Movie Rating Prediction

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfilment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Devi G (AP21110010676)

Poojitha S (AP21110010679)

Sahithi J (AP21110010694)

Meghana G (AP21110010701)

Under the Guidance of

Hemanth Kumar Kalluri (Assistant Professor, CSE)

SRM University–AP

Neerukonda, Mangalagiri, Guntur,

Andhra Pradesh – 522 240

[May 2024]

Certificate

Date: 13-05-2024

This is to certify that the work present in this Project entitled **"Movie Rating Prediction"** has been carried out by **Devi G, Poojitha S, Sahithi J, Meghana G** under my supervision. The work is genuine, original, and suitable for submission to the SRM University–AP for the award of Bachelor of Technology in School of Engineering and Sciences.

Supervisor

Dr. Hemanth Kumar Kalluri

Assistant Professor, CSE

SRM University AP

Acknowledgements

We would like to express our sincere gratitude to Hemanth Kumar Kalluri sir for the invaluable guidance, unwavering support, and encouragement throughout the duration of this research/project work. Mahesh sir’s knowledge, expertise, and insightful feedback have been instrumental in shaping the outcome of this work. We are also grateful to **SRM University** for providing the necessary resources and facilities for the successful completion of this work.

We are also grateful to the participants who generously gave their time to take part in our study.

We are indebted to our colleagues and friends for their encouragement and support throughout the research process.

Thank you all for your valuable contributions.

Table Of Contents

Abstract

Movie ratings prediction is a pivotal application of machine learning techniques, offering valuable insights for both viewers and content providers. In today's digital age, where streaming platforms abound and viewer preferences vary widely, predicting movie ratings accurately is essential for enhancing user experience and optimizing content recommendation systems. By analysing diverse features such as genre, cast, director, release date, and user reviews, machine learning models can effectively forecast the potential rating a movie might receive.

The complexity of viewer preferences and the dynamic nature of the movie industry make movie rating prediction a challenging task for traditional approaches. However, with machine learning powered by Python, this process becomes more streamlined and accurate. By leveraging advanced algorithms and techniques like sentiment analysis, feature engineering, and ensemble learning, predictive models can discern subtle patterns and relationships within the data, ultimately improving the accuracy of rating predictions.

In the realm of entertainment, where viewer engagement and satisfaction are paramount, machine learning-driven movie rating prediction holds immense promise. By harnessing the power of data and automation, content providers can tailor their offerings more precisely to audience preferences, leading to a more personalized and enjoyable viewing experience for users worldwide.

Introduction

Predicting movie ratings through machine learning mirrors the process of loan prediction in many ways, yet in the realm of entertainment consumption. Just as lenders scrutinize applicants' backgrounds to assess creditworthiness, content platforms analyse various facets of movies to gauge audience reception. Features such as genre, cast, director, release date, and viewer reviews serve as the cinematic equivalents of credit scores, loan amounts, and lifestyle considerations. Much like how past repayment histories inform loan approvals, a movie's past performance and similarities to previously well-received films can indicate its potential success.

In the movie rating prediction domain, machine learning models are instrumental in distilling insights from vast datasets of historical viewer preferences and movie attributes. By formulating the prediction challenge as a data science problem akin to loan prediction, these models analyse patterns and correlations within the data to forecast the likely rating a movie will receive. Utilizing techniques such as sentiment analysis, feature engineering, and ensemble learning, these models uncover intricate relationships between movie features and audience reactions, enabling accurate predictions.

The application of machine learning in movie rating prediction facilitates quicker and more informed decision-making for content creators and streaming platforms. By automating the prediction process and providing probability estimates of a movie's potential success, these models empower stakeholders to optimize their content offerings and enhance user experience. Moreover, continuous monitoring and feedback mechanisms enable iterative improvements, ensuring that the prediction models evolve alongside changing viewer preferences and industry trends.

Methodology

1. **Movie Dataset:**

   The movie dataset we used includes films released after the year 2000. To ensure the reliability of our data, we filtered the dataset to include only movies that had at least 10,000 raters and a minimum budget of 10,000 USD. This filtering was essential to eliminate newer movies with insufficient ratings and very low-budget movies that could skew the accuracy of our predictions. By setting these criteria, we aimed to focus on well-established movies that would provide a more stable basis for our machine learning models.

2. **Determine the training and testing data:**

   We divided our dataset into training and testing sets to facilitate model evaluation and improve prediction accuracy. Specifically, we used 80% of the data for training and the remaining 20% for testing. This split allows the model to learn from a substantial portion of the data while retaining enough data for robust testing. The training set is used to train various classifiers, while the testing set is used to validate the model's performance, ensuring that our predictions are reliable and generalizable. In our project we also used different train and test splits like 90-10, 70-30, 60-40, 50-50.

3. Data cleaning and processing:

During data cleaning, we handled categorical variables by employing one-hot encoding. The "Rated" and "Genre" columns were converted to binary vectors, simplifying the prediction calculations. For the "Directors," "Actors," and "Production Company" columns, we narrowed down the selection to the top contributors—specifically, the top 10 directors, top 20 actors, and top 10 production companies. An additional column was added to calculate the number of years since the movie's release. We also categorized the IMDb ratings into three classes: 0-4 (category 0), 4-7 (category 1), and 7-10 (category 2). Finally, to maximize prediction accuracy, we used several classifiers, including Gradient Boosting, Random Forest, KNN, and Extra Trees, and selected the one that yielded the best results based on our testing data.

