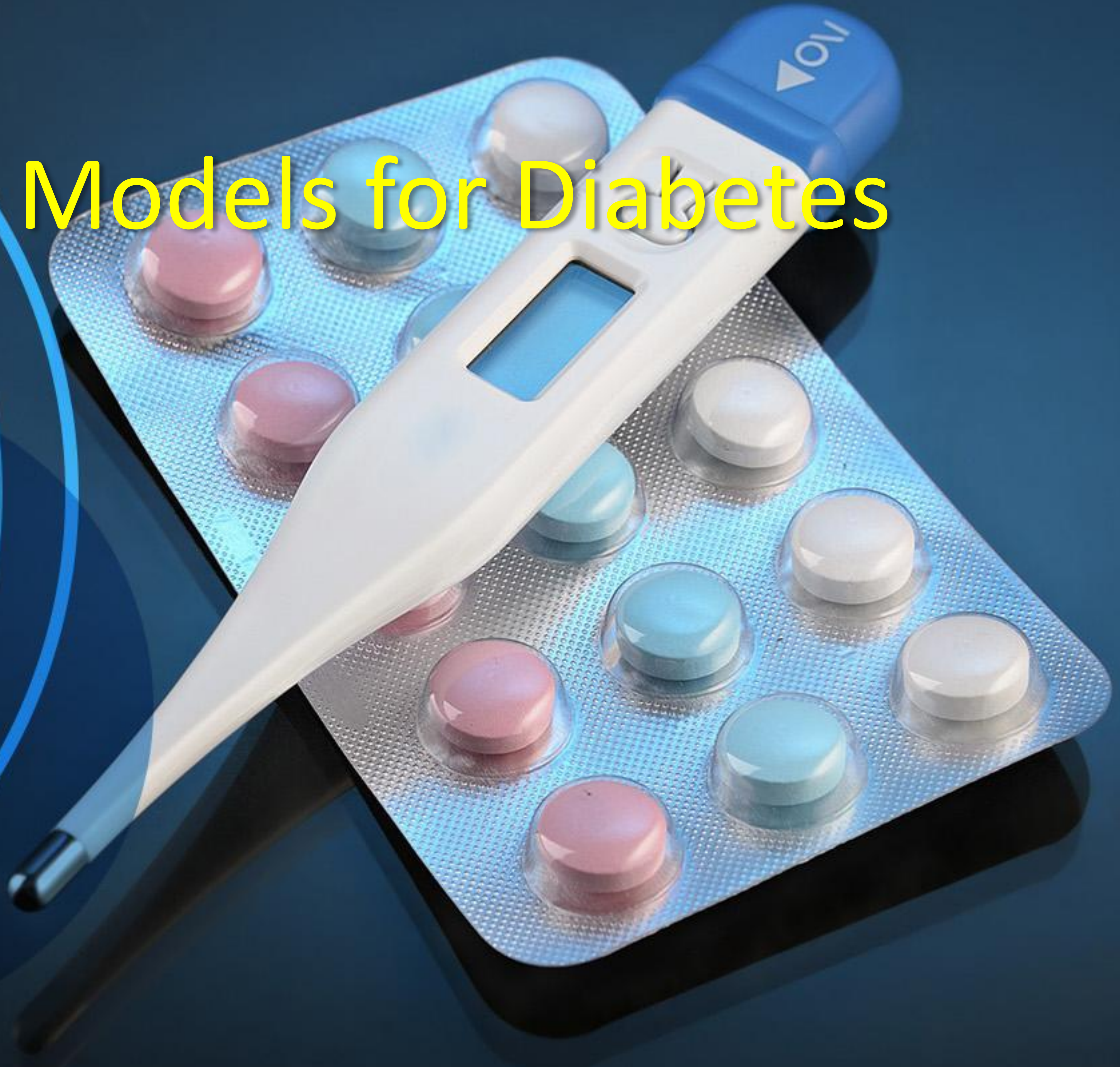
Exploring Predictive Models for Diabetes Diagnosis



By:

STEPHEN MULINGWA

Introduction

Global Health Concern: Diabetes is a growing issue worldwide with significant health and economic impacts.

Importance of Early Detection: Early identification is critical to managing diabetes and preventing complications.

