***Predicting Hospital Readmission Using a Simple Machine Learning Model***

**Executive Synopsis**  
In order to address a binary classification problem, this study applies three machine learning models: RandomForestClassifier, XGBoost, and Logistic Regression. The preprocessing procedures, model training, and accuracy and confusion matrix evaluation are described in full in the study. Using GridSearchCV, hyperparameter tweaking was carried out on XGBoost, yielding an accuracy of 61.92%. A review of the model's performance and prospective areas for development comes to a close in this report.

**1. Overview**  
This study aimed to use multiple machine learning techniques to tackle a binary classification problem. In order to find the most accurate model for outcome prediction, this study applied XGBoost, RandomForestClassifier, and Logistic Regression. To further maximise XGBoost's performance—which is well-known for being effective in structured data classification tasks—hyperparameter tuning was used.  
  
Model correctness serves as the main performance metric in the analysis. To ensure a fair model evaluation, the dataset underwent preprocessing prior to model training and was split into training and testing sets.

**2. Techniques/Process**  
To avoid overfitting and guarantee that the model performs well in general, the dataset was divided into training and test sets at the beginning of the procedure. The following algorithms were then used to train the models:  
  
A basic baseline model for binary classification is logistic regression.  
RandomForestClassifier: An ensemble learning technique that enhances prediction performance by utilising several decision trees.  
XGBoost: A performance- and speed-optimized gradient boosting technique.  
Performance indicators, such as confusion matrices and accuracy, were employed to assess the models' efficacy once they had been trained. Lastly, GridSearchCV was used for hyperparameter tweaking on XGBoost in order to determine the optimal parameters.

**3. Outcomes and Discoveries**  
Based on its accuracy and confusion matrix, each model was assessed:  
  
Logistic Regression: This model did a passable job of providing a baseline accuracy.  
RandomForestClassifier: By utilising an ensemble of decision trees, this model outperformed Logistic Regression in terms of performance.  
XGBoost: XGBoost outperformed RandomForestClassifier and Logistic Regression in terms of accuracy without any adjustment. Using GridSearchCV for hyperparameter tuning, XGBoost was able to attain an accuracy of 61.92%.  
All models made mistakes in both false positives and false negatives, according to the confusion matrices, with XGBoost having the greatest overall performance.

**4. Analysis and Discussion**  
With an accuracy of 61.92%, the findings show that XGBoost was the best-performing model following hyperparameter adjustment. Even if they were capable, the other models' performance fell short. Even while it is useful, the accuracy metric may not accurately capture the predictive ability of the model, especially when dealing with imbalanced datasets. Thus, in subsequent investigations, metrics such as ROC-AUC or F1-score may provide more information.  
  
Because of its capacity to manage missing data, regularisation to avoid overfitting, and effective tree-building, XGBoost performs well. The model's prediction abilities were eventually enhanced through the tuning procedure, which concentrated on characteristics including the maximum depth, learning rate, and number of estimators.

**5. Concluding remarks**  
Three machine learning algorithms were successfully applied for a binary classification problem in this research, with XGBoost outperforming the other two. The accuracy of XGBoost was considerably increased via hyperparameter tweaking. Even if the accuracy of 61.92% indicates a respectable prediction power, more improvement is still achievable.  
  
To enhance model performance, future research could include feature engineering, sophisticated tuning methods, and the addition of new performance indicators. A more thorough grasp of the models' generalisation abilities would also be possible by testing them on untested data and using cross-validation.

***THANK YOU***

**Submitted by: Ayush Kumar Raikwar**

**Date: 11/10/24**