A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR CARDIOVASCULAR DISEASE RISK PREDICTION

SUPERVISOR/INSTRUCTOR

Dr.Meshbauddin Sarkar,Professor

Institute of Information Technology

Jahangirnagar University

PRESENTED BY

Md.Emad Uddin Aksir

ID-233114

Batch: PMIT(30) **Section:A**

Presentation Contents

INTRODUCTION

- Cardiovascular disease (CVD) is the leading cause of death for both men and women in the world. According to the World Health Organization (WHO), nearly 18 million, or 31%, of deaths are caused by this disease globally.

- The goal of this research is to develop a machine learning model that can accurately predict the risk of cardiovascular disease (CVD) in individuals.

- This research is used to identify individuals who are at high risk for cardiovascular disease (CVD), so that they can be screened and treated early, before the disease develops.

MOTIVATION

- Contribute to the advancement of machine learning methodologies in healthcare.
- Identify the most important factors that contribute to CVD risk.
- Validate the accuracy of the machine learning model in a large, independent dataset.
- Find the best machine learning algorithm based on their results.
- To aid clinicians in decision-making by artificial intelligence (AI).
- To develop a user-friendly and accessible risk assessment tool.

OBJECTIVES

- To check how much each type of machine learning model can help us predict the Cardiovascular disease risk.
- Using class imbalance using SMOTE and assess its impact on model performance.
- Compare the performance of ensemble learning methods with traditional machine learning models.
- To find out the most important factors affecting Cardiovascular disease risk prediction using the Boruta Algorithm

RELATED WORKS

Machine Learning Models for Cardiovascular Disease: A Review on Early Diagnosis Performance

RF,KNN and Custom machine learning model ,accuracy gained 88%–89%

Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis

CatBoost machine learning model ,stacking catboost,accuracy gained 90.94%

Clinical Decision Support System for Cardiovascular Disease Prediction Using a Novel Deep Ensemble Learning Approach

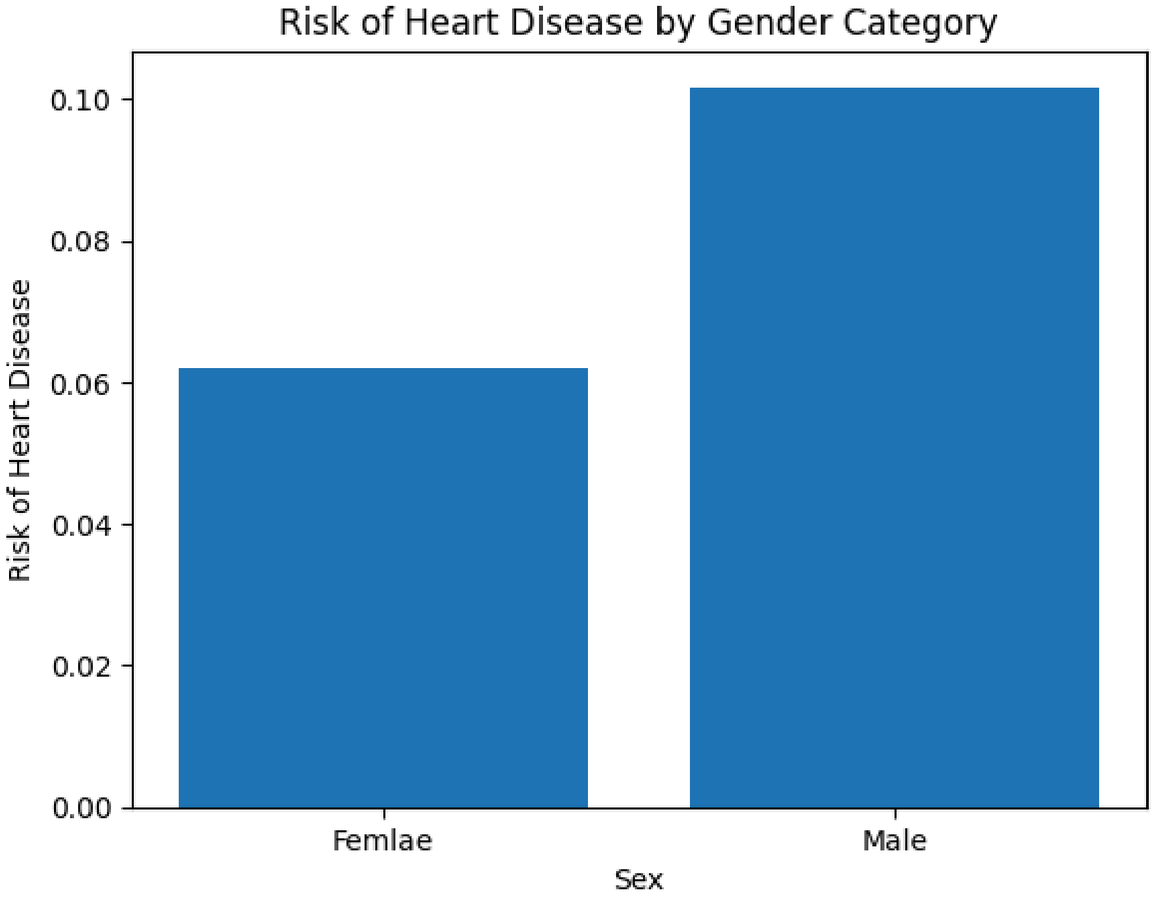
CNN and Stacking using,accuracy found 93.97%