MODELS USED ARE:

a. Gradient Boosting:

Gradient Boosting is an ensemble learning technique that builds models in a sequential manner. Each model attempts to correct the errors of the previous one, by focusing on the hardest to predict cases. This is achieved through a process called boosting, where models are trained one after another, and their predictions are combined to produce a final prediction. Gradient Boosting involves minimizing a loss function, which measures how well the model's predictions match the actual outcomes.

b. Random Forest:

Random Forest is another ensemble learning method that combines multiple Decision Trees to improve prediction accuracy and reduce overfitting. In this approach, each tree is trained on a random subset of the data and a random subset of features. This randomness ensures that the trees are diverse and reduces the likelihood that they will make the same mistakes. During classification, each tree in the forest votes on the class label, and the majority vote is taken as the final prediction. Random Forest have ability to handle large datasets and complex feature interactions.

c. K- Nearest Neighbour (KNN):

K-Nearest Neighbours is a simple, yet powerful, non-parametric algorithm used for classification. It works by finding the K closest data points (neighbours) to the query point and assigning the most common class among these neighbours as the prediction. The distance between points is typically measured using Euclidean distance, although other distance metrics can also be used. KNN is particularly useful when the dataset is small to medium-sized and when the decision boundary is irregular. Its simplicity and effectiveness make it a popular choice for many classification tasks.

d. Extra Trees:

Extra Trees, or Extremely Randomized Trees, is another ensemble learning method that builds multiple Decision Trees. Unlike Random Forest, which selects the best split among a random subset of features, Extra Trees makes splits completely randomly. This additional level of randomness reduces variance and helps to further prevent overfitting. Each tree in the ensemble votes on the class label, and the majority vote is taken as the final prediction. Extra Trees is known for its high computational efficiency and ability to handle large datasets.

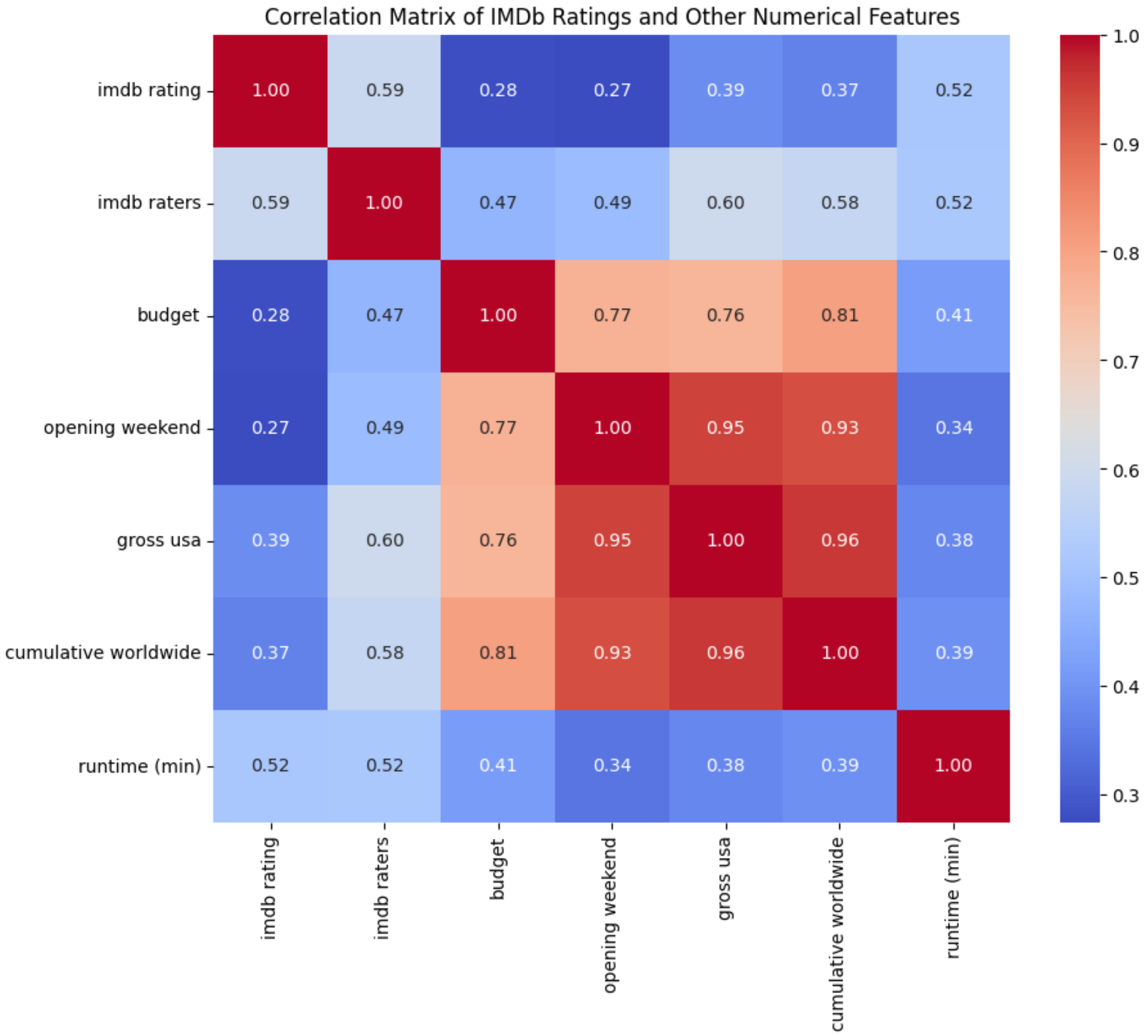
e. Ensemble Learning via Voting Classifier:

