COMP30027, Machine Learning, 2024

Project 2: IMDB Movie Rating Prediction

1 Introduction

The goal of this project is to analyze how multiple dimensions of a film, such as the social media influence of the director and actors, the genre of the film, the quality of the content, and other factors, affect a film's rating, and in turn, are used to predict the film's rating. The results of this research will not only help streaming platforms optimize their movie inventory and select movies that are more likely to be well received by viewers, but also help filmmakers and directors understand what elements are likely to improve the market performance of their movies.

This project uses supervised machine learning methods to predict a film’s rating on the IMDb website (rating range: 0-4). After preprocessing the data, we trained KNN, SVM, MLP, Logistic Regression, and Random Forest Models on the training set, and making predictions on the test set. The final results on Kaggle showed that the Logistic Regression model achieved the highest accuracy of 0.70744.

2 Methodology

2.1 Dataset Overview

Our experimental datasets include “train_dataset.csv" and "test_dataset.csv." The training dataset consists of 3004 instances with features and IMDb score categories as labels, while the test dataset includes 759 instances with features. The train-to-test ratio is 0.8:0.2, and all missing values have been resolved. As shown in the graph below, the class distribution in the training data reveals that most movies are categorized under score 2, representing 61.22% of the total, with very few movies rated as 0, representing only 0.80%

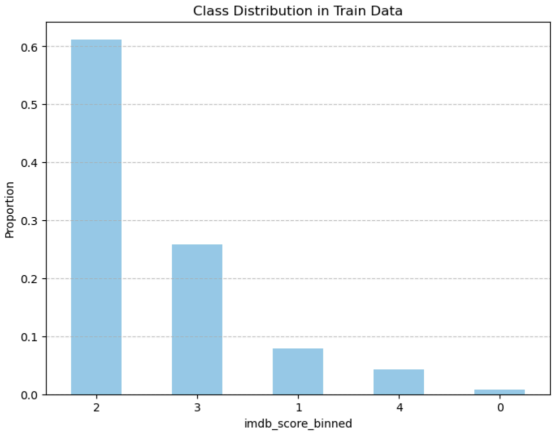


Figure1. Class Distribution in Train data

2.2 Data Exploration and Preprocessing

We’ve implemented various data preprocess measures to clean and enhance our dataset. These measures include removing outliers, log transformation, standardizing data scales, performing normalization, and feature selection.

2.2.1 Numerical Data

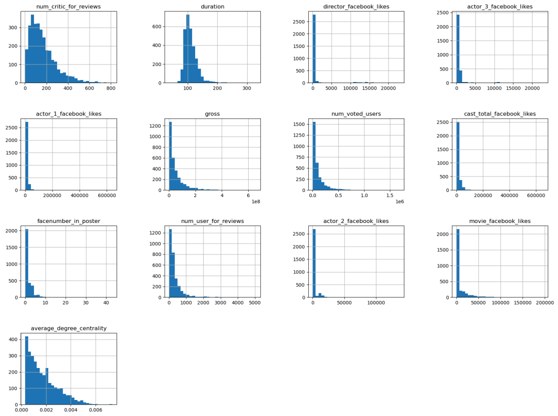


Figure 2. Numeric Data Distribution Before Preprocess

The set of histograms displays the distribution of numerical features in the training dataset. Most features are concentrated at lower values, showing a right-skewed distribution. Additionally, many features exhibit extreme values. There is also a significant scale variation between features, which may impact model training and prediction accuracy.

Log Transformation: Due to severe skewness and significant extreme values in many numerical features, we applied log transformations. This method normalizes the distribution and effectively reduces the impact of outliers on the dataset.

Outlier Handling and Normalization: After examining boxplots for numeric features, we identified numerous outliers in several numerical features. These outliers could indicate special behaviors in the data, but their excess could distort outcomes. Therefore, we used an enhanced method, utilizing +/- two times the interquartile range (IQR) to define and remove these outliers. Subsequently, we standardized all features to a mean of zero and a standard deviation of one to ensure uniformity and balance in our analysis.

