"Performance Comparison of Random Forest, KNN, and SVM on Titanic Dataset Using Various Balancing Techniques and Feature Selection"

ABSTRACT

This study evaluates the performance of three machine learning models—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—on the Titanic dataset. After cleaning the data, the models were trained and tested on unbalanced data and data balanced using under-sampling and over-sampling techniques. Additionally, feature selection methods were applied to optimize model performance. The models were assessed and compared based on accuracy and F1 score, providing insights into the impact of data balancing and feature selection on classification outcomes. The results highlight the strengths and limitations of each model under varying preprocessing conditions.

INTRODUCTION

The Titanic dataset is a well-known benchmark for machine learning, used to predict passenger survival based on factors such as age, gender, and socio-economic status. In this project, we evaluated the performance of three classification models—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). To address challenges like class imbalance, we applied various preprocessing techniques, including data balancing (over-sampling and under-sampling) and feature selection, alongside evaluating the models on unbalanced data. Model performance was assessed using accuracy and F1 score, providing a comprehensive analysis of the impact of different preprocessing strategies on classification results. This study highlights the importance of data preparation and model selection in achieving optimal performance for imbalanced datasets.

METHODS AND MATERIALS

For this project, we utilized Python as the primary programming language and employed Jupyter Notebook and VS Code as development environments to implement and evaluate the machine learning models. The Titanic dataset was preprocessed to handle missing values, encode categorical variables, and scale numerical features. The dataset was then split into training and testing sets to ensure unbiased evaluation. Using Python libraries such as scikit-learn, we trained the Random Forest, SVM, and KNN models under different preprocessing conditions, including unbalanced data, balanced data (using over-sampling and under-sampling techniques), and feature selection. The results were obtained by evaluating the models on the test set, focusing on accuracy and F1 score as performance metrics.

RESULTS

The performance of the three machine learning models—Random Forest, SVM, and KNN—was evaluated using different data preprocessing techniques, with accuracy as the primary metric for comparison.

Unbalanced Data:

- Random Forest achieved the highest accuracy of 81.6%, demonstrating its robustness in handling imbalanced data.
- SVM and KNN performed significantly lower, with accuracies of 65.4% and 64.2%, respectively.

Balanced Data (Over-Sampling):

- Random Forest maintained strong performance with an accuracy of 81.0%, slightly lower than on unbalanced data.
- SVM and KNN showed minor reductions in accuracy, achieving 65.0% and 63.0%, respectively.

Balanced Data (Under-Sampling):

- Random Forest's accuracy dropped to 74.8%, reflecting the learning challenges from reduced data.
- SVM and KNN achieved accuracies of 64.8% and 60.8%, respectively, showing slight declines compared to their performance on unbalanced data.

Feature Selection (Called Data):

- Random Forest demonstrated its adaptability, achieving the highest accuracy of 82.6%.
- SVM and KNN showed significant decreases in performance, with accuracies of 58.5% and 56.9%, respectively, indicating their limitations in leveraging the selected features.

BEST MODEL PERFORMANCE

Random Forest on	Accuracy	F1-Score
Unbalanced Data	81.6	74.8
Balanced Data (Over-Sampling)	81.0	76.7
Balanced Data (Under-Sampling)	82.6	78.3
Feature Selection (Scaled Data)	74.8	69.3

Table 1. Random Forest.

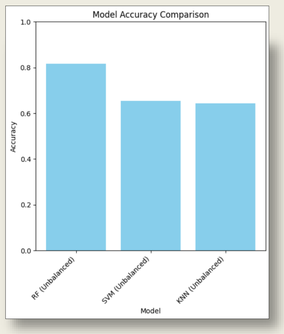

Figure 1. For Unbalanced data.