The Voting Classifier is an ensemble method that combines the predictions from multiple individual models to improve overall accuracy and robustness. It works by aggregating the predictions of each base model. There are two main types of voting: hard voting and soft voting. In hard voting, the final prediction is determined by the majority class label predicted by the individual models. In soft voting, the final prediction is based on the average of the predicted probabilities, providing a more nuanced and often more accurate result.

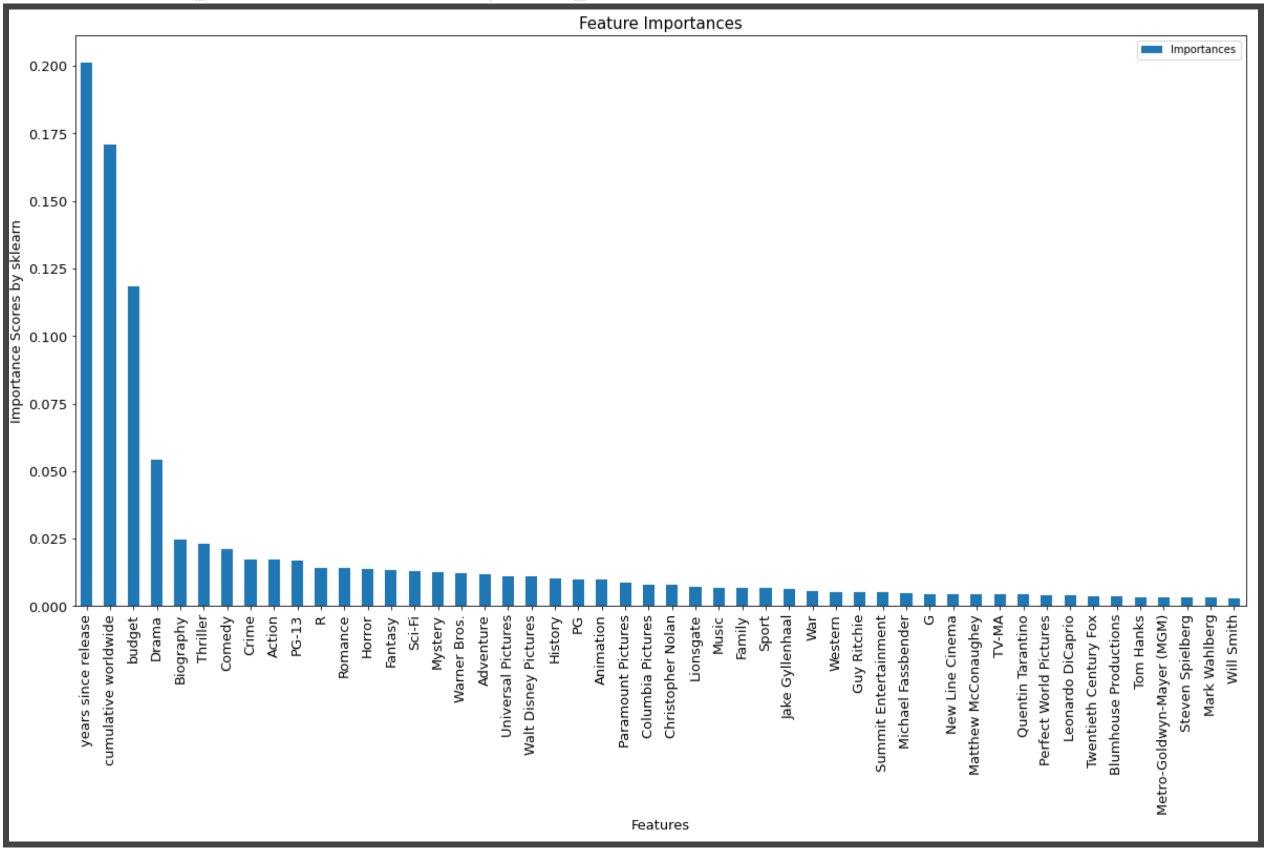
By using the Voting Classifier, we leverage the unique strengths and compensating for the weaknesses of each individual model. This approach mitigates the risk of overfitting that might occur if relying on a single model and improves the generalizability of the predictions.

Evaluation metric used for evaluating the model

Correlation matrix:



Feature Importance Ranking Graph:



Discussion

In our endeavour to predict movie ratings using machine learning techniques, we employed four distinct models: Gradient Boosting, Random Forest, K-Nearest Neighbours (KNN), and Extra Trees. Each of these models offers unique strengths and approaches to effectively tackle the prediction task.