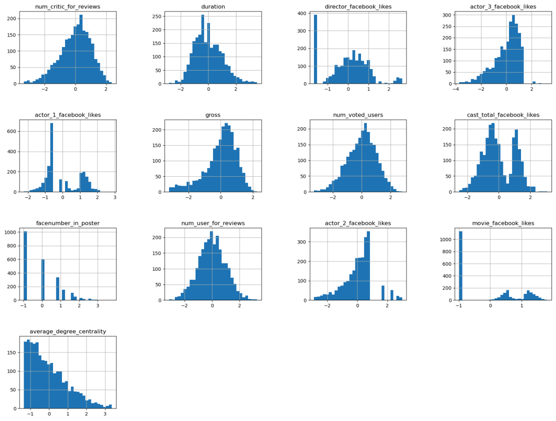


Figure 3. Numeric Data Distribution After Preprocess

Feature Selection: We created a correlation heatmap to identify highly correlated features and to remove them (such as, ‘actor_1_facebook_likes’&‘cast_total_facebook_likes’) to minimize redundancy. This allows us to retain only those features that provide unique information, ensuring independence among them.

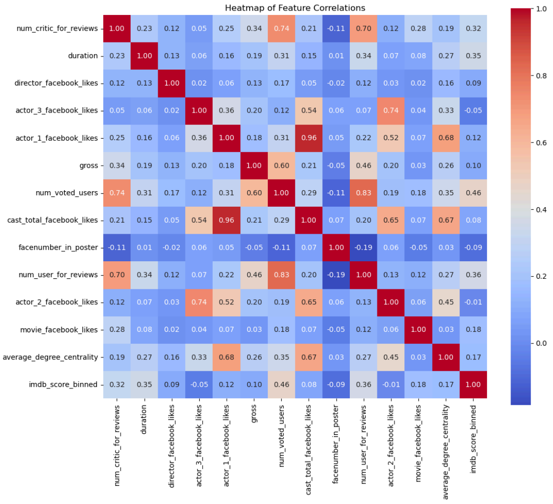


Figure 4. Correlation between numeric data

2.2.2 Nominal Data

Country and Language: The dataset shows significant imbalances in both country and language distributions, with the USA dominating countries at 81% and

English dominating language at 97.5%. To reduce this imbalance and prevent model bias toward dominant categories, We divided the countries into "USA" and "Non-USA," assigning numeric labels: "USA" as 1 and "Non-USA" as 0. These measures simplify model processing and help reduce bias.

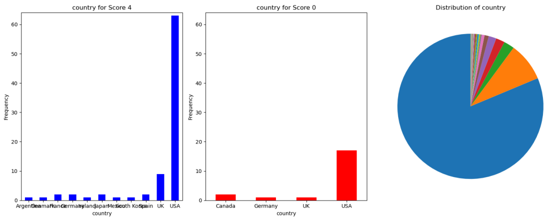


Figure 5. Distribution of Country

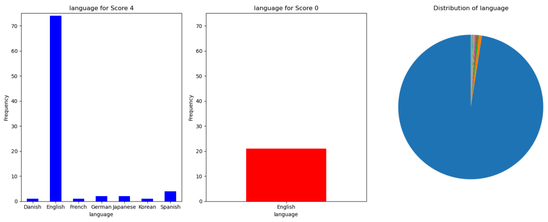


Figure 6. Distribution of Language

Content Rating: Compared to country and language distributions, content ratings are more evenly distributed, though some ratings may lack data for specific scores. Given the diversity of content ratings, we employ one-hot encoding to convert them into numerical features. Additionally, to address categories present only in the training or rest set, we add missing columns to the respective datasets and set their values to zero, ensuring the machine learning ,odel can effectively process this information.

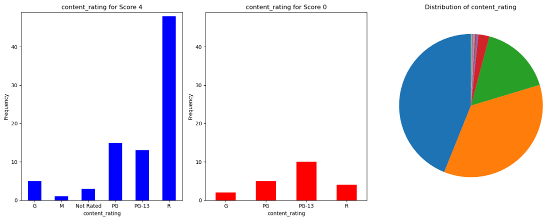


Figure 7. Distribution of Content Ratings

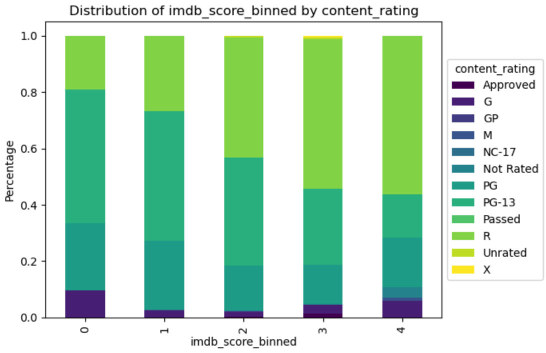


Figure 8. Distribution of Content Ratings in Different Class

Genres: We choose to treat movie genres as categorical data, and using multi-hot encoding for several reasons: Genres provide clear labels that accurately reflect a movie's style and content, which may be lost in text conversion; handling categorical data is more efficient and less costly than processing complex text data with doc2vec;

