Predicting IMDb Movie Rating using Machine Learning

Ananya Jain | ananya19408@iiitd.ac.in
Manasvi Singh | manasvi19369@iiitd.ac.in
Pritish Wadhwa | pritish19440@iiitd.ac.in
Yash Bhargava | yash19289@iiitd.ac.in

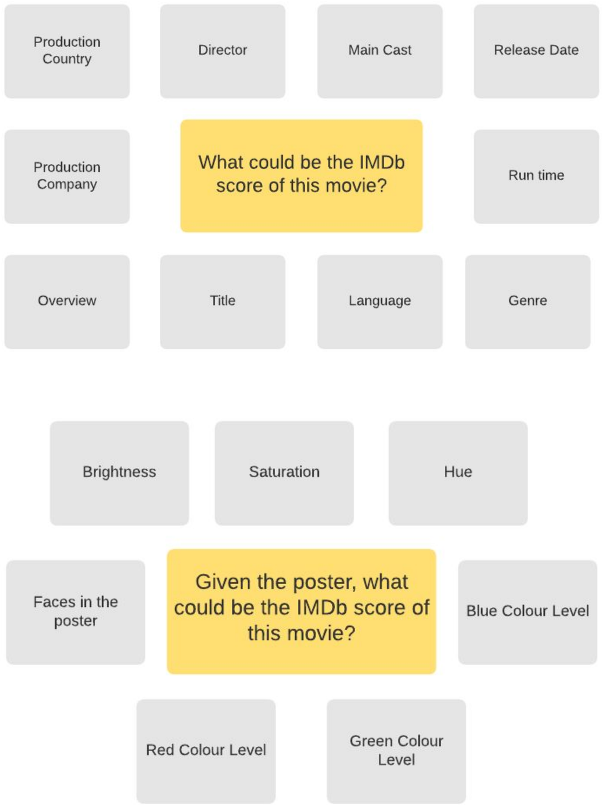
INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Motivation[1]

- Is it possible to predict, **ratings of a movie prior to its release or production**?
- **IMDb rating** is the single most influential factor in deciding consumer’s opinion and inherently the blockbuster success of a movie
- With the machine learning techniques at our disposal, our aim to predict the **IMDb rating of any movie before its theatrical release**
- Successfully predicting IMDb rating is **beneficial to both producers** (from a financial standpoint) and **consumers** (from an entertainment standpoint) alike.
- We aim to identify the **features** that may cause the making of a high rated movie.

Motivation[2]

- Answering the questions:
  - Given pre release features, what could be the IMDb Score of this movie?
  - Given the poster of the movie, what could be it's IMDb Score?
  - Which features are important in this prediction?
  - How accurate can we be?
  - Which is the best model for this purpose?



Literature Review [1]

“**Predicting IMDB Movie Ratings Using Social Media**” by Oghina, Andrei & Breuss, Mathias & Tsagkias, Manos & Rijke, Maarten

- Addressed the task of predicting Imdb movie ratings using data from social media .
- They identified qualitative and quantitative activity indicators for a movie in social media,and extracted two sets of surface and textual features.
- They discovered on training various models and analysis that the fraction of the number of likes and dislikes on YouTube, combined with textual features from Twitter lead to the best performing model.
- The models had strong agreement with the observed ratings and high predictive performance

Literature Review [2]

"**A machine learning approach to predict movie box-office success**" by N. Quader, M. O. Gani, D. Chaki and M. H. Ali,

- Proposed a decision support system for movie investment sector using machine learning techniques.
- The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic.
- They discovered that *budget, IMDb votes and no. of screens* are the most important features which play a vital role while predicting a movie’s box-office success

Dataset Description[1]

Numerical Data

1. release date: Date of release
2. runtime : Duration of the movie

Image Data:

1. num_faces : Number of faces in the poster
2. saturation : Saturation level of the movie poster
3. hue : Hue level of the movie poster
4. brightness_sd : standard deviation of brightness level of the movie poster
5. saturation_sd :standard deviation of saturation level of the movie poster
6. hue_sd :standard deviation of hue level of the movie poster
7. blue_sd : standard deviation of blue colour level in the movie poster
8. Green :green colour level in the movie poster
9. green_sd: standard deviation of green colour level in the movie poster
10. red :red colour level in the movie poster
11. red_sd :standard deviation of red colour level in the movie poster