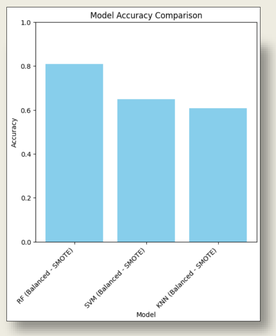

Figure 2. Balanced data (SMOTE)

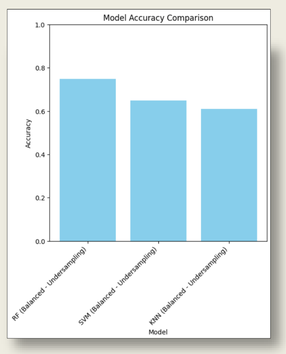

Figure 3. Balanced Under sampling

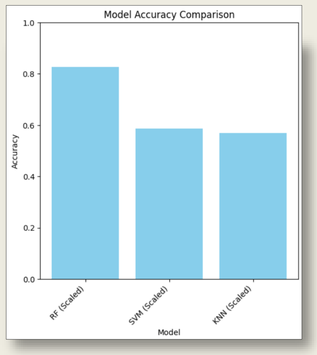

Figure 4. For Scaled data

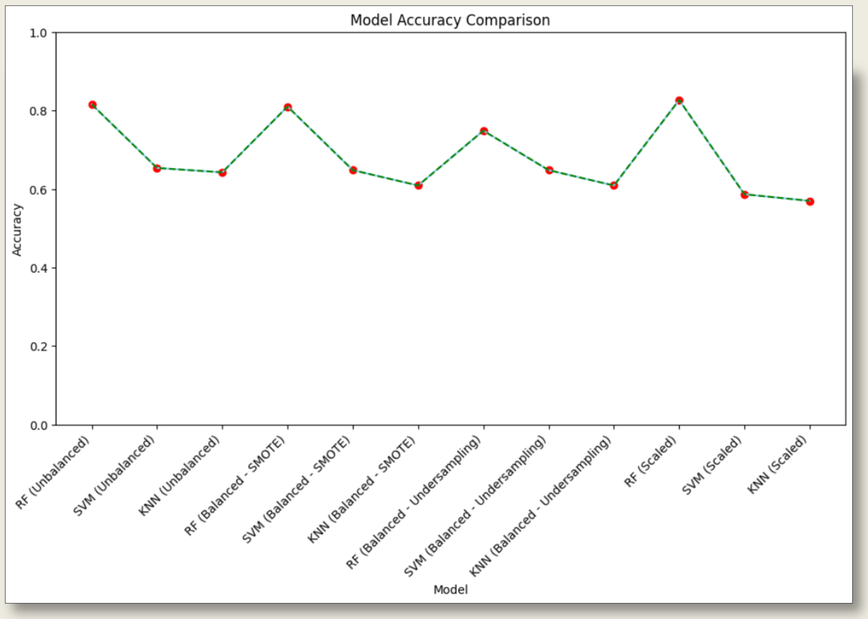

Chart 1. All results.

DISCUSSION

The results of this study demonstrate that data preprocessing significantly impacts the performance of machine learning models. Random Forest consistently outperformed SVM and KNN across all scenarios, showcasing its robustness and adaptability to different preprocessing techniques. While balancing the data through over-sampling and under-sampling helped mitigate class imbalance, it slightly affected the accuracy of Random Forest and did not significantly improve SVM or KNN. Feature selection led to the highest accuracy for Random Forest, emphasizing the importance of focusing on relevant predictors. However, SVM and KNN struggled with feature-selected data, likely due to their sensitivity to the reduced feature set. These findings highlight the importance of aligning preprocessing strategies with the strengths and limitations of each model to achieve optimal results, particularly when dealing with imbalanced datasets.

CONCLUSIONS

This project evaluated the performance of Random Forest, SVM, and KNN on the Titanic dataset under various preprocessing techniques, including handling unbalanced data, balancing through over-sampling and under-sampling, and applying feature selection. The results showed that Random Forest consistently outperformed the other models, demonstrating its robustness and adaptability. Data balancing had a mixed impact, improving class distribution but slightly reducing overall accuracy in some cases. Feature selection proved highly effective for Random Forest but highlighted the limitations of SVM and KNN in leveraging reduced feature sets. These findings underscore the importance of selecting appropriate preprocessing strategies and models tailored to the dataset characteristics to achieve optimal classification performance.