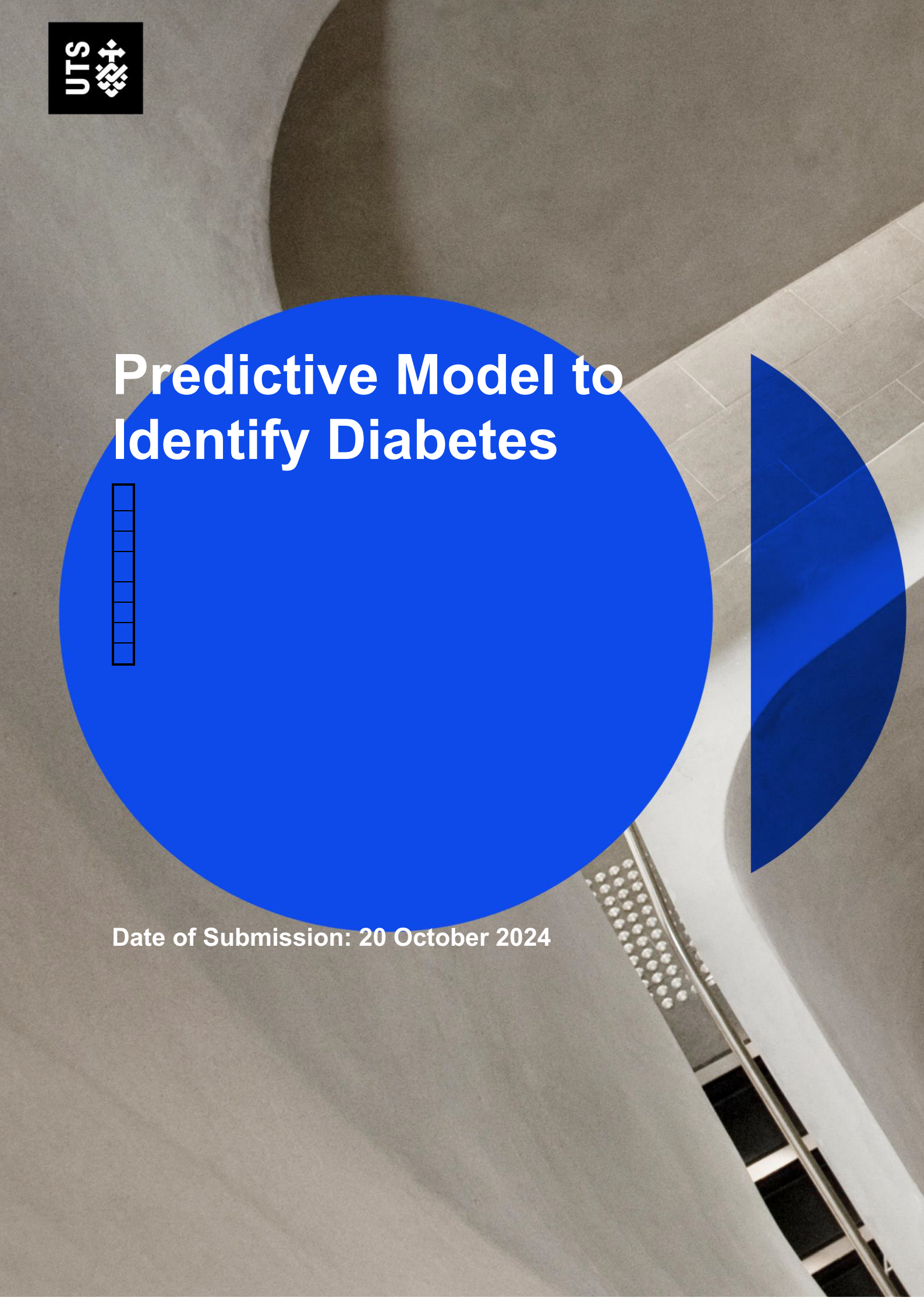


Predictive Model to Identify Diabetes

Date of Submission: 20 October 2024

Executive Summary

Diabetes is one of the world's most pervasive and chronic health conditions, affecting approximately 422 million people globally (World Health Organisation, n.d.). The disease inflicts a significant financial burden on nation-states, many of which have marginal (low and middle-income) economies. The condition impairs the life expectancy and the quality of life for those diagnosed. To put this into context, there are approximately 1.5 million deaths directly linked to diabetes each year.

