Exploring Machine Learning Approaches for Medical Condition Classification

Farah Walid Abdelsalam

Software Engineer

Uneeq interns Machine learning task extra

***Abstract*— This report presents a comparative analysis of four machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naive Bayes—on a healthcare dataset. The primary goal is to predict the medical condition of patients based on various features. The models were evaluated based on accuracy, precision, recall, and F1 score. The results show that each model has it's strengths and weaknesses, with Random Forest achieving the highest performance across most metrics.**

Keywords— Logistic Regression, Random forest, SVM, Naïve Bayes, healthcare dataset, accuracy, precision, recall, f1-score, classification models.

I. Introduction

Accurate medical condition prediction has the potential to greatly enhance patient outcomes and healthcare delivery. Strong tools for analyzing large, complicated datasets and generating predictions based on patterns in the data are provided by machine learning. In this paper, we examine the results of four widely used classification algorithms on a healthcare dataset: Random Forest, SVM, Naive Bayes, and Logistic Regression. Finding the model that makes the most accurate forecasts in this situation is the goal.

II. Data Description and Preprocessing Steps

Preprocessing Steps:

- **Loading and Inspecting Data:** The dataset was loaded using pandas, and initial exploration was done to understand the data structure.

- **Feature Selection:** The features 'Date of Admission' and 'Discharge Date' were dropped as they were deemed irrelevant for the prediction task.

- **Encoding Categorical Variables:** Categorical features such as 'Name', 'Gender', 'Blood Type', etc., were encoded using LabelEncoder.

- **Train-Test Split:** The dataset was split into training and testing sets with a test size of 45.5%.

- **Feature Scaling:** To ensure that features contribute equally to the model, standardization was applied using StandardScaler.

III. Using Methodology/Approach

Four machine learning models were implemented:

- Logistic Regression: A linear model used for binary and multi-class classification, optimized using the liblinear solver.

- Random Forest: An ensemble learning method that builds multiple decision trees and merges them for better accuracy and stability.

- Support Vector Machine (SVM): A model that finds the hyperplane that best separates different classes.

- Naive Bayes: A probabilistic model that assumes independence among predictors.

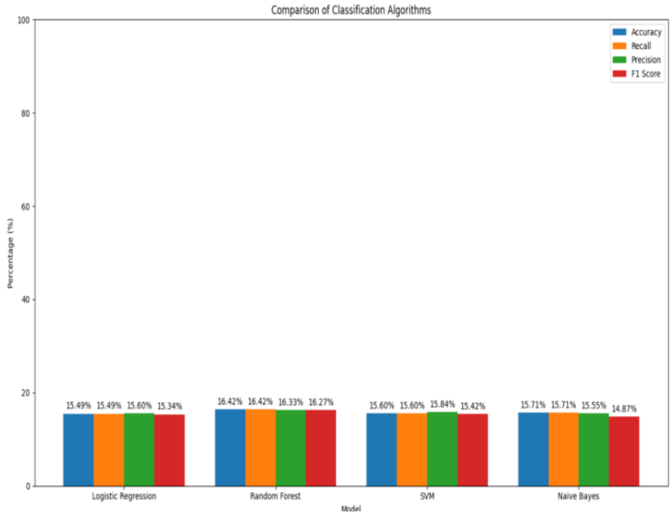
Each model was trained on the preprocessed training data and then evaluated on the test data. The metrics used for evaluation included accuracy, precision, recall, and F1 score, calculated using the sklearn.metrics module.

IV. Results

The results obtained from the four models are as follows:

Model Comparisons

	Logistic Regression	SVM	Random Forest	Naïve Bayes
Accuracy	15.49%	15.60%	16.42%	15.7%
Recall	15.49%	15.60%	16.42%	15.7%
Precision	15.60%	15.84%	16.33%	15.5%
F1-Score	15.34%	15.42%	16.27%	14.8%



Among the models evaluated, the Random Forest algorithm performed the best with an accuracy of 16.42% and an F1 score of 16.27%. This indicates a marginally better balance between precision and recall compared to the other models. The Support Vector Machine also showed slightly better precision than the others but overall, all models exhibited relatively low performance, suggesting that further optimization or alternative approaches may be necessary for improved predictive accuracy.

V. CONCLUSION

This study explored the application of various machine learning models to predict medical conditions using a healthcare dataset. The performance of the models—Logistic Regression, Support Vector Machine, Random Forest, and Naive Bayes—was generally low, with accuracy, precision, recall, and F1 scores hovering around 15-16%. These results suggest that the models, in their current form, are not effective at making reliable predictions for this dataset. The Random Forest model showed slightly better performance compared to the others, but the overall low metrics indicate a need for further improvement. Future work should focus on optimizing the models through hyperparameter tuning, addressing potential data quality issues, or exploring alternative algorithms. Additionally, techniques such as feature engineering, data augmentation, or the use of ensemble methods may help to enhance predictive accuracy and model robustness.