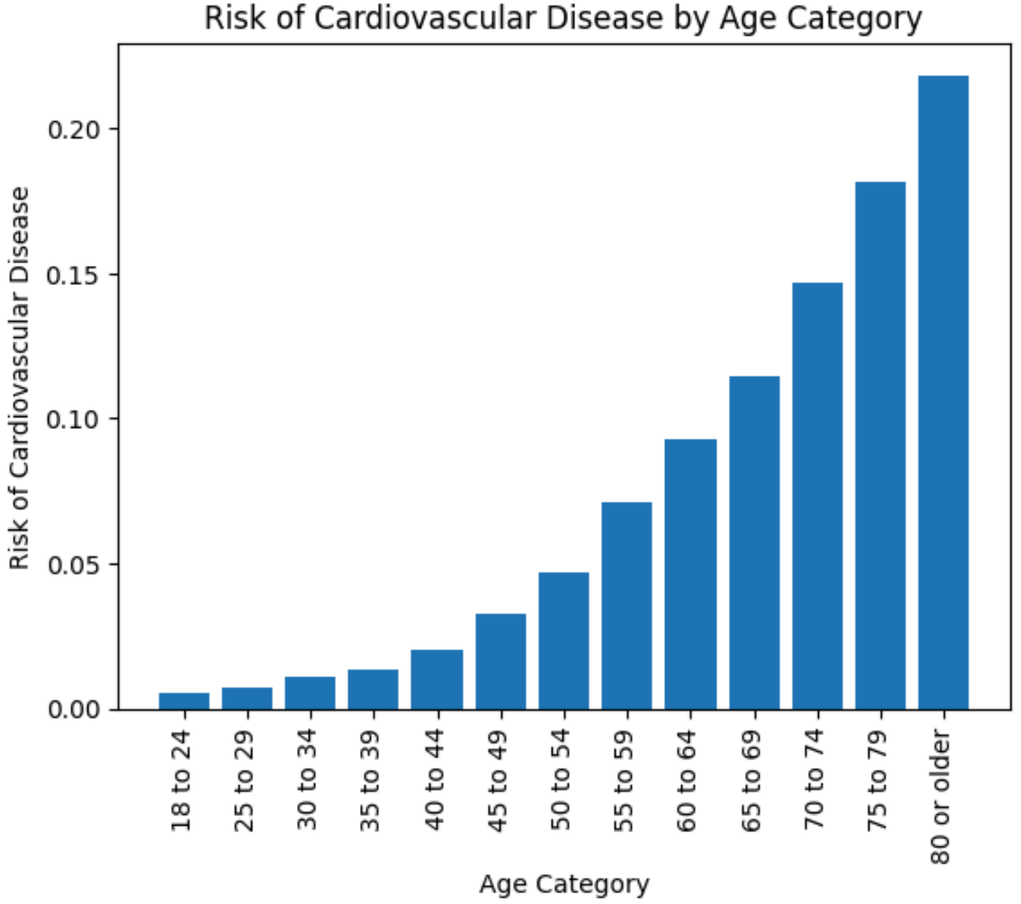
DATA PROCESSING

To commence the model evaluation process, a unique categorical dataset comprising **308,070 rows** and **17 columns** was utilized. The dataset was collected from the **Kaggle** online platform. To enrich the usability of dataset more accurately Data Cleaning (removing duplicate values, null values), Data Encoding (Label Encoding), Data Oversampling (SMOTE) and Data Standardization (Standard Scaler) applied.

```
Index: 308070 entries, 0 to 308853
Data columns (total 17 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   General_Health                308070 non-null  object
 1   Exercise                      308070 non-null  object
 2   Skin_Cancer                   308070 non-null  object
 3   Other_Cancer                  308070 non-null  object
 4   Depression                    308070 non-null  object
 5   Diabetes                      308070 non-null  object
 6   Arthritis                     308070 non-null  object
 7   Sex                           308070 non-null  object
 8   Age_Category                  308070 non-null  object
 9   Height_(cm)                   308070 non-null  int64
 10  Weight_(kg)                   308070 non-null  float64
 11  BMI                           308070 non-null  float64
 12  Smoking_History               308070 non-null  object
 13  Alcohol_Consumption           308070 non-null  int64
 14  Fruit_Consumption             308070 non-null  int64
 15  Green_Vegetables_Consumption  308070 non-null  int64
 16  Cardio_Disease                308070 non-null  object
dtypes: float64(2), int64(4), object(11)
```