Project Goal: Develop a machine learning model to predict the risk of diabetes in individuals.

Value to Stakeholders:

- **Healthcare Providers**: Equip them with a tool to prioritize intervention for high-risk patients.
- **Healthcare Systems**: Enable efficient resource allocation to reduce the burden of diabetes.
- **Policymakers and Organizations**: Support initiatives focused on diabetes prevention and management.

Data Overview

Dataset Source: CDC’s BRFSS2015 dataset.

Sample Size: Over 250,000 individuals with diverse health indicators.

Key Features:

- **Health Metrics**: BMI, physical activity, mental health, smoking status, and blood pressure and **Target Variable**: Classification of individuals as diabetic, prediabetic, or non-diabetic.

Challenges Identified:

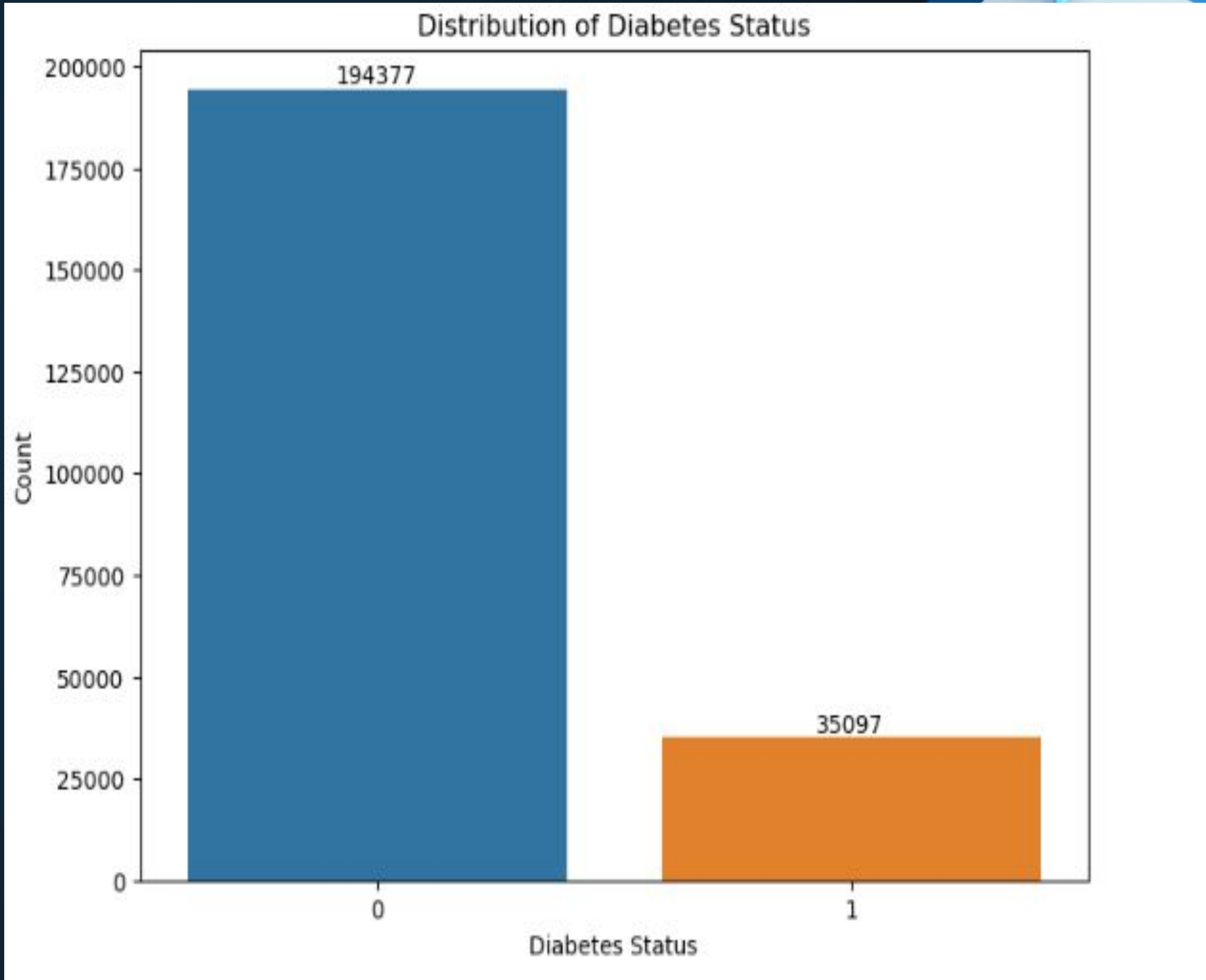
- **Class Imbalance**: Significantly more non-diabetic cases, impacting model accuracy.
- **Skewed Distributions**: Continuous variables, such as BMI, required scaling and normalization during preprocessing.