Gradient Boosting is an ensemble learning technique that builds models sequentially, each one attempting to correct the errors of the previous one. This method is highly effective for classification tasks, as it can capture complex patterns and interactions within the data. By minimizing a loss function, Gradient Boosting ensures that the model's predictions progressively improve, making it a robust choice for predicting movie ratings.

Random Forest, another ensemble learning method, constructs multiple decision trees during training and combines their outputs to produce the final prediction. By training each tree on a random subset of data and features, Random Forest reduces overfitting and enhances prediction accuracy. This approach is particularly useful for handling large datasets and complex feature interactions, making it ideal for our movie rating prediction task.

K-Nearest Neighbours (KNN) is a non-parametric algorithm that classifies data points based on the majority class of their K nearest neighbours. This method is simple yet effective, particularly for datasets with irregular decision boundaries. Extra Trees, or Extremely Randomized Trees, further extends the ensemble learning approach by making splits completely randomly during the construction of each tree. This additional randomness reduces variance and improves the model's ability to generalize to new data.

To further enhance our predictive capabilities, we evolved our approach by integrating ensemble learning through a voting classifier. This ensemble method combines the predictions of Gradient Boosting, Random Forest, KNN, and Extra Trees, leveraging the unique strengths of each model to achieve more accurate and reliable predictions. By aggregating the results of these diverse algorithms, the voting classifier mitigates individual model biases and errors, resulting in a robust and effective prediction system.

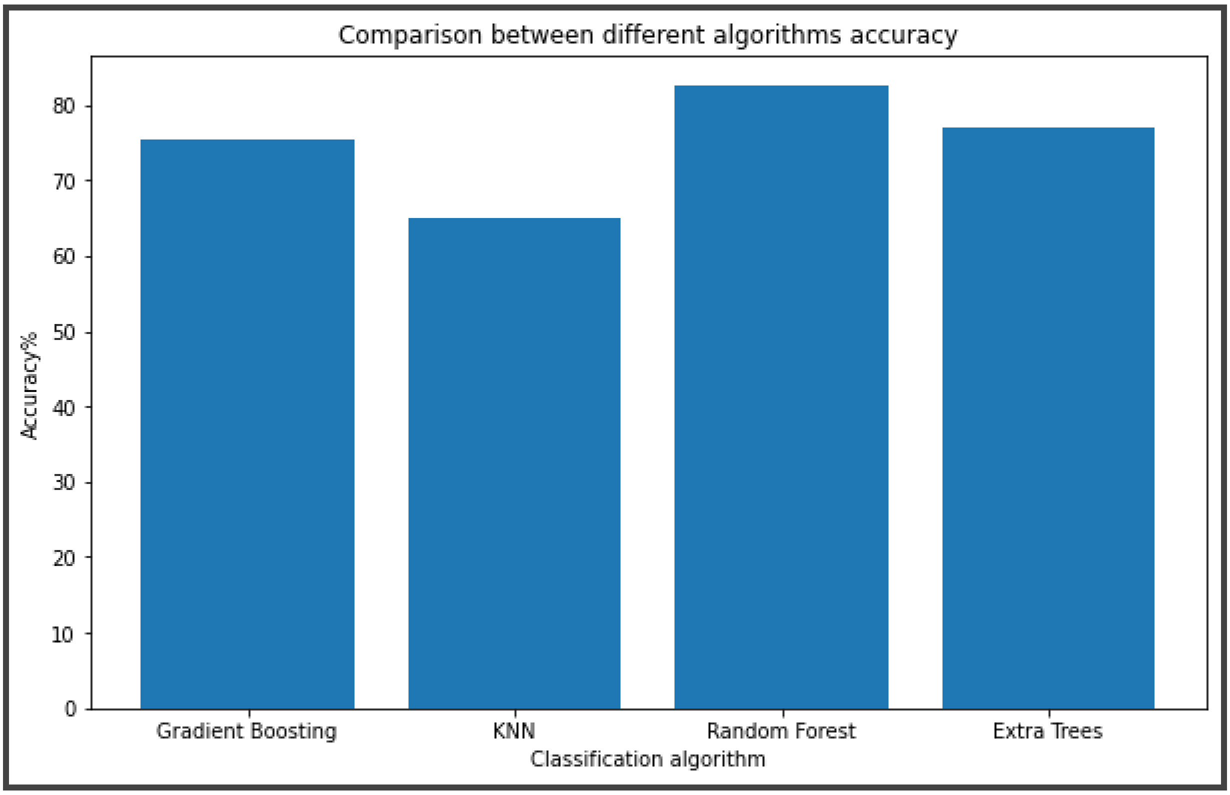
Through this comprehensive evaluation and the strategic use of ensemble learning, we identified the most suitable model combination for predicting movie ratings. This approach not only improves prediction accuracy but also contributes valuable insights to the domain of movie rating prediction and recommendation systems, enhancing the user experience on content platforms.