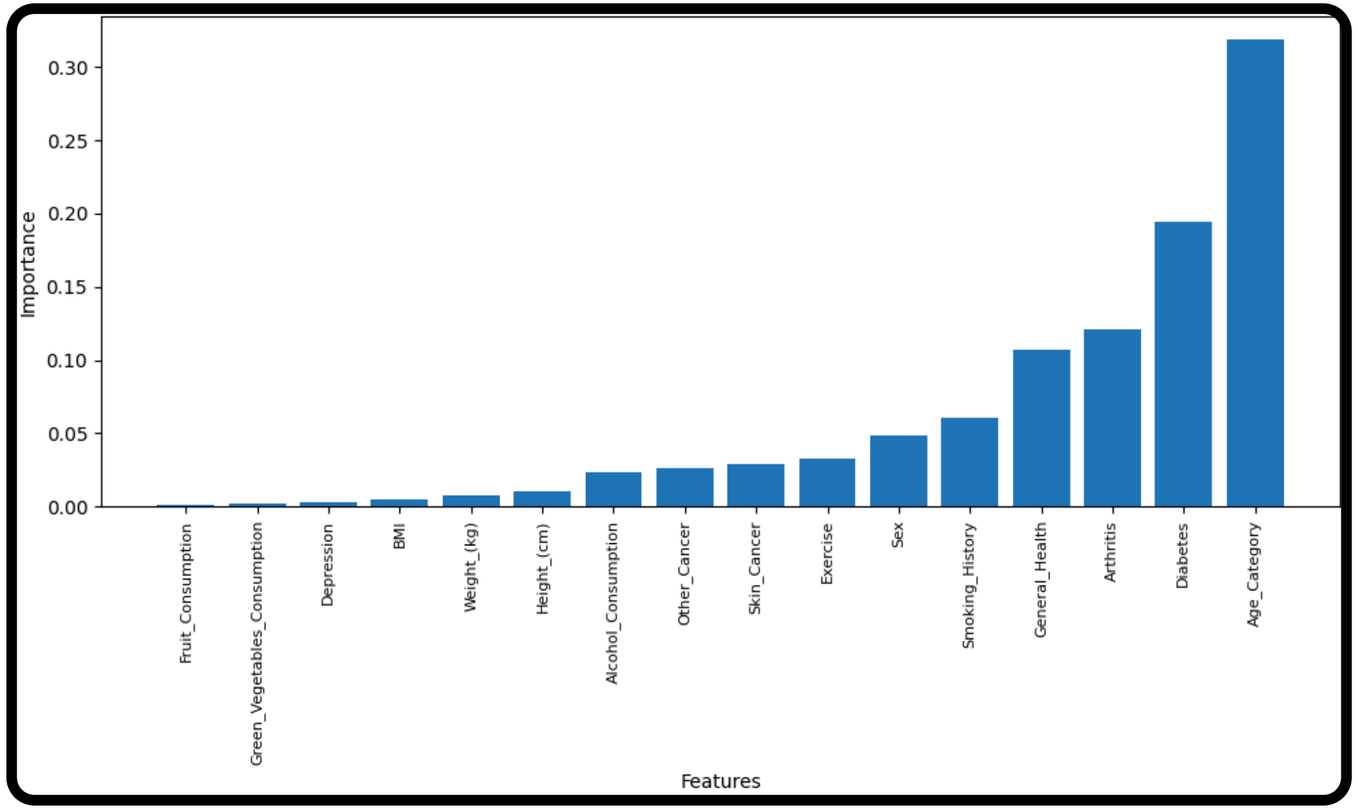
DATA OVERVIEW





FEATURE IMPORTANCE

The most influential predictor groups identified by **Boruta** consist of Age_Category followed by Diabetes and Arthritis and General Health. Smoking History and Sex together with Exercise demonstrate weak but additional predictive strength among the behavioural factors. The other lifestyle variables including cancer indicators along with alcohol usage and body size measurements become negligible when Age Category, Diabetes, Arthritis, and General Health remain in the predictive model.



WORKING OVERVIEW

To prepare the dataset for machine learning, initial preprocessing involved converting all categorical features into numerical representations using label encoding. An examination of the target variable, 'Cardio_Disease', revealed a significant class imbalance, with 283,883 instances of 'No' and 24,971 instances of 'Yes', indicating a skew towards the negative class. To address this imbalance and ensure robust model training, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, resulting in a balanced dataset with 283,103 instances of 'No' and 226,482 instances of 'Yes' for 'Cardio_Disease'. The balanced dataset was then partitioned into training (80%) and testing (20%) sets. To explicitly demonstrate the impact of the data balancing procedure, each machine learning model was trained and evaluated both with and without the application of SMOTE. Furthermore, to optimize the performance of each model and achieve higher accuracies, a RandomizedSearchCV approach with 5-fold cross-validation was employed to explore a range of hyperparameters.

MODEL SCHEMES

- Logistic Regression
- Decision Tree
- Support Vector Machine
- K-Nearest Neighbor
- Random Forest
- Gaussian Naive Bayes
- Gradient Boosting
- Adaboosting
- XGB
- Stacking