The goal of this study is to create a predictive model to identify people who are at the most risk of developing diabetes. Binary and numerical characteristics, including BMI, hypertension, cholesterol, physical activity, and lifestyle variables, were included in the dataset. Developing predictive models that reliably predict diabetes was the main objective with the intention to assist treatment and medical practitioners identify and identify the key physical signs early on. Several machine learning models were trained and assessed by an extensive investigation that comprised feature engineering, data cleansing, and model experimenting. Random Forest, XGBoost, K-Nearest Neighbours, Decision Trees, and Logistic Regression were some of the models employed. Metrics like accuracy, precision, recall, and F1-score were used to gauge performance; recall was given special attention because it is crucial for identifying people who are at risk.

The analysis's main conclusions show that ensemble models—Random Forest and XGBoost in particular—performed better in terms of overall accuracy and recall. With the highest F1 score and the best overall performance, XGBoost demonstrated a balance between recall and precision. Additionally, recall was enhanced by about 3% with model modification, which is important for detecting more people who might have diabetes. This work has important implications for health management authorities and practitioners. When people at risk for diabetes are identified early and accurately, prompt therapies may be implemented, which lowers the long-term expenses of treating advanced-stage diabetes. Furthermore, by classifying patients based on risk, these models can assist healthcare organisations in allocating resources more effectively and facilitating more effective screening and preventative strategies.

In conclusion, this project's predictive models—XGBoost and Random Forest are useful resources for medical and health practitioners who want to improve diabetic screening and early diagnosis, which can be used to eventually improve patient outcomes and save expenses.

Table of Contents

1. Business Understanding

Business Use Cases

This project aims to develop a model that predicts the likelihood of a person having diabetes based on various health indicators. Healthcare providers and insurance companies can use this predictive model to identify at-risk individuals early, allowing for preventive measures and personalised healthcare services.

Key Objectives

- Understand which factors (i.e. blood pressure, cholesterol, smoking habits) are most influential in predicting diabetes.
- Develop a model that can detect people with diabetes early with high accuracy with lifestyle and health data.
- Provide healthcare organisations with a tool to guide resource allocation and intervention programs.

2. Data Understanding

Dataset Overview

The dataset is from the Behavioural Risk Factor Surveillance System (BRFSS), a survey-collected dataset regarding health from the United States of America through telephone (Centers for disease control and prevention & Dane, n.d.). The dataset is then initially cleaned by Teboul (n.d.) to include only diabetes-related features. There are 253,680 entries with 22 features, including health metrics such as blood pressure, cholesterol, physical activity, and diabetes diagnosis. Key features include:

- **HighBP** (binary): High blood pressure status.
- **BMI**: Body mass index.
- **Smoker**: Smoking status.
- **PhysActivity**: Physical activity status.
- **CholCheck** (binary): Cholesterol check status.
- **Diabetes binary** (target): Indicates whether a person has been diagnosed with diabetes or not (1 = Yes, 0 = No).

Ethical Considerations

The dataset contains data that may be considered sensitive, such as Age, Income and Gender. However, age and income have been grouped into bins to avoid revealing exact values, ensuring better privacy protection and reducing the risk of identification.

The dataset is available from Kaggle under a public license and does not contain Personal Identification Information (PII).