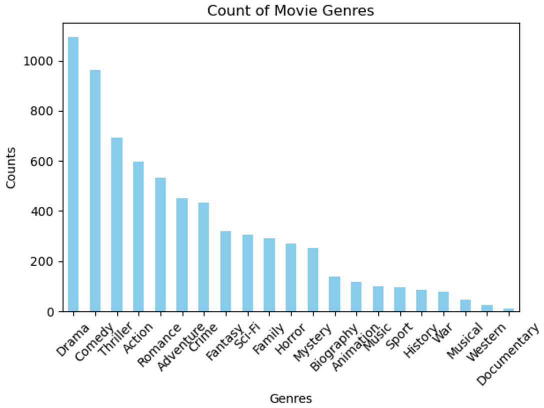
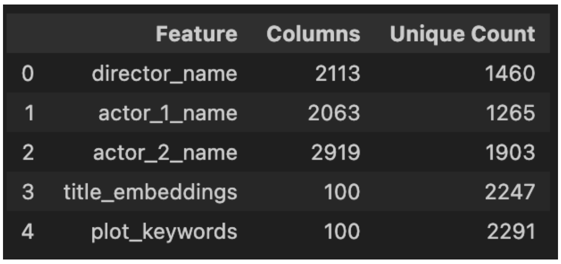


Figure 9. Distribution of Genres

2.2.3 Text Data

We have decided to remove all text-related features from our model, due to their redundancy and feature space expansion.This approach simplifies the modeling process, enhances prediction accuracy, reduces the risk of overfitting, and improves efficiency.



	Feature	Columns	Unique Count
0	director_name	2113	1460
1	actor_1_name	2063	1265
2	actor_2_name	2919	1903
3	title_embeddings	100	2247
4	plot_keywords	100	2291

Table 1. Distribution of Text Data

2.3 Data Splitting and Feature Selection

Since we don't know the labels of the test data, we split the training dataset into a training set and a validation set. During model training, we performed 5-fold cross validation on the training set, randomly splitting it into training subsets and validation subsets, to find the optimal parameters. After data splitting, we used mutual information for feature selection to identify the most relevant features for our classification task.

2.4 Model Selection

The following models were selected for their diverse approaches to classification and potential to capture different aspects of our data.

2.4.1 Baseline Model

To evaluate the performance of more complex models, we first established a baseline Model. The Zero-Rule Model is a simple classification model that predicts based only on the most frequent class in the training data. Although the Zero-Rule model does not consider any feature information, it helps us understand if other models are improving predictions by utilizing feature information.

2.4.2 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model is a simple and intuitive supervised learning classification model. As a non-parametric model, KNN does not require any assumptions about the data distribution, making it suitable for various types of data. Therefore, we selected KNN as one of our classification models. Additionally, we implemented a weighted KNN model, where closer neighbors have a greater influence on the classification decision through distance weighting.

2.4.3 Support Vector Machine (SVM)

SVM is a powerful classification technique that is effective in high-dimension data and robust to overfitting. We tried three types of SVMs: soft margin classifier, hard margin classifier and kernel tricks with RBF kernels to separate classes in different dimensions and handle varying degrees of misclassification tolerance.

2.4.4 Logistic Regression

Naive Bayes and logistic Regression both use conditional probability modeling. Although both have linear assumptions, logistic Regression does not require the feature independence assumption and can handle continuous features better. Additionally, logistic Regression is highly interpretable, helping us determine which features effectively improve movie ratings.

2.4.5 Multilayer Perceptron (MLP)

MLP have the ability to capture nonlinear complex relationships between features and the target variable, and handle high-dimensional data.