Methodology

- The **NearMiss algorithm** was applied to address class imbalance, ensuring balanced representation of diabetic and non-diabetic classes.
- Exploratory Data Analysis revealed significant predictors of diabetes, including **BMI**, **mental health status**, and **difficulty in walking/climbing**.
- Multiple machine learning models were built and evaluated, including **Logistic Regression**, **Decision Trees**, **Random Forest**, and **XGBoost**.
- Model performance was assessed using metrics such as **Accuracy, Confusion Matrix**, **F1-score**, and **AUC** to ensure reliability.
- The objective was to select the most effective model to provide accurate predictions and meet stakeholder requirements.

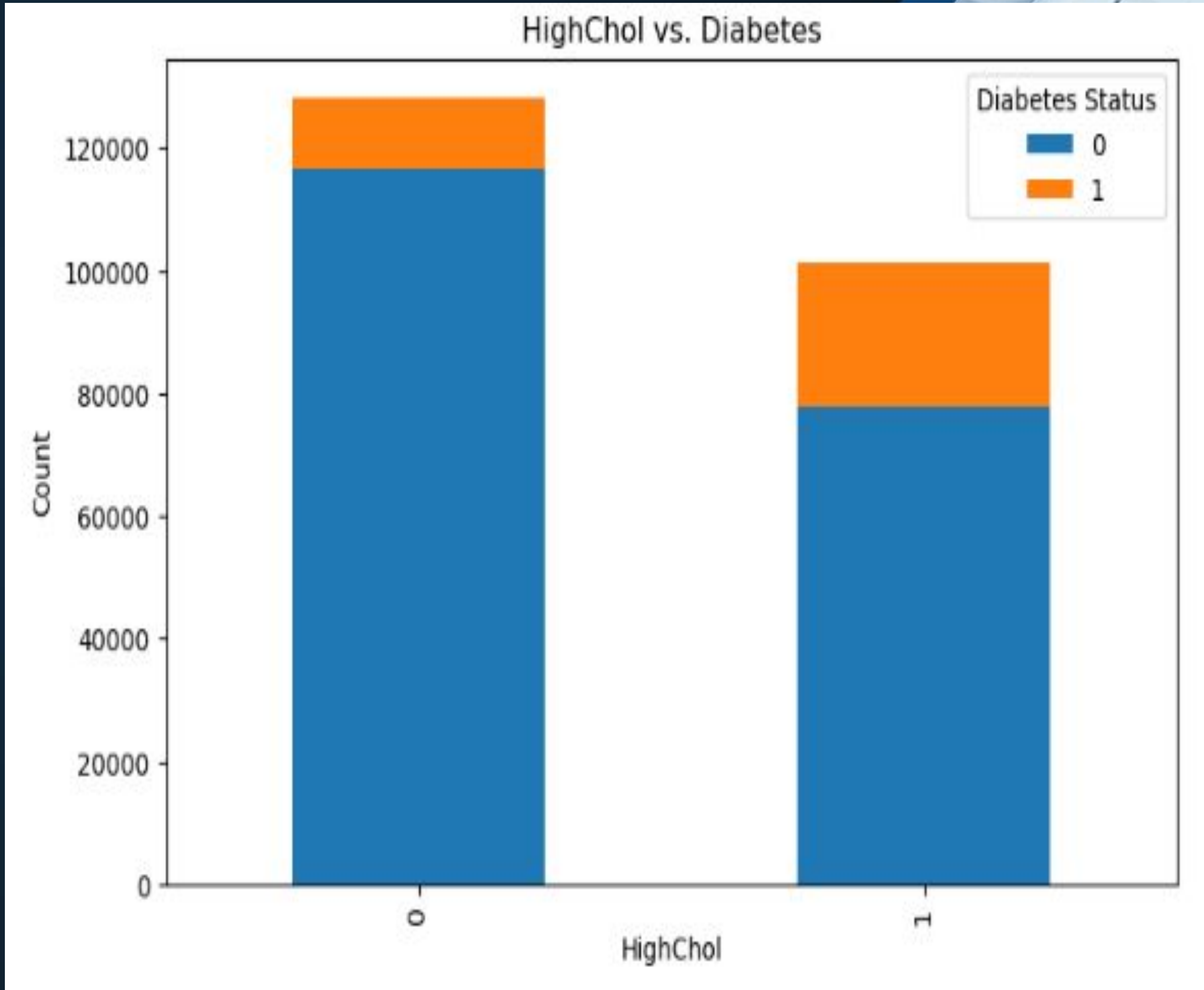
Analysis and Results

From the graph there was class imbalance with class 0 (people with no diabetes) being high. I balanced the data using nearmiss and had 70,194 0bservarions.