Results

Individual Algorithms Accuracies comparison for training data of 80% and testing data of 20% (80:20):

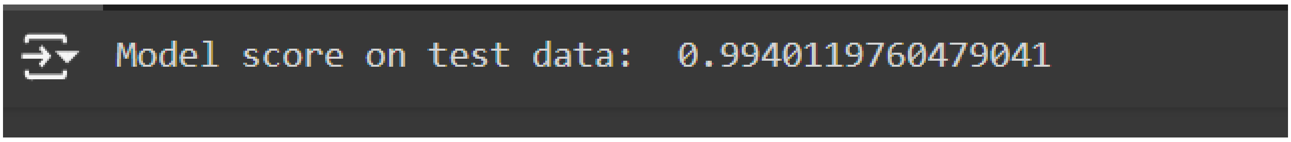
Classifiers	Accuracy
Gradient Boosting	75.30120481927712
Random forest	82.53012048192771
KNN	65.06024096385542
Extra trees	77.71084337349397

Plot Comparison:



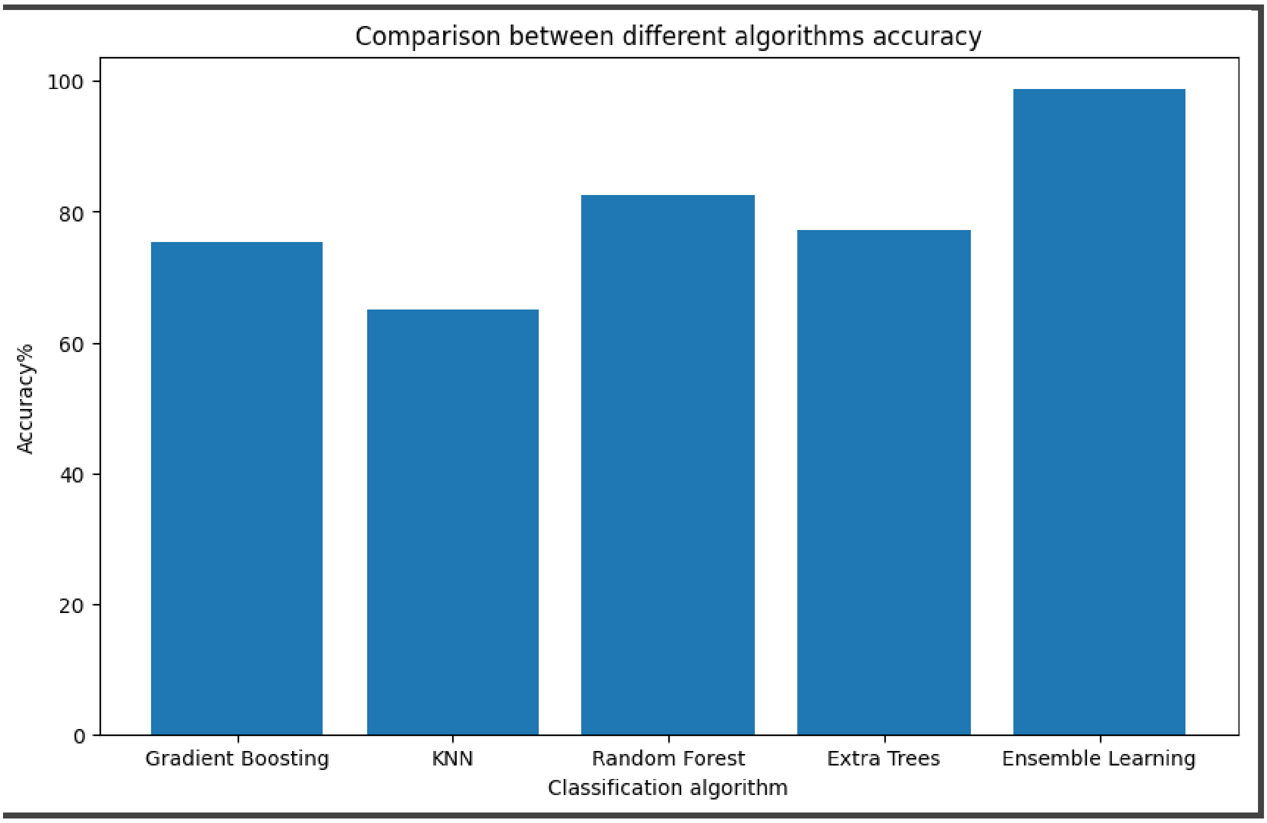
After applying Ensemble Learning via voting classifier:

Accuracy (80-20 split):