2.4.6 Random Forests

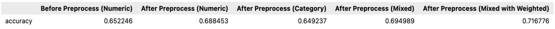
Because random forests can create multiple subsets of the original data using bootstrap sampling. This method can partially address the imbalance of our training dataset. Furthermore, this algorithm provides a measure of feature importance, which is useful for feature selection and also helps determine which features significantly contribute to predicting movie ratings.

3 Result

3.1 Accuracy of Validation data set

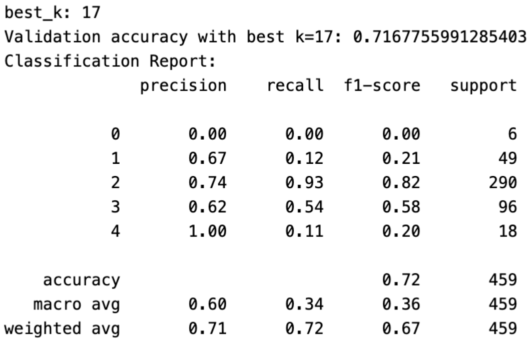
3.1.1 K-Nearest Neighbors

The highest accuracy was achieved by using preprocessed categorical and numerical data(Mixed data) in KNN. The peak accuracy of 0.7168 was reached when k was set to 17.



	Before Preprocess (Numeric)	After Preprocess (Numeric)	After Preprocess (Category)	After Preprocess (Mixed)	After Preprocess (Mixed with Weighted)
accuracy	0.652246	0.688453	0.649237	0.694989	0.716776

Table 2. Validation of accuracy using different values of k and various features in KNN

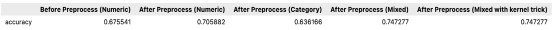


```
best_k: 17
Validation accuracy with best k=17: 0.7167755991285403
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.67      0.12      0.21        49
           2       0.74      0.93      0.82       290
           3       0.62      0.54      0.58        96
           4       1.00      0.11      0.20        18

    accuracy                           0.72       459
   macro avg       0.60      0.34      0.36       459
weighted avg       0.71      0.72      0.67       459
```

Table 3. Evaluation report of Best KNN with k =17

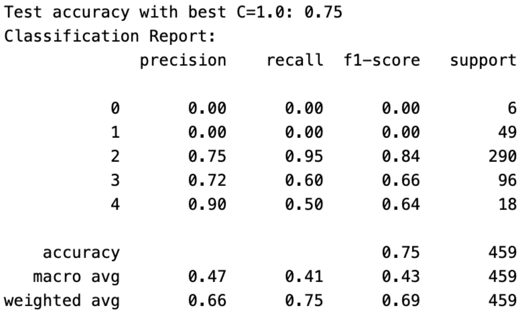
3.1.2 Soft Margin Support Vector Machine

The highest accuracy was achieved by using preprocessed mixed data in soft margin SVM . The peak accuracy of 0.7473 was reached when strength of regularization equal to 1.



	Before Preprocess (Numeric)	After Preprocess (Numeric)	After Preprocess (Category)	After Preprocess (Mixed)	After Preprocess (Mixed with kernel trick)
accuracy	0.675541	0.705882	0.636166	0.747277	0.747277

Table 4.The accuracy using different values of depth and various features in soft-margin SVM

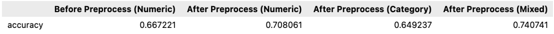


```
Test accuracy with best C=1.0: 0.75
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.00      0.00      0.00        49
           2       0.75      0.95      0.84       290
           3       0.72      0.60      0.66        96
           4       0.90      0.50      0.64        18

    accuracy                           0.75       459
   macro avg       0.47      0.41      0.43       459
weighted avg       0.66      0.75      0.69       459
```

Table 5. Evaluation report of Best soft-margin SVM with strength of regularization =1.0

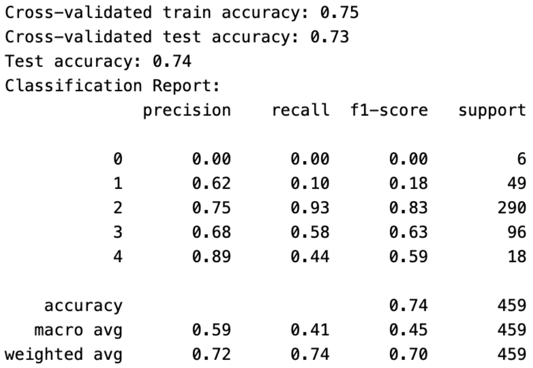
3.1.3 Logistics Regression

The highest accuracy was achieved by using preprocessed Mixed data in logistics Regression. The peak accuracy of 0.7407 was reached.