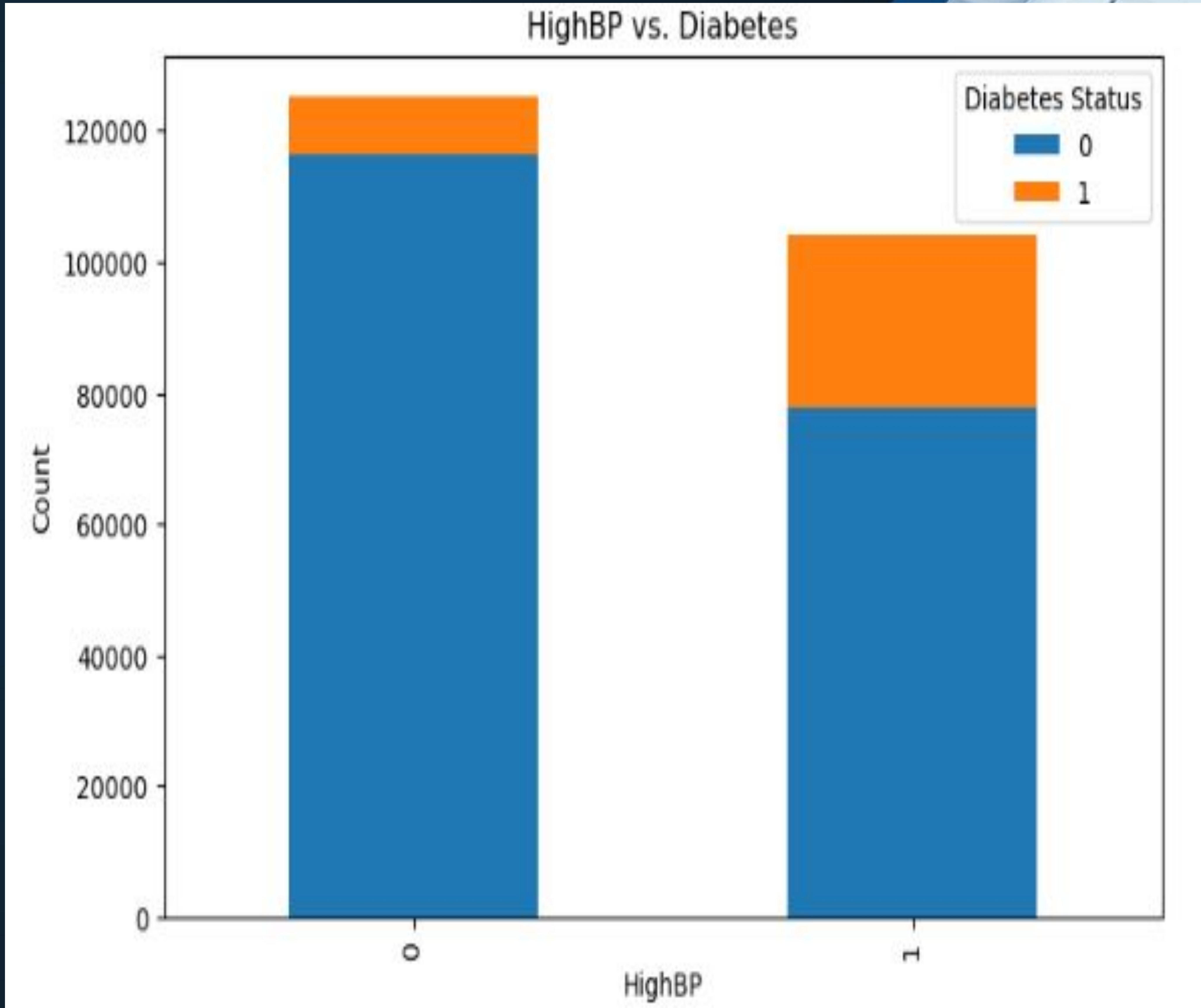
High Cholesterol Vs Diabetes

Based on the barplot people with High cholesterol are at higher risk of developing Diabetes.