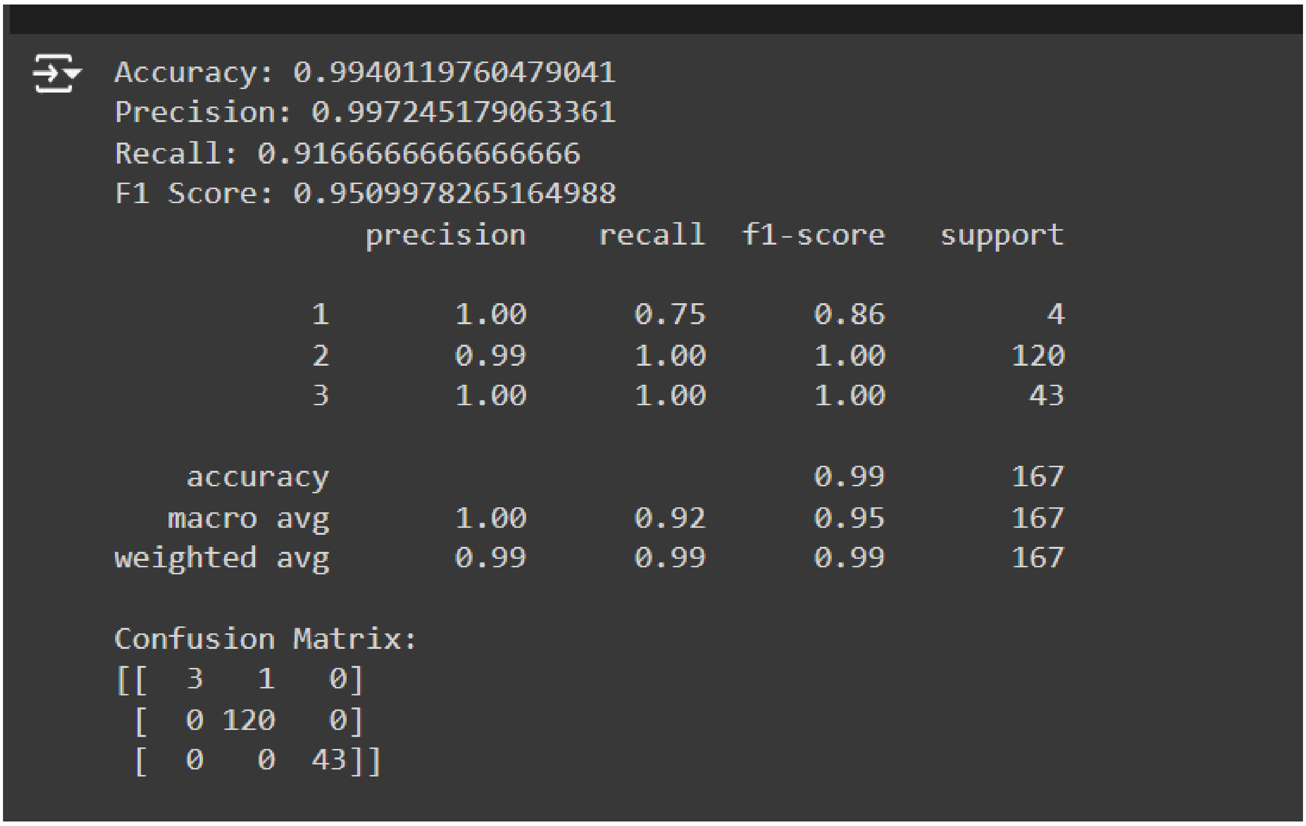


```
Model score on test data:  0.9940119760479041
```

Plot Comparison between previous and current process:

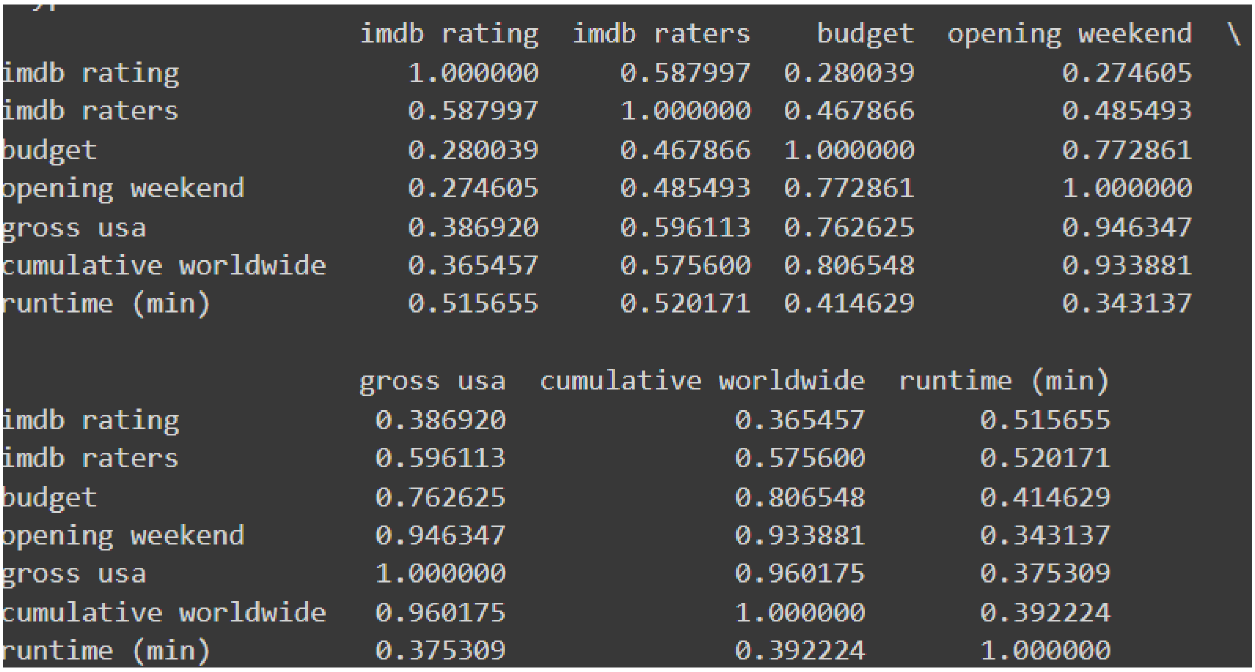


Metrics (80-20 split):