Results analysis and discussion

RESULTS

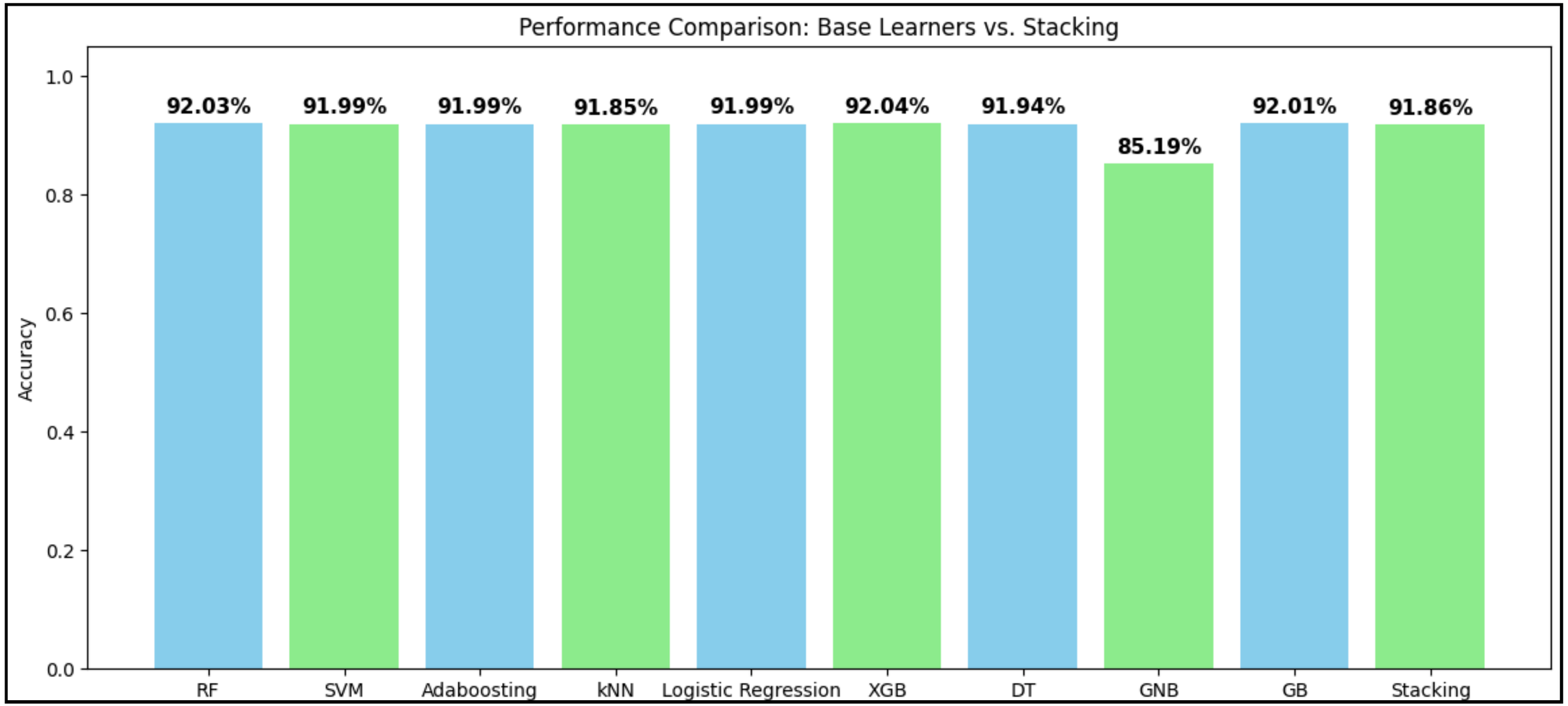


Fig: Accuracy without SMOTE Method

RESULTS

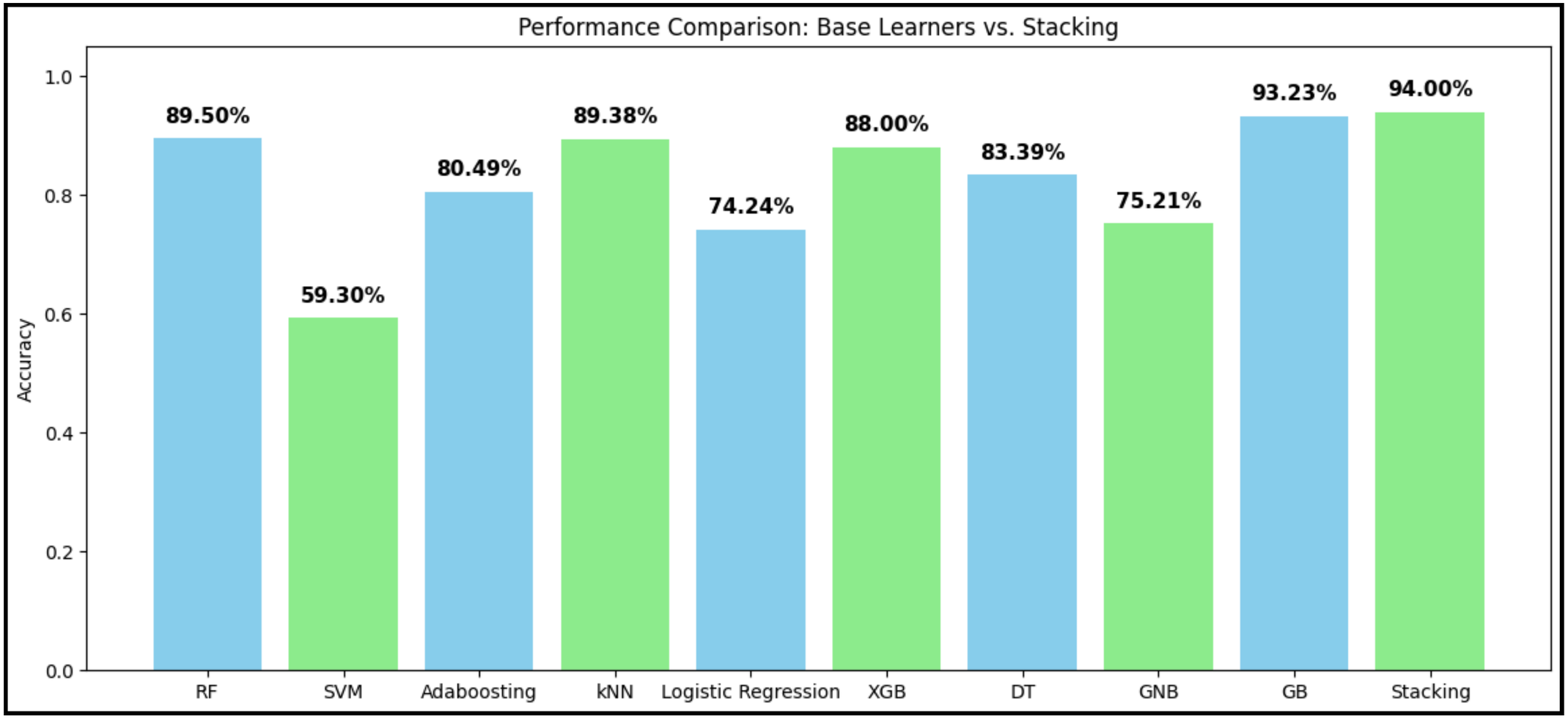


Fig: Accuracy with SMOTE Method

DISCUSSION

We balanced the data with SMOTE, expecting accuracy to rise. Strangely, the raw accuracy was actually higher before oversampling. That “extra” accuracy is misleading because the model was simply predicting the majority class over and over. When we check the real quality measures—precision, recall, and F1-score—many models score close to zero without SMOTE, proving they miss almost every minority-class case. SMOTE creates synthetic examples of the rare class, giving the model more to learn from, so these scores jump sharply afterward. Although the headline accuracy number drops a bit, the model now spots positive cases it once ignored, making its predictions far more trustworthy in practice. In short, the post-SMOTE results give a clearer, fairer view of the model’s true skill.

PRECISION, RECALL, F1-SCORE