Data Limitations

- **Imbalance in target variable**: There may be an imbalance between positive and negative diabetes cases, which could affect model performance.
- **Binary representation**: Several health factors are binary (e.g., HighBP, Smoker), which could oversimplify some underlying patterns.
- **Potential for feature correlation**: Health indicators such as high blood pressure, cholesterol, and BMI may be correlated, potentially leading to multicollinearity.

Exploratory Data Analysis (EDA)

Class Distribution: As we see in Figure 1, the diabetes diagnosis (Diabetes binary) distribution shows 86% of individuals with diabetes and 14% without diabetes, indicating an imbalanced dataset since there is a significant disparity between the two classes.

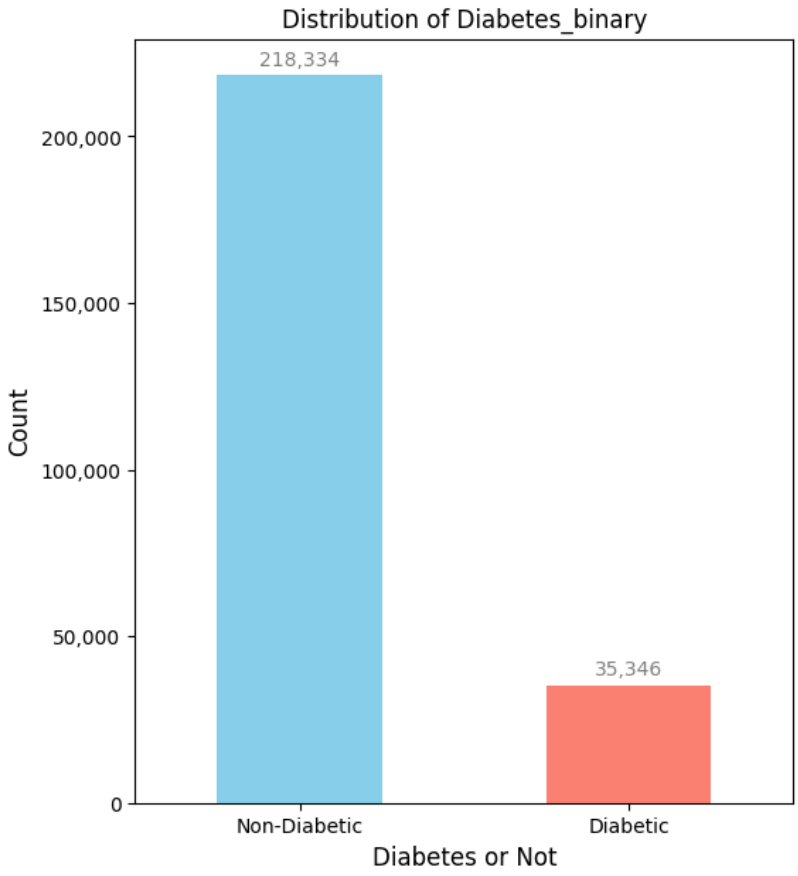


Figure 1: Diabetes_binary distribution

Distribution of Categorical Values: This section focuses on analysing the distribution of categorical variables by the target variable Diabetes binary. By visualising the frequency and proportions of diabetic/non-diabetic individuals within each category, the analysis uncovers potential correlations and provides insights into risk factors related to diabetes.

