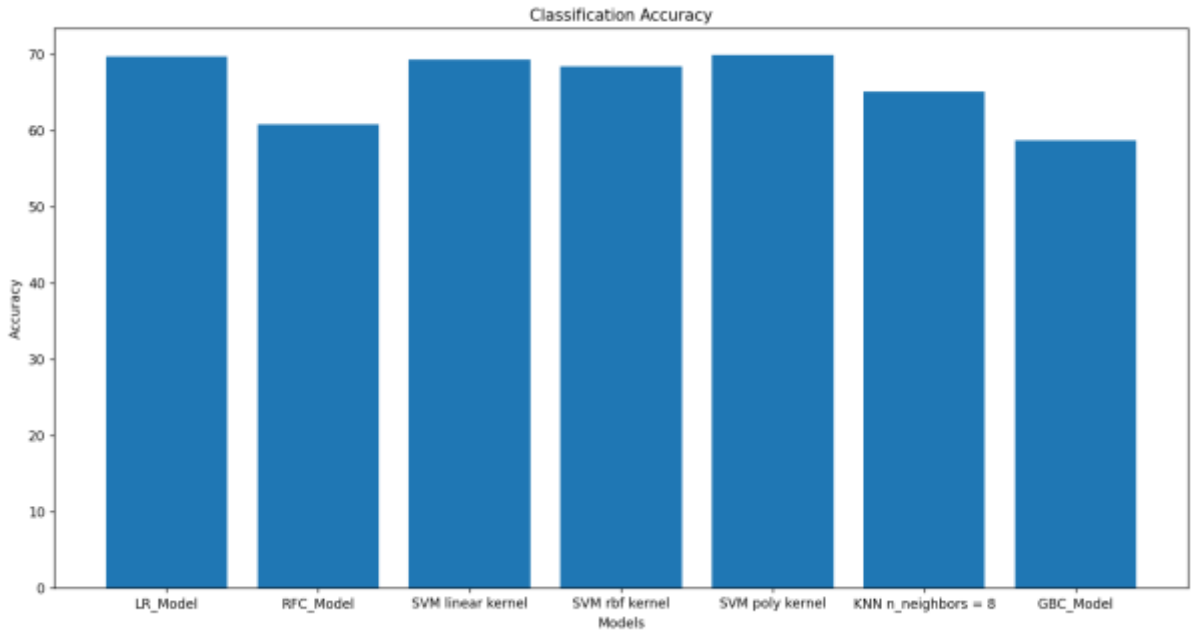
Movie Popularity Prediction

Milestone 2 Report

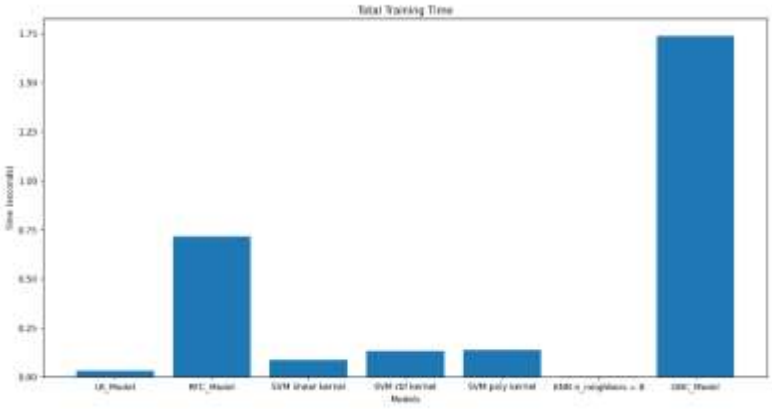
---

1. The classification accuracy is under range from 60% to 70 % whenever we change the type of classification or we change the hyperparameter of any classification and this is evident in his bar graph.



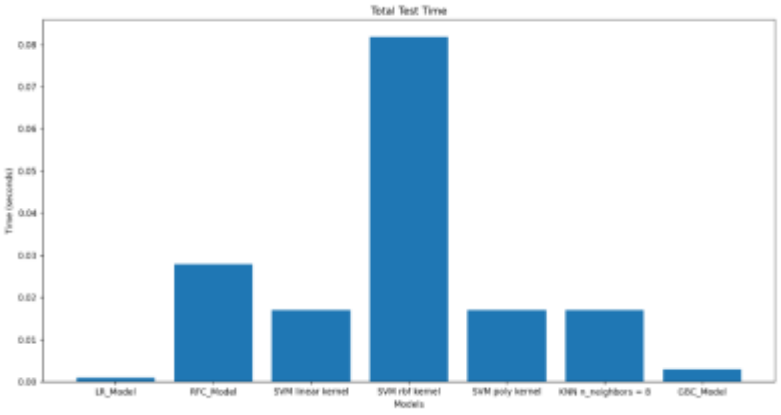
2. We used the time before and after each fit to calculate The total training time , We notice that it changes significantly with each classification

   and It changes from classifier to classifier and also changes with the change of the hyperparameters of each classifier.