Model Name	Precision		Recall		F1 Score	
	0	1	0	1	0	1
Random Forest	0.92	0.56	1.00	0.02	0.96	0.04
Support Vector Machine	0.92	0.00	1.00	0.00	0.96	0.00
Logistic Regression	0.92	0.00	1.00	0.00	0.96	0.00
Gaussian Naïve Bayes	0.94	0.23	0.89	0.36	0.92	0.28
Adaboosting	0.92	0.00	1.00	0.00	0.96	0.00
Extreme Gradient Boosting	0.92	0.32	1.00	0.02	0.96	0.03
Decision Tree	0.92	0.39	1.00	0.01	0.96	0.02
Gradient Boosting	0.92	0.52	1.00	0.03	0.96	0.05
KNN	0.90	0.78	0.78	0.99	0.87	0.87
Stacking	0.92	0.52	1.00	0.01	0.96	0.02

Table 1 : Without SMOTE

Model Name	Precision		Recall		F1-Score	
	0	1	0	1	0	1
Random Forest	0.91	0.88	0.90	0.88	0.91	0.88
Support Vector Machine	0.58	0.82	0.98	0.11	0.73	0.19
Logistic Regression	0.79	0.69	0.72	0.77	0.76	0.73
Gaussian Naïve Bayes	0.79	0.71	0.75	0.75	0.77	0.73
Adaboosting	0.84	0.79	0.84	0.80	0.84	0.80
Extreme Gradient Boosting	0.90	0.91	0.93	0.87	0.92	0.89
Decision Tree	0.87	0.79	0.81	0.85	0.84	0.82
Gradient Boosting	0.92	0.94	0.96	0.90	0.94	0.92
KNN	0.90	0.78	0.78	0.99	0.87	0.87
Stacking	0.95	0.92	0.94	0.94	0.95	0.93

Table 2 : With SMOTE

CONCLUSION

This study aimed to predict early-stage cardiovascular disease using data from 308,070 patients. After cleaning and encoding, the Boruta algorithm identified key features. Due to data imbalance, SMOTE was applied. Models trained without SMOTE showed high accuracy but poor minority class performance, with very low precision, recall, and F1-scores. Random Forest, XGB, and GB had high overall accuracies (~92%) but poor minority class prediction. After applying SMOTE, minority class metrics improved significantly across most models, though overall accuracy slightly dropped. Stacking achieved the best results with 94% accuracy, followed by Gradient Boosting, Random Forest, KNN, and XGB. SMOTE effectively balanced the dataset, leading to more reliable model evaluations.

FUTURE WORKS

- We will apply for the Deep Learning Model
- Will Apply the Model Optimization Techniques
- We will apply it to the geographical dataset
- Will include local geolocated dataset
- Develop a Web based application

THANK YOU