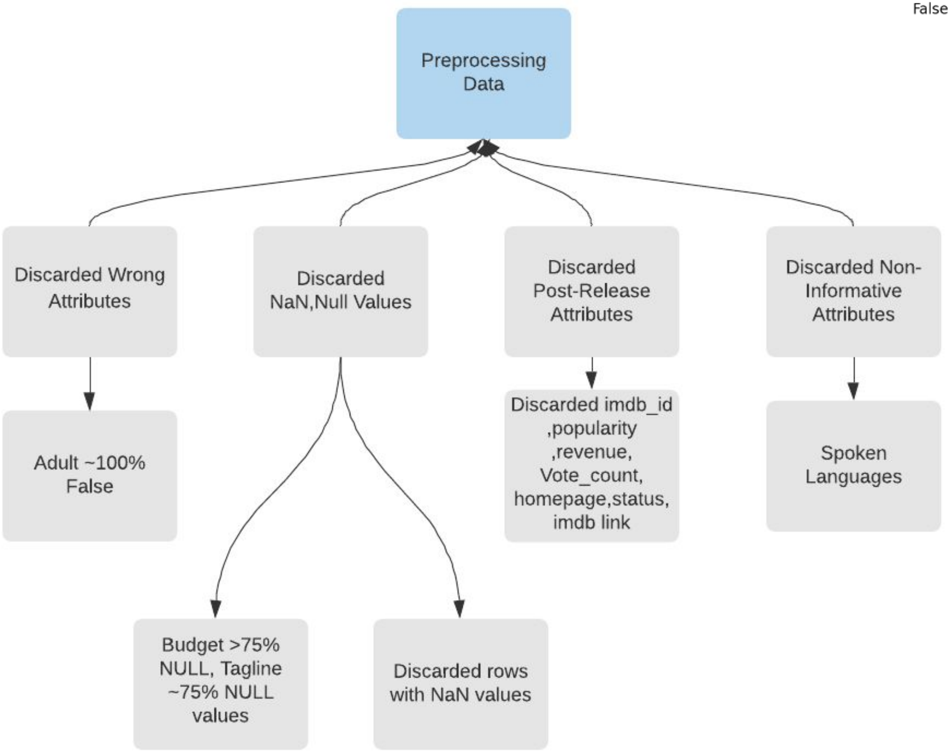
Categorical Data

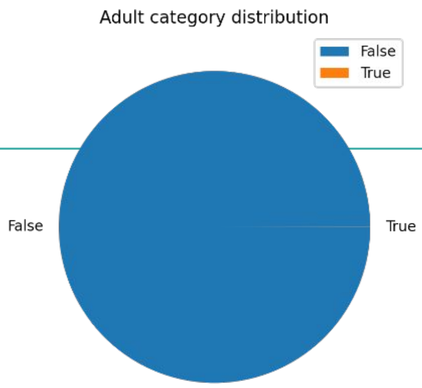
1. Genre : The different genres of the movies
2. Language : The main language spoken in the movie
3. Production Companies : The company producing the movie
4. Production Countries :The main countries where the movie was produced

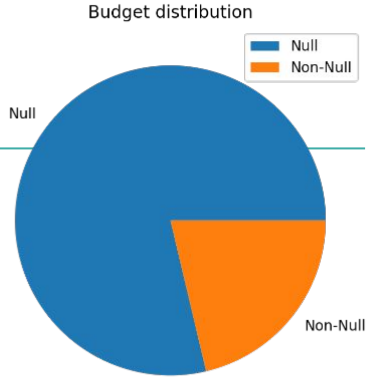
Text Data

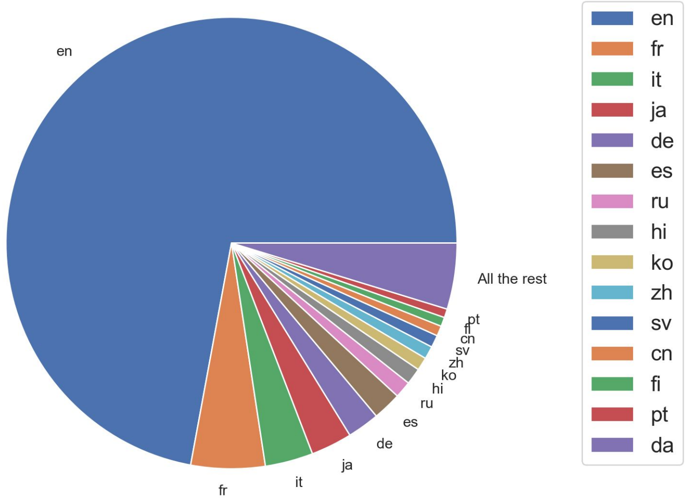
1. title : The title of the movies
2. Keywords: The major themes in the movie
3. Overview : The summary of the movie
4. cast :The actor and actresses in the movie
5. crew :The every department involved in movie production

Dataset Description[2]