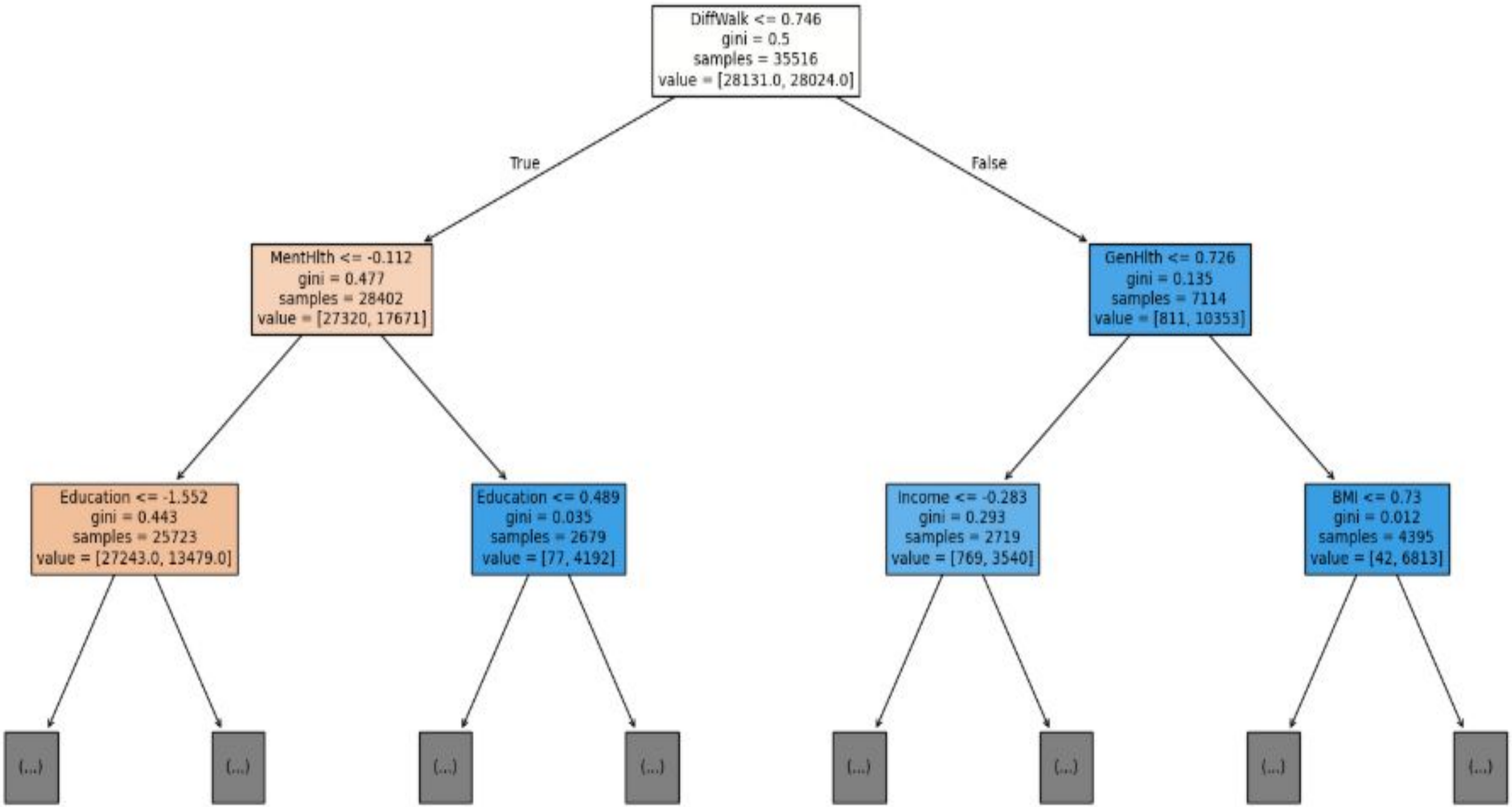
High Blood Pressure Vs Diabetes

Based on the barplot people with High blood pressure are at higher risk of developing Diabetes.