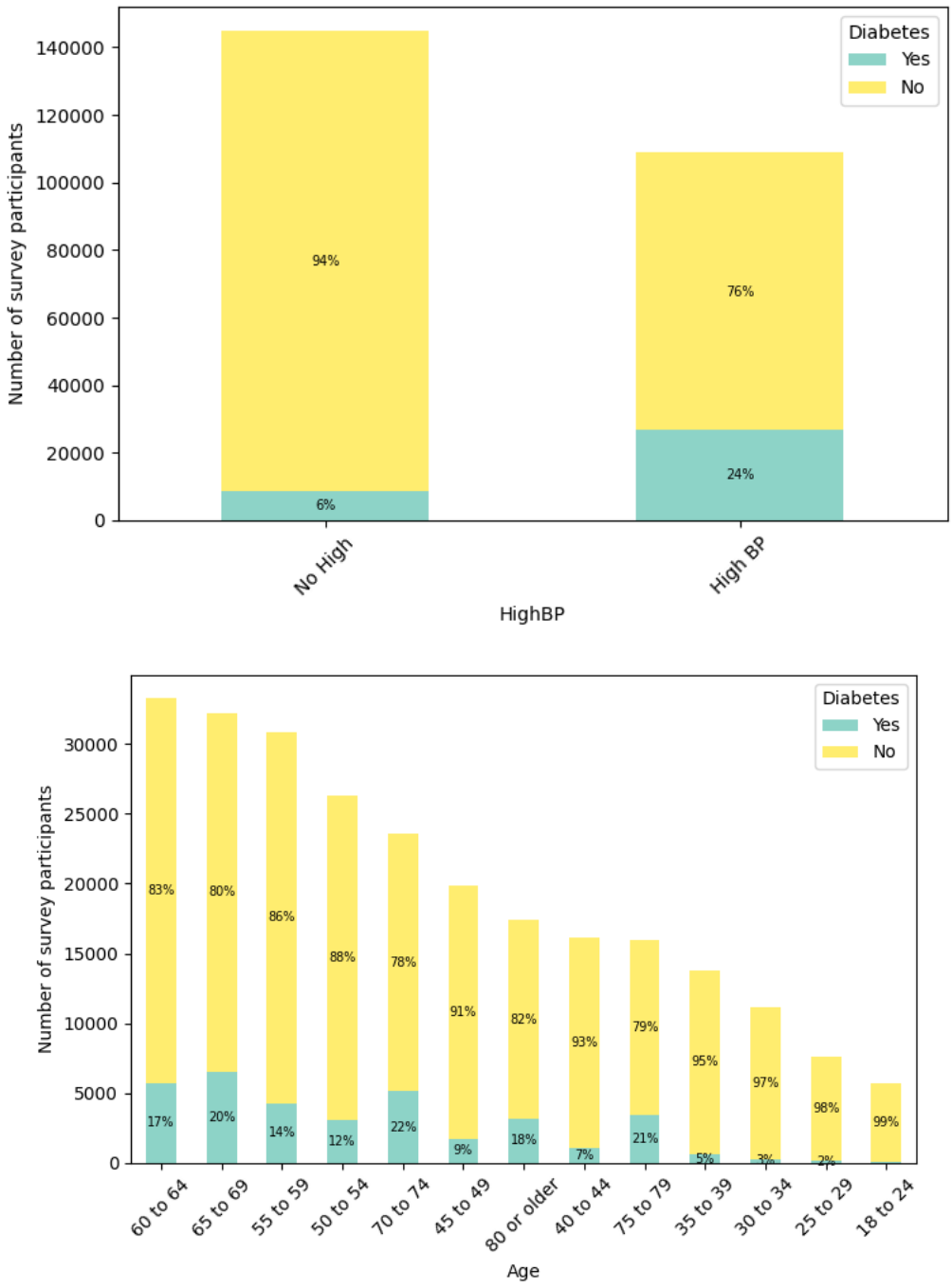


Figure 2: Distributions of Categorical Variables by target

Histograms for Numerical Features: Here, histogram plots visualise the distributions of numerical columns in a dataset by plotting individual histograms for each column in a grid format.