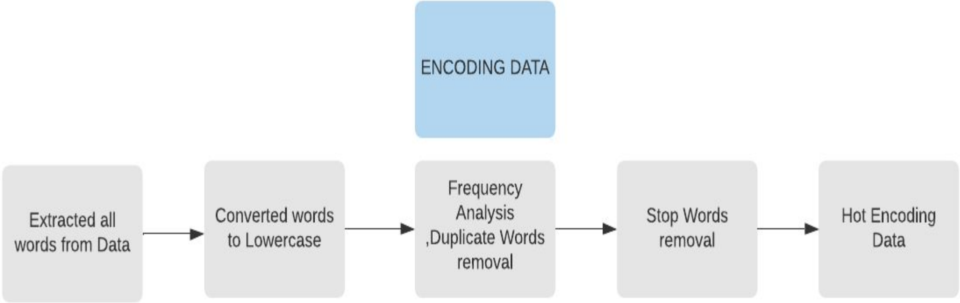


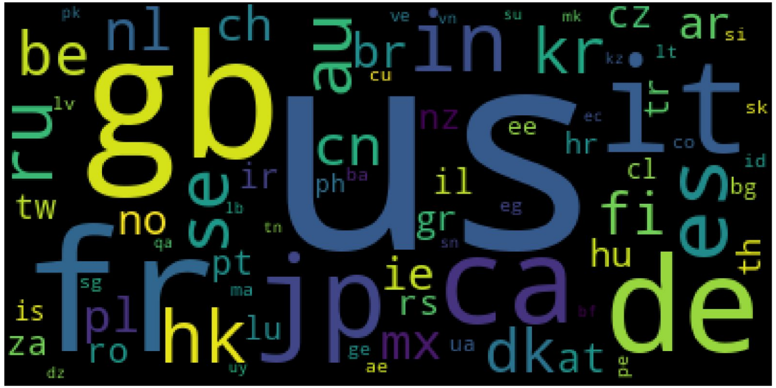


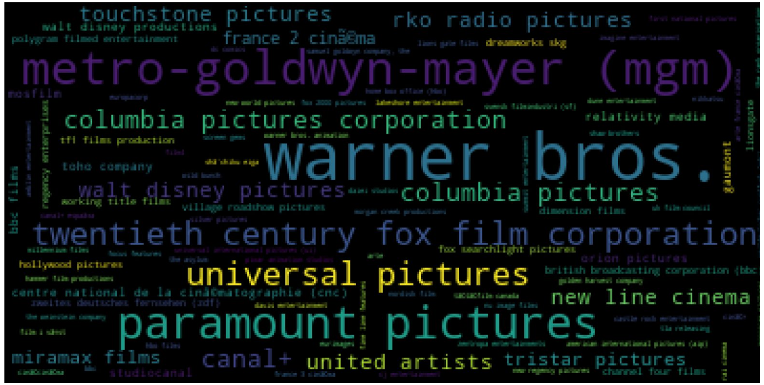




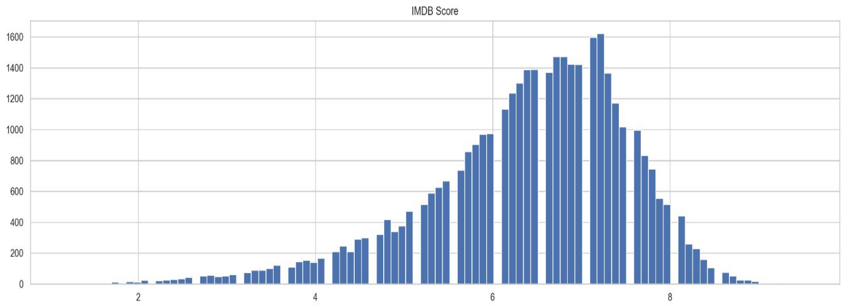
Dataset Description[3]