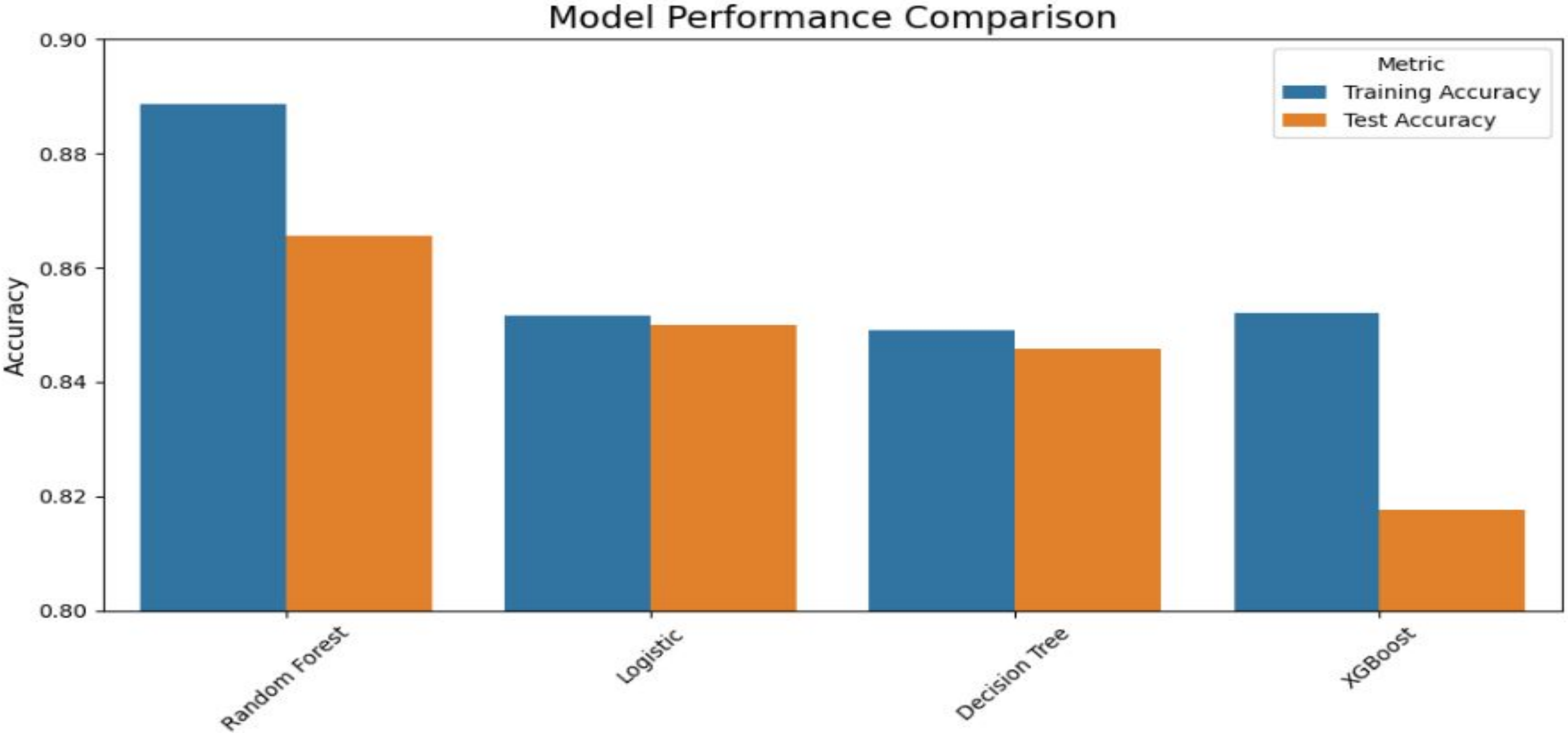


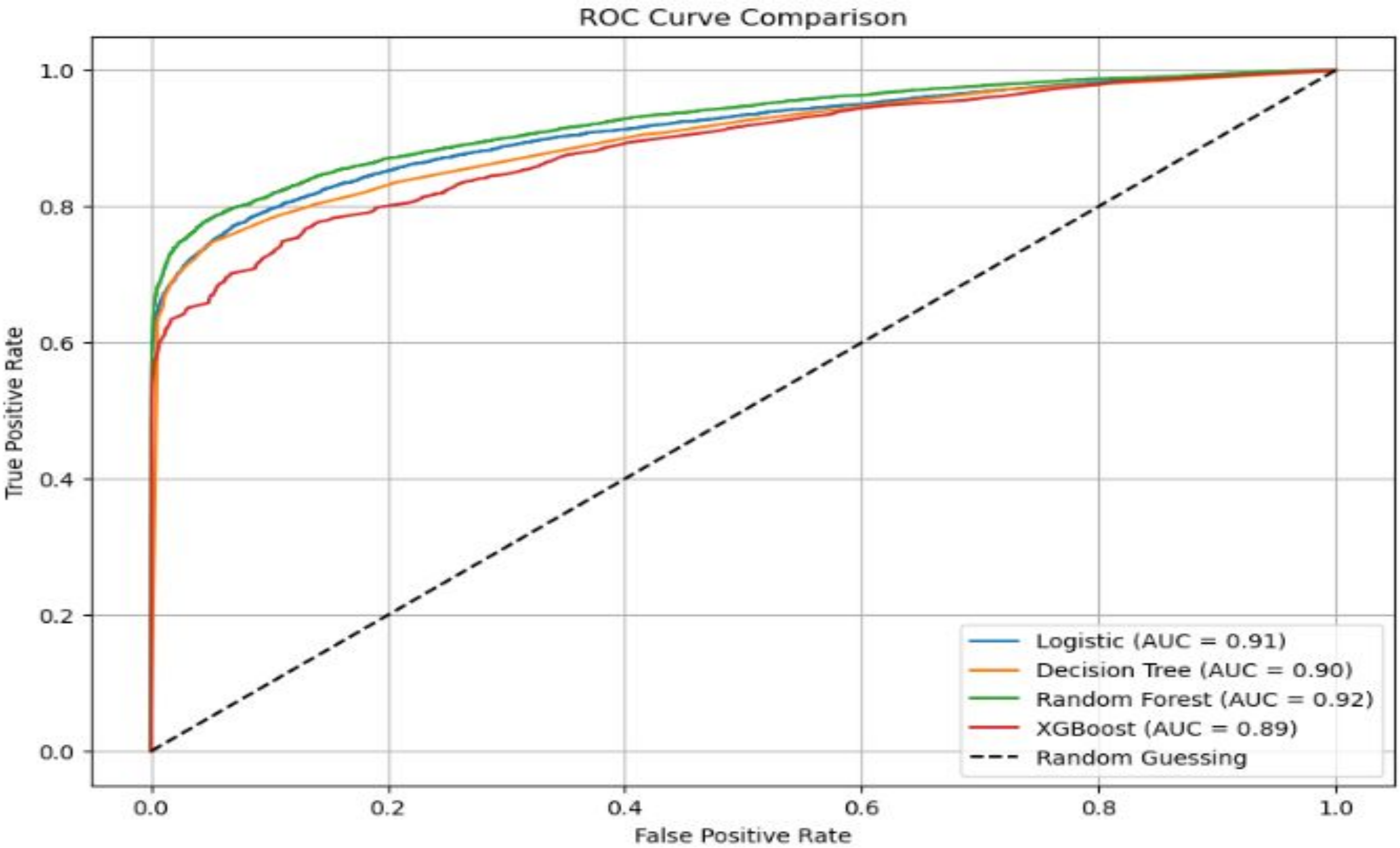
Random Forest

- **Random Forest** was the best-performing model, with a snippet of the decision tree shown in the next slide.
- Key predictors of diabetes risk included **difficulty in walking/climbing**, **poor mental health**, and **high BMI**.
- Individuals reporting **poorer general health** were also found to be at greater risk of developing diabetes.
- The findings emphasize the importance of **targeted interventions**, such as **weight management programs**, **mental health support**, and **mobility assistance** for at-risk individuals.
- A significant proportion of individuals also had **high blood pressure**, highlighting the need for **comprehensive care** to address comorbid conditions associated with diabetes.


DiffWalk <= 0.746
gini = 0.5
samples = 35516
value = [28131.0, 28024.0]
True
False
MentHlth <= -0.112
gini = 0.477
samples = 28402
value = [27320, 17671]
GenHlth <= 0.726
gini = 0.135
samples = 7114
value = [811, 10353]
Education <= -1.552
gini = 0.443
samples = 25723
value = [27243.0, 13479.0]
Education <= 0.489
gini = 0.035
samples = 2679
value = [77, 4192]
Income <= -0.283
gini = 0.293
samples = 2719
value = [769, 3540]
BMI <= 0.73
gini = 0.012
samples = 4395
value = [42, 6813]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Model Performance Comparison




ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic (AUC = 0.91)
Decision Tree (AUC = 0.90)
Random Forest (AUC = 0.92)