3. We used the time before and after each predict to calculate The total testing time , We notice that it changes significantly with each classification

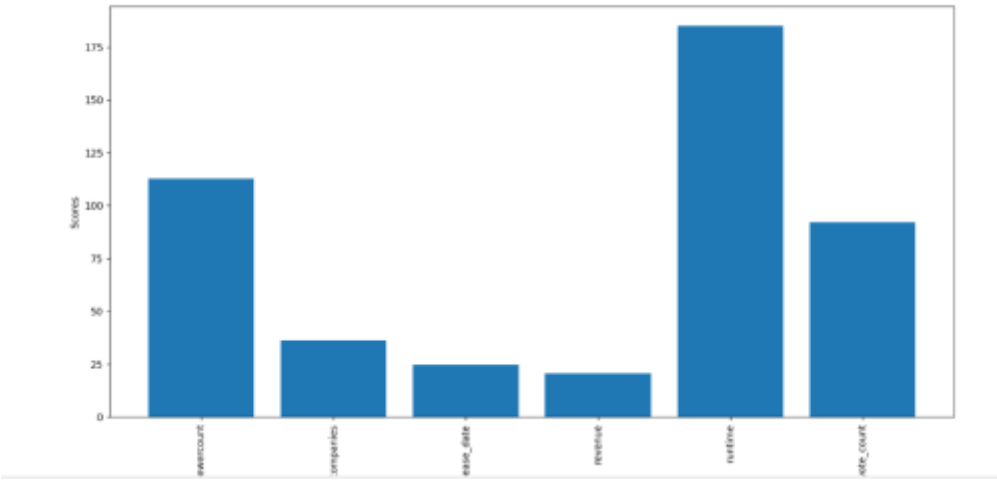
   and It changes from classifier to classifier and also changes with the change of the hyperparameters of each classifier.



4. In the classification phase, the feature selection process may differ from the previous regression phase.

In regression, the feature selection process involved using correlation analysis to identify relevant features. However, in classification, a different approach was adapted.

For classification, the feature selection process utilized the K-best feature selection method, This technique selects the K best features based on statistical significance and ability to contribute to the classification task. The specific criteria used to determine the best features may vary, but common approaches include chi-squared test, mutual information, or ANOVA F-value.
By adopting the K-best feature selection approach in the classification phase, the feature set was refined, focusing on the most informative and discriminative features. This process aimed to enhance the model's ability to accurately classify instances and improve overall classification performance.



5. Hyperparameter tuning plays a crucial role in optimizing the performance of machine learning models. In the case of classification models like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), tuning specific hyperparameters such as 'C' (regularization parameter) and 'kernel' for SVM, and 'n_neighbors' for KNN can have a significant impact on the models' performance.
    a. SVM:
        1- C Hyperparameter: The C hyperparameter controls the trade-off between maximizing the margin and minimizing the classification error. Higher values of C result in a narrower margin and can lead to overfitting, while lower values encourage a wider margin but may result in underfitting.
        2- Kernel Hyperparameter: The kernel determines the type of decision boundary created by the SVM model. Common kernel functions include linear, polynomial, and radial basis function (RBF). Choosing the appropriate kernel can greatly affect the model's ability to capture complex patterns in the data.
    b. KNN:

        1-n_neighbors Hyperparameter: The n_neighbors hyperparameter defines the number of neighbors considered for classifying a new data point. A higher value of n_neighbors can smooth out the decision boundary, reducing the model's

sensitivity to noisy data but potentially oversimplifying the classification. On the other hand, a lower value of n_neighbors can lead to a more complex decision boundary but may be prone to overfitting.

Overall, by tuning the hyperparameters of classification models such as SVM and KNN, we aim to find the optimal configuration that maximizes the models' performance on the given task. The tuning process involves selecting appropriate ranges for the hyperparameters, exploring different combinations, and evaluating the models' performance using suitable evaluation metrics. Hyperparameter tuning can significantly improve the models' performance by finding the best trade-off between underfitting and overfitting and adapting the models to the specific characteristics of the dataset.

6. Conclusion

- In this phase of the project, the same preprocessing techniques used in the previous phase were applied. The K-best feature selection method was employed and proved to be effective in selecting the most relevant features for the classification task.
- Multiple classification models, namely Logistic Regression, SVM, KNN, Random Forest, and Gradient Boosting, were

evaluated, and their performances were compared. The intuition behind this approach was to explore different models with varying characteristics and capabilities, expecting that some models would outperform others in terms of accuracy.

- The hyperparameter tuning process was also conducted, aiming to optimize the performance of each classification model. By adjusting the hyperparameters, such as regularization parameters for SVM, the number of neighbors for KNN, respectively, the models' performance could be further improved.
- Through experimentation and evaluation, conclusions can be drawn about the effectiveness of each classification model. The accuracy of each model was monitored and compared, and it was observed how the hyperparameter tuning influenced the models' performance.