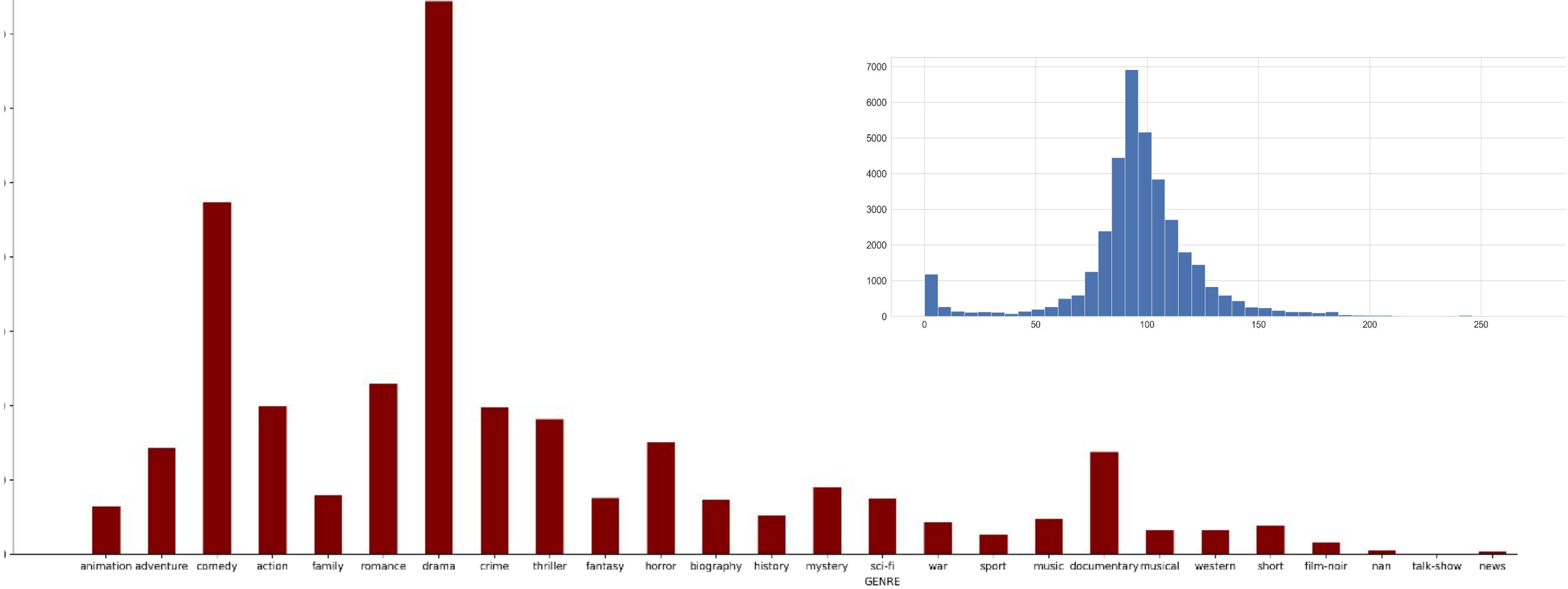




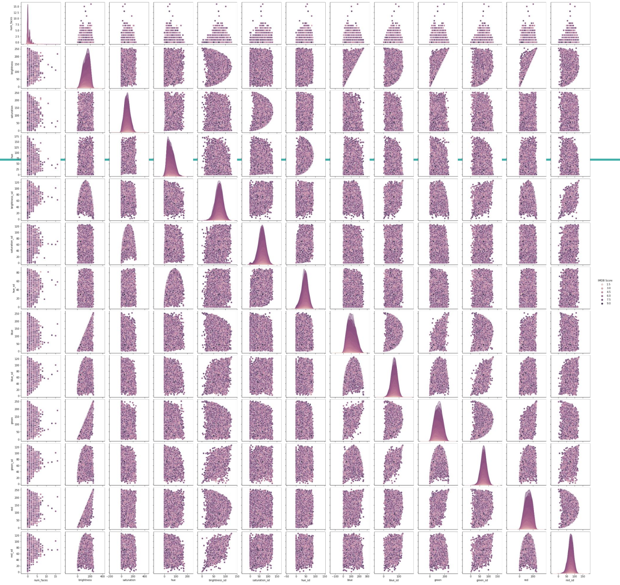


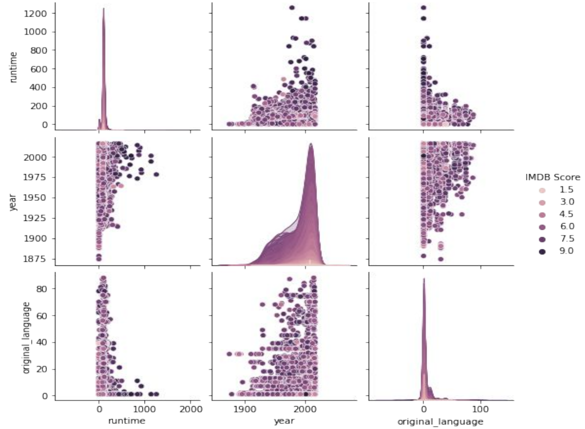
Dataset Description[4]



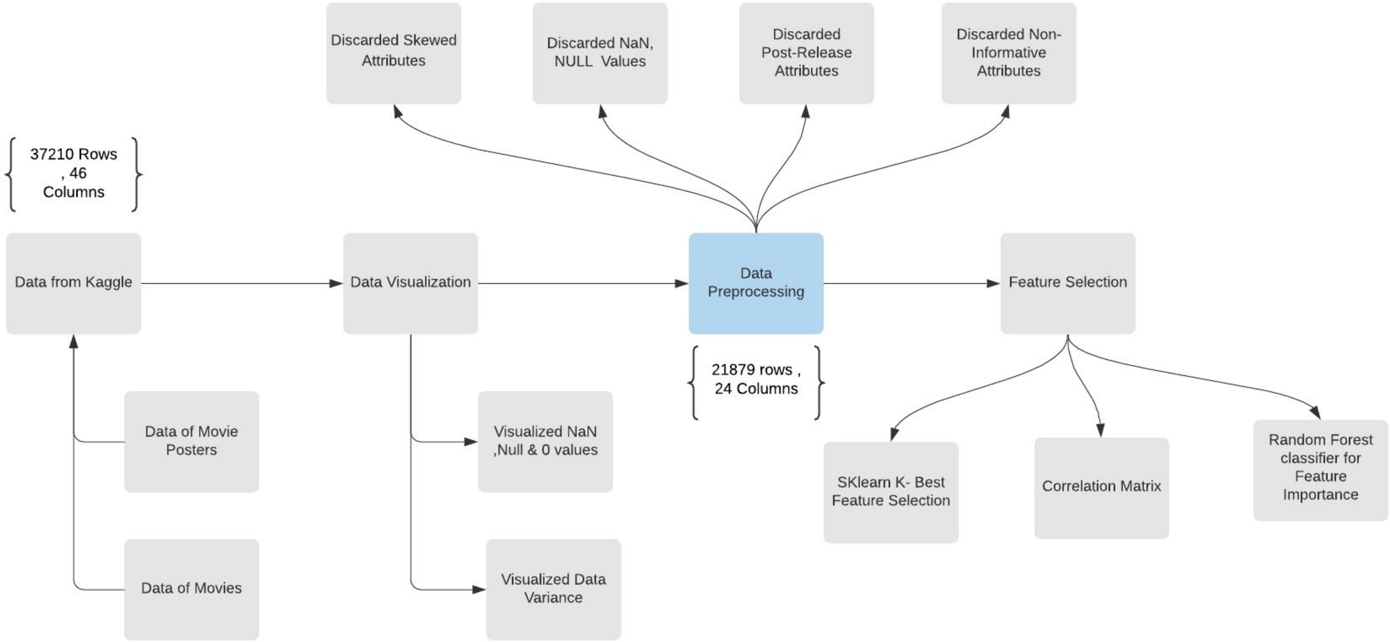


Dataset Description[5]