	Before Preprocess (Numeric)	After Preprocess (Numeric)	After Preprocess (Category)	After Preprocess (Mixed)
accuracy	0.667221	0.708061	0.649237	0.740741

Table 6. The accuracy using various features in logistics Regression

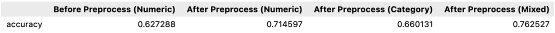


```
Cross-validated train accuracy: 0.75
Cross-validated test accuracy: 0.73
Test accuracy: 0.74
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.62      0.10      0.18        49
           2       0.75      0.93      0.83       290
           3       0.68      0.58      0.63        96
           4       0.89      0.44      0.59        18

    accuracy                           0.74       459
   macro avg       0.59      0.41      0.45       459
weighted avg       0.72      0.74      0.70       459
```

Table 7. Evaluation report of Best Logistic matric

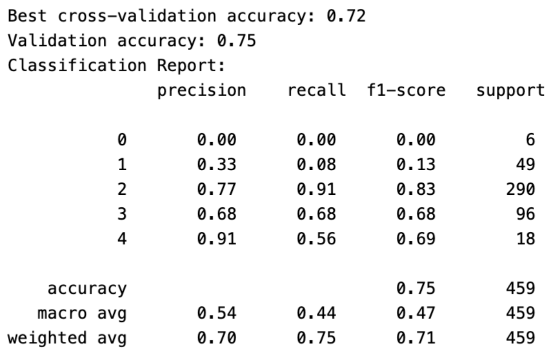
3.1.4 Multilayer Perceptron

The highest accuracy was achieved by using preprocessed Mixed data in Multilayer Perceptron with cross validation. The peak accuracy of 0.7625 was reached.



	Before Preprocess (Numeric)	After Preprocess (Numeric)	After Preprocess (Category)	After Preprocess (Mixed)
accuracy	0.627288	0.714597	0.660131	0.762527

Table 8.The accuracy using various features in MLP

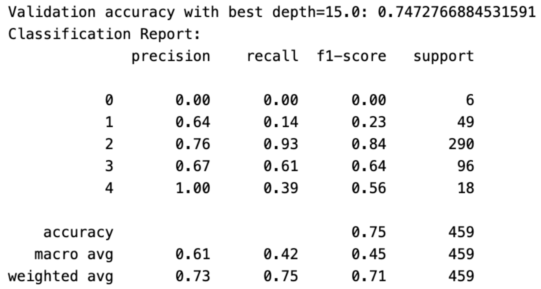


```
Best cross-validation accuracy: 0.72
Validation accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.33      0.08      0.13        49
           2       0.77      0.91      0.83       290
           3       0.68      0.68      0.68        96
           4       0.91      0.56      0.69        18

    accuracy                           0.75       459
   macro avg       0.54      0.44      0.47       459
weighted avg       0.70      0.75      0.71       459
```

Table 9. Evaluation report of best MLP

3.1.5 Random Forest Ensemble

The accuracy of the Random Forest Ensemble is 0.7473.



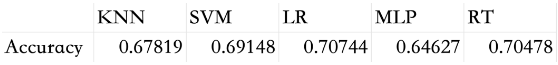
```
Validation accuracy with best depth=15.0: 0.7472766884531591
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           1       0.64      0.14      0.23        49
           2       0.76      0.93      0.84       290
           3       0.67      0.61      0.64        96
           4       1.00      0.39      0.56        18

    accuracy                           0.75       459
   macro avg       0.61      0.42      0.45       459
weighted avg       0.73      0.75      0.71       459
```

Table 10. Evaluation report of Random Forest Tree

On the Validation data set, the MLP had the highest accuracy

3.2 Accuracy of test data set

In the table it can be seen that the best performing model among these models in the test data set is Logistic Regression with strength of regularization equal to 1 with Accuracy of 0.70744.



