look at the conclusion.txt file.