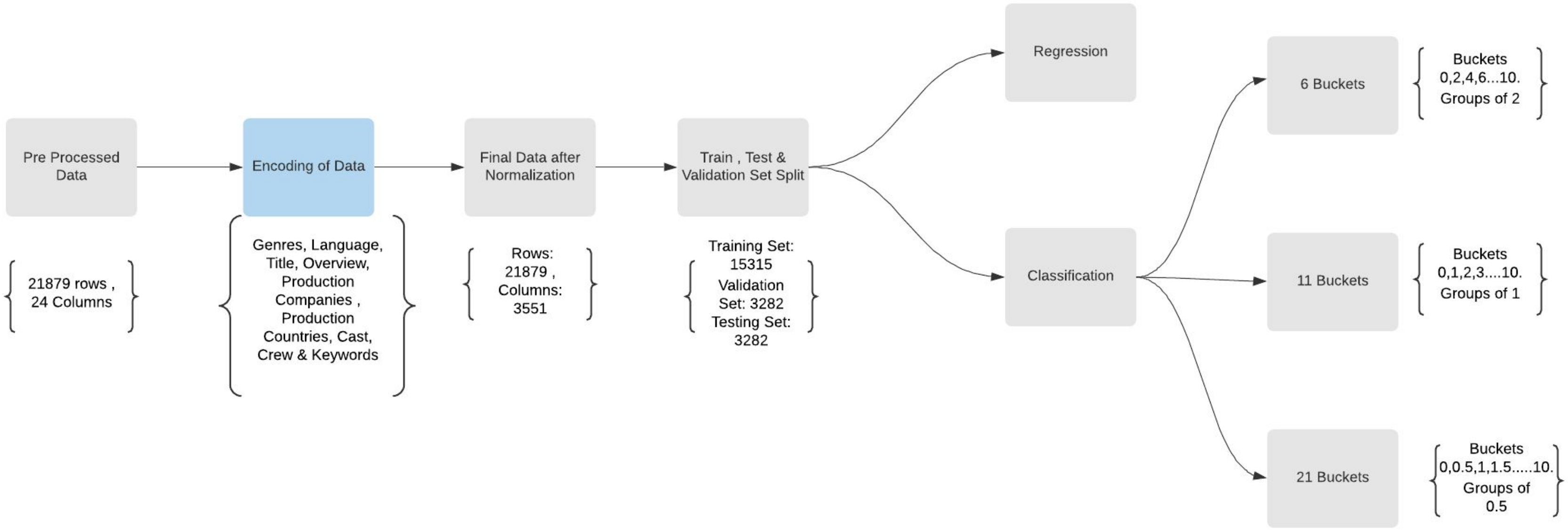




Methodology[1]



Methodology[2]