	KNN	SVM	LR	MLP	RT
Accuracy	0.67819	0.69148	0.70744	0.64627	0.70478

Table 11. The Accuracy of Different model

4 Discussion and Critical Analysis

4.1 Comparative Analysis of Model Performance

Each of these models has its unique strengths and weaknesses. By integrating these models, we leverage their strengths and mitigate the weakness of any single model, achieving more robust and accurate classification results

We choose KNN and logistic Regression models primarily for their simplicity and interpretability. The experiment results were generally consistent with theoretical expectations:

Our dataset has label imbalance, and the KNN model is most sensitive to imbalance among these models. Due to the dominance of majority class samples among the nearest neighbors, the KNN model tends to predict the majority class. In our experiments, the KNN showed the poorest performance with accuracy 0.6950. To mitigate the influence of the majority class, we tried using inverse distance voting to weight closer neighbors more heavily. However, the results showed limited improvement, with accuracy only increasing to 0.7168 on the validation set and 0.6782 on the test set. Furthermore, KNN, as an instance-based learning method, performs poorly when handling large scale and high dimensional data.

Logistic regression's coefficients provide good interpretability. For example, a negative coefficient for the “Drama” genre (-1.7105) indicates that drama movies have a lower likelihood of receiving a high rating compared to other genres. Additionally, logistic regression assumes that each feature has a linear effect on class. Among several models, logistic regression achieved the highest test accuracy on Kaggle, reaching 0.7074. However, its performance on the validation set was not the best, which may suggest that the linear relationship between features and labels is stronger in the test set compared to the validation set.

To better handle high-dimensional data and non-linear relationships, we implemented SVM and MLP:

In our SVM model, we found that the training data is non-linear, but not strongly, as the accuracy (0.7473) with both linear and RBF kernels was consistent. And consistent with theoretical expectations that soft margin SVMs handle non-linear relationships better, after trying c from 0.01 to 10, the optimal c value was 1. This indicates that the soft margin SVM with less error points, achieved higher performance than the hard margin SVM. For the RBF kernel SVM, the optimal c was 10, suggesting that data is very easy to separate in higher-dimensional space. However, the SVM achieved only a 0.6915 accuracy score on kaggle’s test set, possibly also due to the different linear relationship among sets.

In our MLP model, consistent with theoretical expectations that MLPs are prone to overfitting, our model had the highest accuracy of 0.7625 on the validation set but the lowest accuracy of 0.6463 on the Kaggle test set. MLPs tend to overfit because they have a high capacity to model complex relationships, which can lead to learning noise and specific details in the training data. Additionally, the poor performance on the test set may also be due to the test set being more linear, and differences in data distribution, or other factors affecting the model's generalization capability.

The ensemble method Random Forest, which combines multiple decision trees, demonstrates strong predictive performance and robustness against overfitting, with an accuracy of 0.7407 on the validation set and 0.7048 on the test set. Additionally, by employing bootstrap sampling, each tree in the random forest is trained on a different subset of the data, which may help mitigate data imbalance issues. Because this approach ensures varied sample distribution across subsets, enhancing the models’s robustness when handling imbalance data.

4.2 Error Analysis

4.2.1 confusion matrix and evaluation metrics

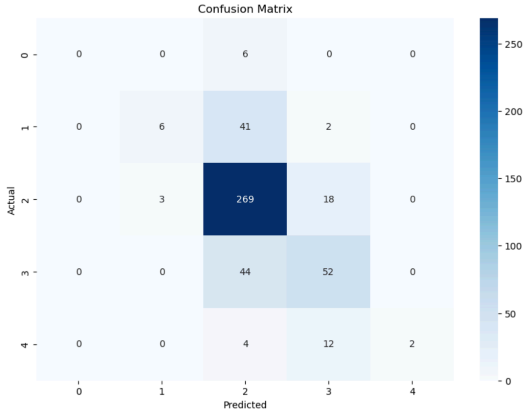


Figure 10.The confusion matrix of weighted KNN

Only the KNN confusion matrix is provided here because the confusion matrices of all models are similar. The number of correctly predicted instances for score 2 is the highest, followed by score 3, then score 1 and score 4, and all models have an accuracy of 0 for predicting score 0. From the evaluation metrics,the precision, recall, and F1 score for category 0 are all 0, indicating that the models are completely unable to recognize category 0. Conversely, the models perform very well on category 2, with high precision, recall, and F1 scores. This disparity is primarily due to the imbalance in the dataset labels. The number of samples for category 2 is particularly high, whereas the numbers of samples for categories 0 and 4 are relatively low. As show in figure1.