XGBoost (AUC = 0.89)
Random Guessing

Model Performance comparison

- Among the models tested, the **Random Forest classifier** delivered the best performance with a **test accuracy of 86.57%** and an **AUC of 0.92**.
- These results demonstrate the Random Forest model's effectiveness in distinguishing between **diabetic and non-diabetic individuals**.
- Other models, including **Logistic Regression** and **XGBoost**, performed well but did not match the Random Forest model’s **generalization ability** to unseen data.
- The **Random Forest model's performance** underscores its potential for **real-world application** in healthcare, where **accurate predictions** are critical for timely and effective interventions.

Conclusion

- This project successfully developed a **predictive model for diabetes risk** using machine learning techniques.
- The **Random Forest model**, with its **high accuracy and AUC**, emerged as the most reliable tool for predicting diabetes in this context.
- Key predictors of diabetes risk, including **difficulty in walking/climbing**, **BMI**, **mental health**, and **income**, were identified.
- By integrating this model into **healthcare systems**, providers can **prioritize individuals for diabetes screening and intervention**, leading to **improved patient outcomes** and a reduction in **healthcare costs** related to diabetes.

Recommendations

- **Integrate the predictive model** into routine healthcare screening processes to **identify high-risk individuals early**.
- **Focus interventions** on addressing key risk factors such as **high BMI**, **mental health**, and **mobility issues**.
- **Optimize the model** further by exploring additional **data balancing techniques** and more complex **ensemble methods** to improve its predictive power.
- **Monitor class imbalance** and incorporate **synthetic data** to enhance model training in future iterations.
- **Track the impact** of the model on **diabetes prevention** and **resource allocation** to ensure its effectiveness in real-world healthcare settings.

Next Steps

- **Deploy the predictive model** into healthcare systems with close **collaboration with healthcare providers** to ensure smooth integration into existing workflows.
- **Develop an easy-to-use interface or application** to allow healthcare professionals to quickly assess diabetes risk and make informed decisions.
- **Monitor and update the model** continuously as new data becomes available, ensuring its **accuracy** and **reliability** over time.
- **Ensure long-term utility** of the model by refining it based on feedback from healthcare professionals and real-world usage.
- **Establish the model as a vital tool** in managing **diabetes risk** and improving **patient care** across healthcare settings.

Thank you

QUESTIONS

Feel free to reach out with any questions.

📞 **TelePhone: 0111224952**

Gmail: mulingwastephen200@gmail.com

Linked Profile:

https://www.linkedin.com/in/stephen-mulingwa-105522205/