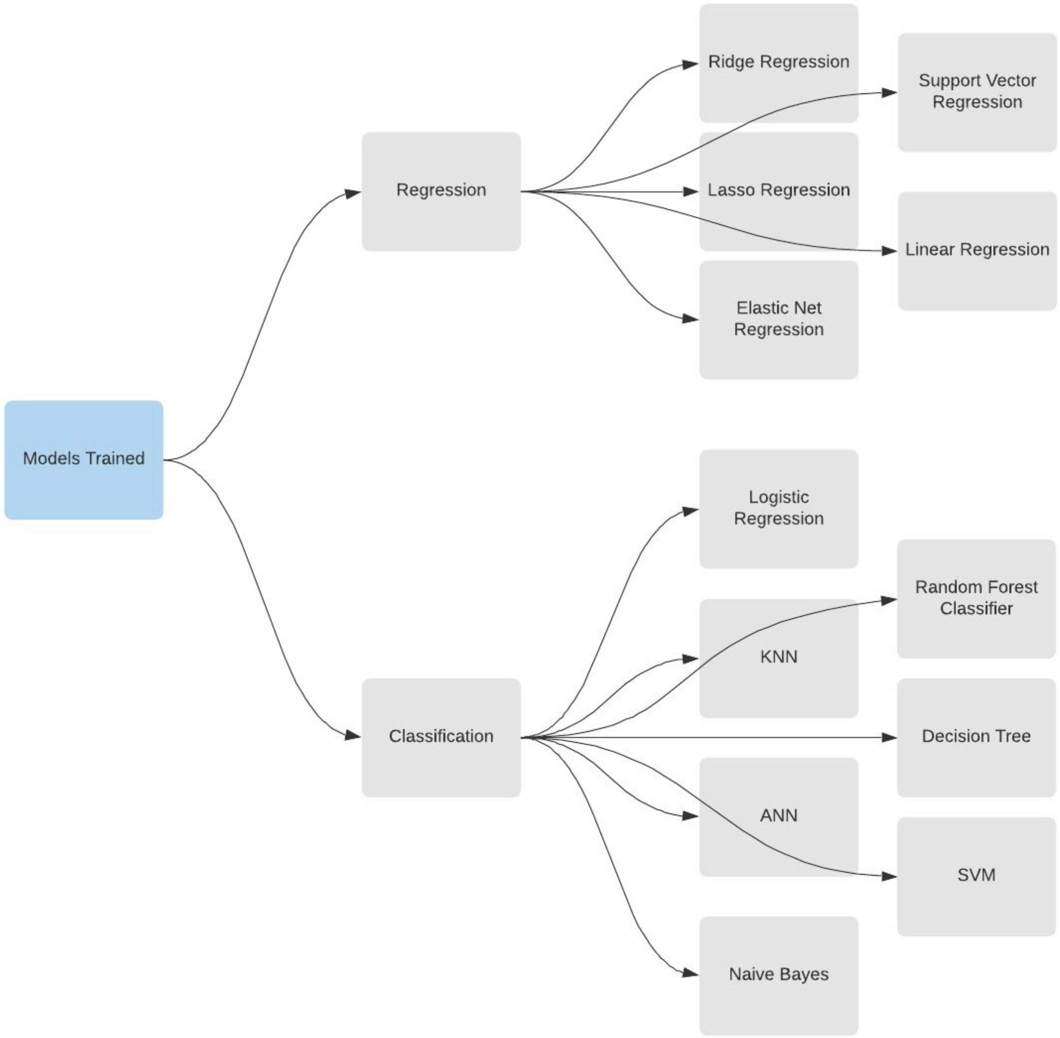
Metrics for Analysis

For Regression we have used Mean Squared Error (MSE) and R^2 metrics. MSE is used for calculating loss function. R^2, the coefficient of determination is a measurement of proportion of variance for predicted IMDb ratings.

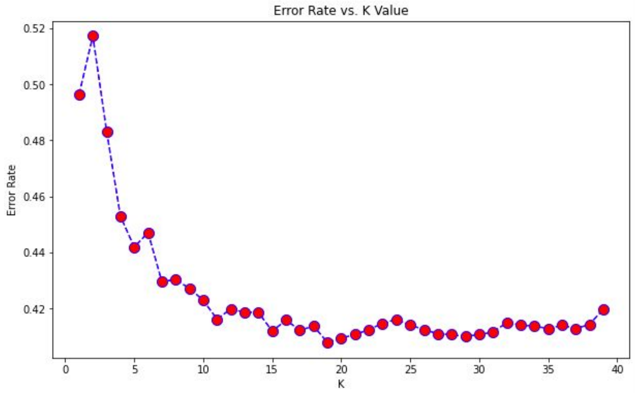
For classification we have used Accuracy,weighted Precision, Recall and F1 to measure the performance of our models. While accuracy gives us a measure of how accurate our predictions are, precision indicates the number of true predictions to all predicted true predictions. Recall indicates the number of true predictions to all actual true prediction and F1 helps score the precision and recall cumulatively.

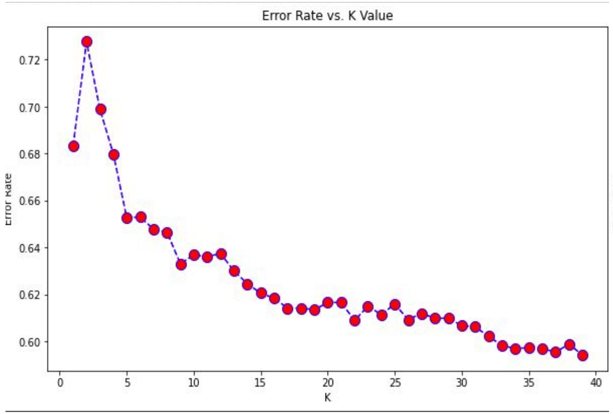
Model Details

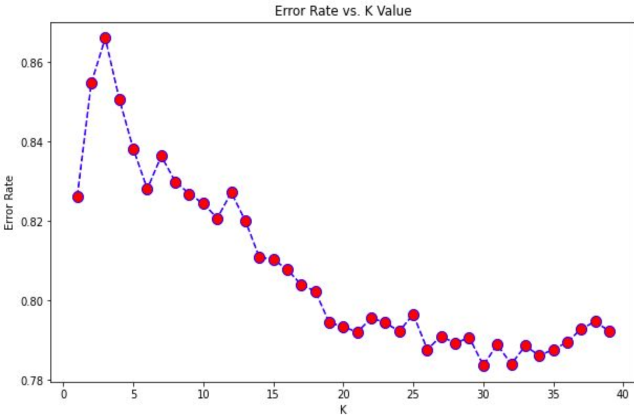
Hyper Parameter tuning was performed in all models to obtain best results using GridSearchCV.