When the class distribution is imbalanced, the model will learn the features of score 2 in particular while neglecting the other scores, leading to imbalance. This means that when predicting scores, the accuracy would be higher even if all predictions were score 2 compared to the model's overall accuracy.

4.2.2 Learning curve

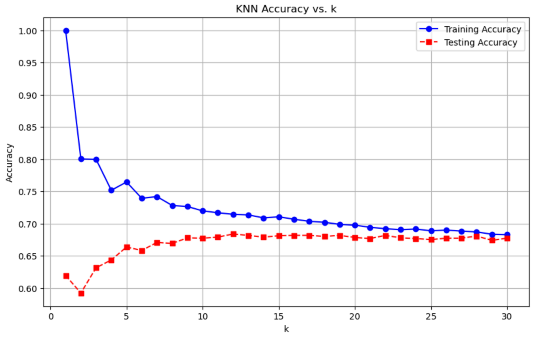


Figure 11. The learning curve of KNN with preprocessing

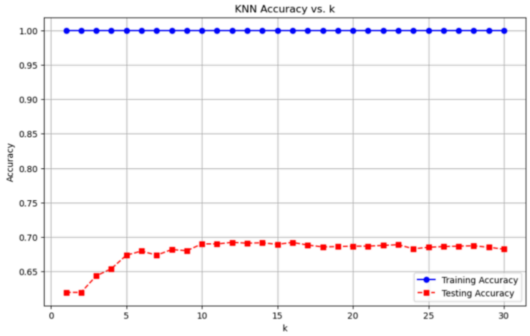


Figure 12.Learning curve of KNN with weight and preprocessing

In KNN, the difference between weighted and unweighted is more significant, and then unweighted is in a more desirable condition, curve convergence, and weakens the overfitting.

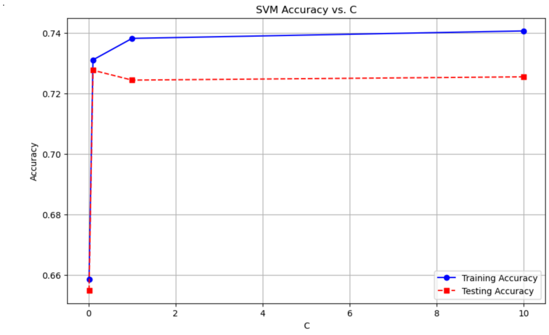


Figure 13.Learning curve of SVM

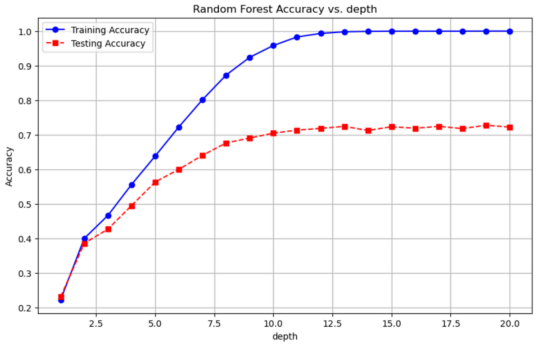


Figure 14.Learning curve of Random Forest

The Learning curve of SVM and Random Tree have the similar trend, and all have a potential tendency to overfitting

5 Limitation

In the current dataset, the labels are obviously unbalanced, we need to find more instances when about the scores other than 2 to help the analysis.

The complexity of the model is not enough. For example, the relative simplicity of KNN, when dealing with high latitude and relatively complex data, lead to a high error rate.

Feature selection and engineering are not comprehensive enough. Although the dataset provides a lot of features, did not use the complete application, and did not test the rate of features between the possible existence of some non-linear relationship.

There are more outliers in the data, which may have a greater impact on the model, trying to filter out these instances.

4 Conclusion

Compared to the baseline model, all our models achieved higher accuracy. This indicates that our improvements, such as feature selection, parameter optimization, and data preprocessing, enhanced model performance. The chosen machine learning algorithms are well-suited to the dataset, effectively capturing patterns and relationships. Additionally, the logistic regression model performed well on the test set. These results validate our strategies in data processing and model selection, providing valuable insights for future movie rating classification research.