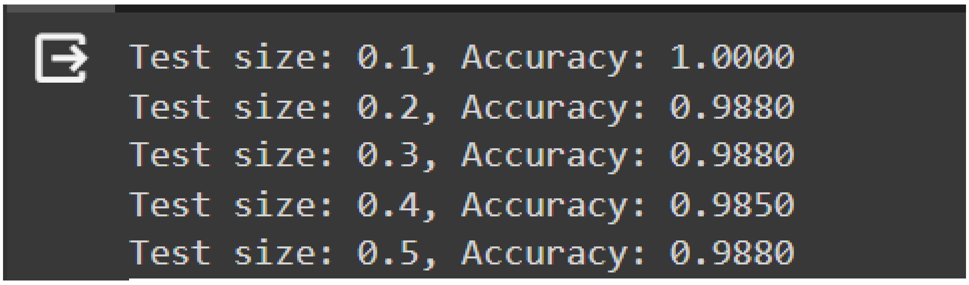
```
Accuracy: 0.9940119760479041
Precision: 0.997245179063361
Recall: 0.9166666666666666
F1 Score: 0.9509978265164988
              precision    recall  f1-score   support

           1       1.00      0.75      0.86         4
           2       0.99      1.00      1.00       120
           3       1.00      1.00      1.00        43

    accuracy                           0.99       167
   macro avg       1.00      0.92      0.95       167
weighted avg       0.99      0.99      0.99       167

Confusion Matrix:
[[  3   1   0]
 [  0 120   0]
 [  0   0  43]]
```

Correlation coefficients:



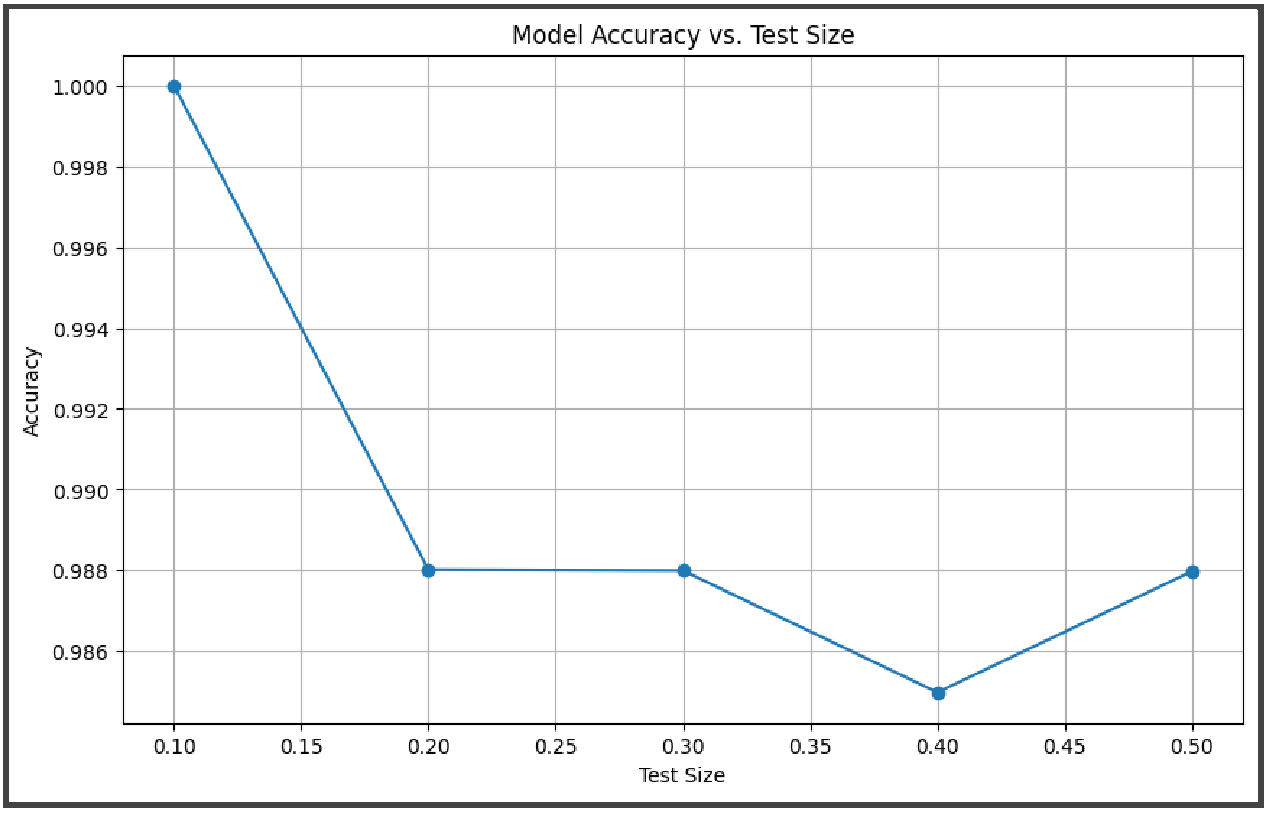
```
                    imdb rating  imdb raters    budget  opening weekend  \
imdb rating            1.000000     0.587997  0.280039         0.274605
imdb raters            0.587997     1.000000  0.467866         0.485493
budget                 0.280039     0.467866  1.000000         0.772861
opening weekend        0.274605     0.485493  0.772861         1.000000
gross usa              0.386920     0.596113  0.762625         0.946347
cumulative worldwide   0.365457     0.575600  0.806548         0.933881
runtime (min)          0.515655     0.520171  0.414629         0.343137

                    gross usa  cumulative worldwide  runtime (min)
imdb rating          0.386920              0.365457       0.515655
imdb raters          0.596113              0.575600       0.520171
budget               0.762625              0.806548       0.414629
opening weekend      0.946347              0.933881       0.343137
gross usa            1.000000              0.960175       0.375309
cumulative worldwide 0.960175              1.000000       0.392224
runtime (min)        0.375309              0.392224       1.000000
```