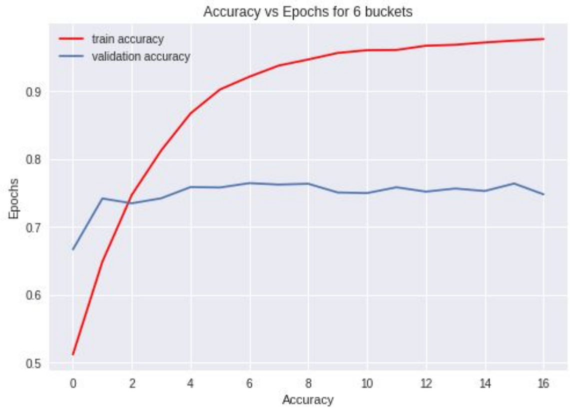
Graphs for KNN

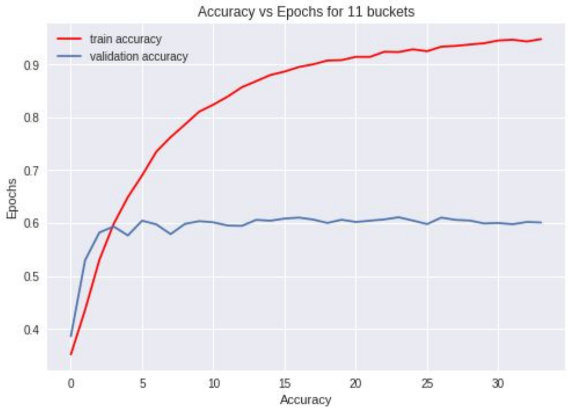


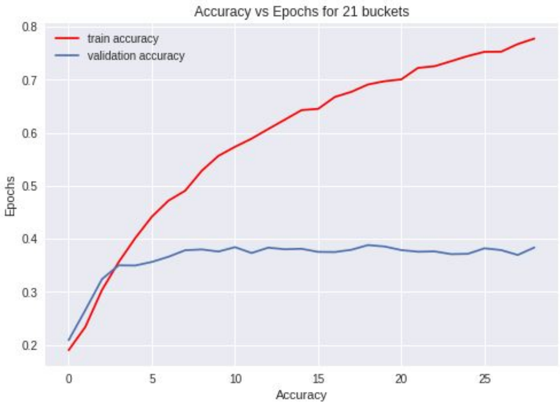


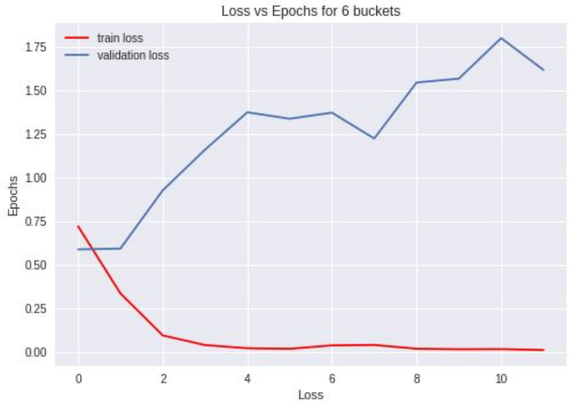


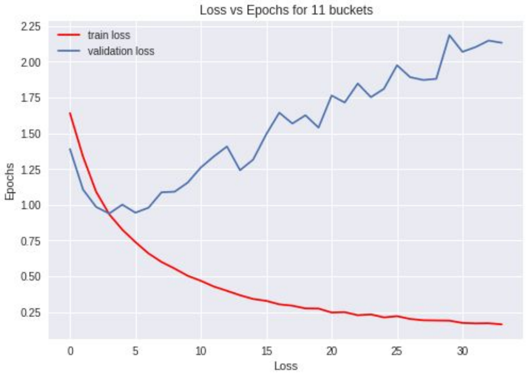
Accuracy and Loss Graphs for MLP

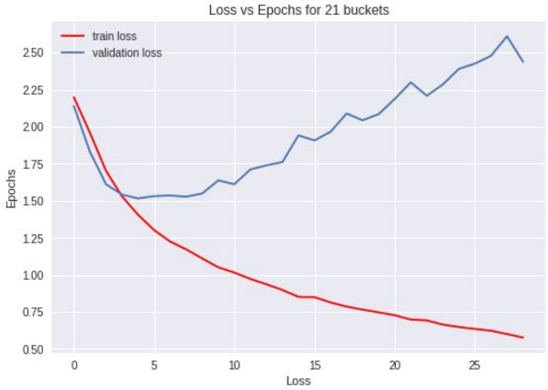










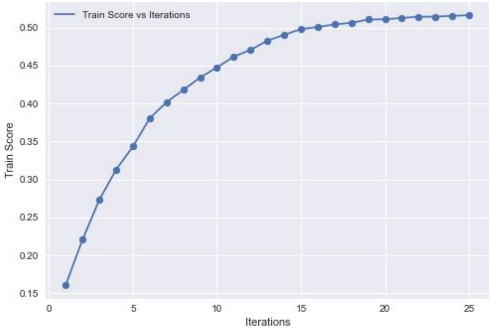


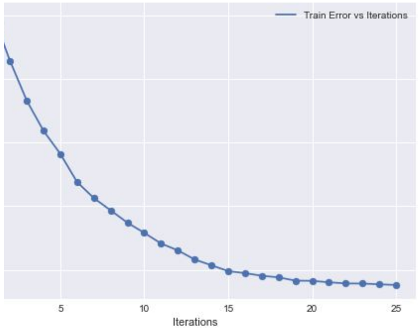
Result / Analysis /Conclusion [2]

Model	R^2	MSE
Linear Regression		
Training Set	0.5951	0.4824
Testing Set	0.3462	0.7739
Lasso Regression		
Training Set	0.3860	0.7314
Testing Set	0.3860	0.7268
Ridge Regression		
Training Set	0.5175	0.5748
Testing Set	0.4236	0.6823
ElasticNet Regression		
Training Set	0.3811	0.7374
Testing Set	0.3821	0.7314
Support Vector Regression		
Training Set	0.4081	0.7605
Testing Set	0.3087	0.8843

Table 2. Results of Regression Models

For regression the best performing model was **Ridge.** We infer that most of the features selected after processing data are important as Ridge outperforms other linear regressors.





Result / Analysis /Conclusion [2]

Model	Accuracy	Precision	Recall	F1
Logistic Regression				
Training	0.79235	0.79859	0.79235	0.7854
Testing	0.6315	0.6144	0.6315	0.6165
Random Forest Classifier				
Training	1.0	1.0	1.0	1.0
Testing	0.6588	0.6588	0.6588	0.6182
K-Nearest Neighbours (K=18)				
Training	0.6157	0.6498	0.6157	0.5562
Testing	0.5862	0.5741	0.5862	0.5227
Decision Tree				
Training	0.66	0.65	0.66	0.64
Testing	0.65	0.63	0.65	0.63
Artificial Neural Network(MLP)				
Training	0.98	0.97	0.98	0.98
Testing	0.78	0.77	0.78	0.77
Support Vector Machine				
Training	0.9920	0.9847	0.9919	0.9882
Testing	0.9787	0.9721	0.9786	0.9753
Naive Bayes				
Training	0.6630	0.6769	0.6630	0.6669
Testing	0.5865	0.5977	0.5865	0.5899

Table 3. Results of Classification Models with 6 buckets

Model	Accuracy	Precision	Recall	F1
Logistic Regression				
Training	0.66047	0.6767	0.66047	0.6532
Testing	0.4323	0.4169	0.4323	0.4136
Random Forest Classifier				
Training	1.0	1.0	1.0	1.0
Testing	0.4596	0.4783	0.4596	0.3848
K-Nearest Neighbours (K=38)				
Training	0.4593	0.4502	0.4593	0.3949
Testing	0.4011	0.3582	0.4011	0.3356
Decision Tree				
Training	0.47	0.43	0.47	0.43
Testing	0.46	0.42	0.46	0.42
Artificial Neural Network				
Training	0.99	0.99	0.99	0.99
Testing	0.61	0.61	0.61	0.60
Support Vector Machine				
Training	0.9925	0.9872	0.9824	0.9766
Testing	0.9538	0.9412	0.9537	0.9470
Naive Bayes				
Training	0.5523	0.5515	0.5523	0.5493
Testing	0.4121	0.4025	0.4121	0.4042

Table 4. Results of Classification Models with 11 buckets

Model	Accuracy	Precision	Recall	F1
Logistic Regression				
Training	0.6421	0.6607	0.64	0.64
Testing	0.2270	0.2170	0.2270	0.2171
Random Forest Classifier				
Training	1.0	1.0	1.0	1.0
Testing	0.7261	0.7158	0.7261	0.6685
K-Nearest Neighbours (K=29)				
Training	0.3078	0.3078	0.2919	0.2601
Testing	0.2093	0.1968	0.2093	0.1809
Decision Tree				
Training	0.34	0.36	0.34	0.32
Testing	0.25	0.23	0.25	0.22
Artificial Neural Network				
Training	0.99	0.99	0.99	0.99
Testing	0.39	0.38	0.39	0.38
Support Vector Machine				
Training	0.9589	0.9587	0.9588	0.9519
Testing	0.8773	0.8572	0.8773	0.8655
Naive Bayes				
Training	0.2197	0.2097	0.2197	0.2121
Testing	0.2294	0.2173	0.2294	0.2215

Table 5. Results of Classification Models with 21 buckets

SVM outperformed them all. This might be because of SVM's ability of extrapolating the features to higher dimensions thus making the classification more prominent.

Conclusion [1]

We had started off by taking up the daunting challenge of predicting IMDb ratings of any movie even before they are released.

To do this we considered various features like length of the movie, the director of the movie and movie genres. These were the most vital features to make any kind of predictions.

We also observed certain other features like the poster of the movie, the movie overview(summary), language of the movie, and country of production.

These features played an important role when taken together, but individually they were not as strong as the features above.

Conclusion [2]

The problem was divided into 2 parts:

1. Regression

   Ridge Regression performed the best giving the R^2 score of .42 on the testing set.

2. Classification

   SVM performed exceedingly well(even better than the most basic neural network) on the classification task, giving the accuracies as 98, 95 and 88% on the 6 bucket, 11 bucket and 21 bucket classification respectively.

Future Work & Learning

There are a lot of aspects which can be further explored.Use of deep learning model is one of them. Music of the movie along with its posters, can also play a huge part and this aspect can also be explored.

We all learnt how to curate a dataset, perform EDA on a large dataset, how to preprocess data, handle null values,create sparse matrices and train machine learning models.

Timeline

This was the last proposed timeline, and we followed it precisely!

Week 6,7 : Decision Tree, Random Forest, SVM

Week 8 : Basic Neural Networks

Week 9-10 : Comparative analysis of various machine learning models, Selecting Best Model

Week 11: Final Report *[Present]*

Individual Contribution

- Each member played a crucial part in discussions, analysis & making this report.
- Ananya & Manasvi performed EDA & Preprocessing.
- Pritish & Yash curated dataset, feature selection & its analysis.
- Regression models were trained & explored by Ananya & Yash.
- Classification models were looked over by Manasvi & Pritish.