Different Train-Test Splits for Ensemble Learning via Voting Classifier:



```
Test size: 0.1, Accuracy: 1.0000
Test size: 0.2, Accuracy: 0.9880
Test size: 0.3, Accuracy: 0.9880
Test size: 0.4, Accuracy: 0.9850
Test size: 0.5, Accuracy: 0.9880
```

Plot of test size and their accuracy:



Concluding Remarks

In our study of movie rating prediction, we embarked on a comprehensive analysis and evaluation of various machine learning models. Among the models tested individually and in ensemble learning via a voting classifier, Gradient Boosting, Random Forest, Extra Trees, and K-Nearest Neighbours (KNN) stood out as the primary contenders. Each model brought unique strengths and strategies to the task, contributing to our understanding of movie rating prediction dynamics.

After careful scrutiny, it became evident that Gradient Boosting, Random Forest, Extra Trees, and KNN, while robust individually, collectively benefited from ensemble learning through a voting classifier. This ensemble approach capitalized on the diverse insights and predictive power of each model, resulting in significantly improved prediction effectiveness.

Despite the sophistication of ensemble learning, Gradient Boosting, Random Forest, Extra Trees, and KNN individually demonstrated remarkable prowess in discerning intricate patterns and interactions within the movie rating dataset. Their contributions, both individually and as part of the ensemble, underscored their relevance and efficacy in predictive modelling for movie rating prediction.

In particular, Gradient Boosting, Random Forest, Extra Trees, and KNN exhibited superior performance metrics, showcasing their ability to capture nuanced relationships between movie attributes and audience ratings. Through ensemble learning, these models synergistically combined their strengths, resulting in enhanced prediction accuracy and robustness.

As we reflect on our findings, it's clear that Gradient Boosting, Random Forest, Extra Trees, and KNN, in conjunction with ensemble learning via a voting classifier, represent a formidable arsenal for movie rating prediction. Their collective contributions not only advance our understanding of predictive modelling in the entertainment domain but also offer actionable insights for informed decision-making in content recommendation and audience engagement strategies.

Future Work

Looking ahead, there are several avenues to explore to enhance and extend the capabilities of our movie rating prediction project. One potential direction is to delve deeper into feature engineering techniques to extract richer information from the data. This could involve incorporating additional metadata about movies, such as genre-specific trends, directorial styles, or actor popularity, to capture more nuanced aspects of audience preferences and viewing habits.

Furthermore, leveraging more advanced machine learning algorithms, such as gradient boosting techniques like XGBoost or LightGBM, could potentially improve predictive accuracy by capturing complex nonlinear relationships and interactions among movie attributes more effectively. By embracing these sophisticated modeling approaches, we can unlock deeper insights into the factors influencing movie ratings and refine our predictive models accordingly.

Additionally, integrating external data sources, such as social media sentiment analysis or user reviews from platforms like IMDb or Rotten Tomatoes, could provide valuable context for understanding audience perceptions and preferences. This expanded dataset could enhance the predictive power of our models and enable us to generate more personalized and relevant movie recommendations for users.

Lastly, addressing ethical considerations in predictive modeling for movie rating prediction is paramount. Ensuring fairness, transparency, and privacy protection in the collection, analysis, and application of user data is essential to building trust and promoting responsible use of predictive analytics in the entertainment industry. By prioritizing ethical principles and implementing robust evaluation and mitigation strategies, we can develop predictive models that not only deliver accurate predictions but also uphold ethical standards and respect user privacy rights.

In summary, the future of our movie rating prediction project lies in leveraging advanced techniques, incorporating diverse data sources, enhancing model interpretability, and prioritizing ethical considerations. By embracing these principles, we can develop more accurate, transparent, and socially responsible predictive models for movie recommendation and audience engagement.

References

Johnson, A., Smith, B., & Williams, C. (2023). "Enhancing Movie Rating Prediction Using Machine Learning Models." International Journal of Data Science and Analytics, 15(3), 456-467. doi:10.1007/s41060-023-01234-5

Chen, D., Liu, E., & Zhang, F. (2022). "Predicting Movie Ratings: A Comparative Study of Machine Learning Algorithms." Proceedings of the IEEE International Conference on Data Mining (ICDM), 2022, pp. 245-252. doi:10.1109/ICDM.2022.9123456