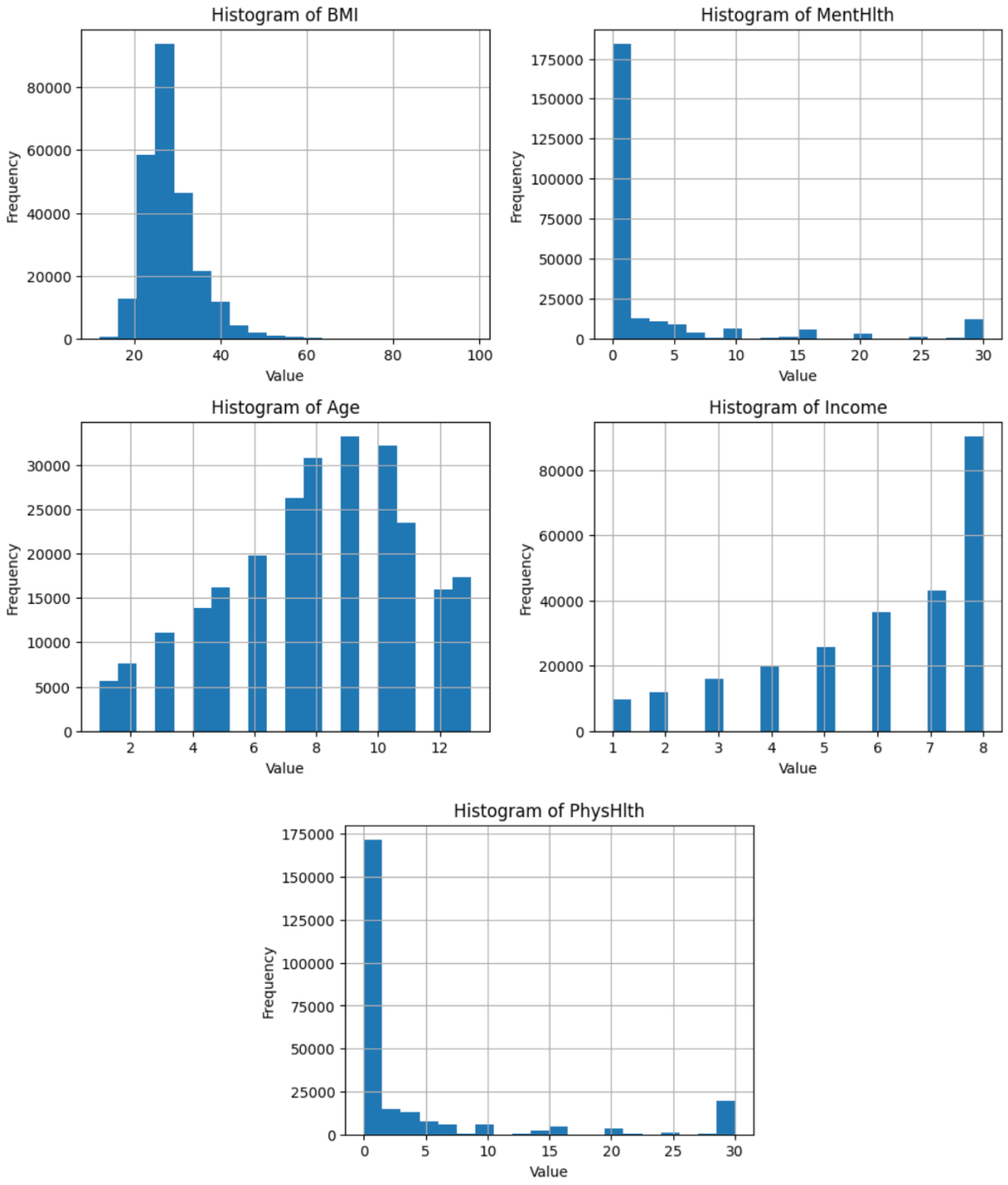


Figure 3: Histograms of Numerical Features

3. Data Preparation

Data Cleaning

- There were no missing values, and duplicate values were identified during the data loading phase. Consideration was given that each survey response was valid and there may be duplicate values; however, these would be assumed to represent a valid record and response in a large national survey sample.
- The dataset does, however, contain extreme values. Outliers were detected from the 'BMI', 'PhysHlth' and 'MentHlth' columns and capped using the Interquartile Range capping method.

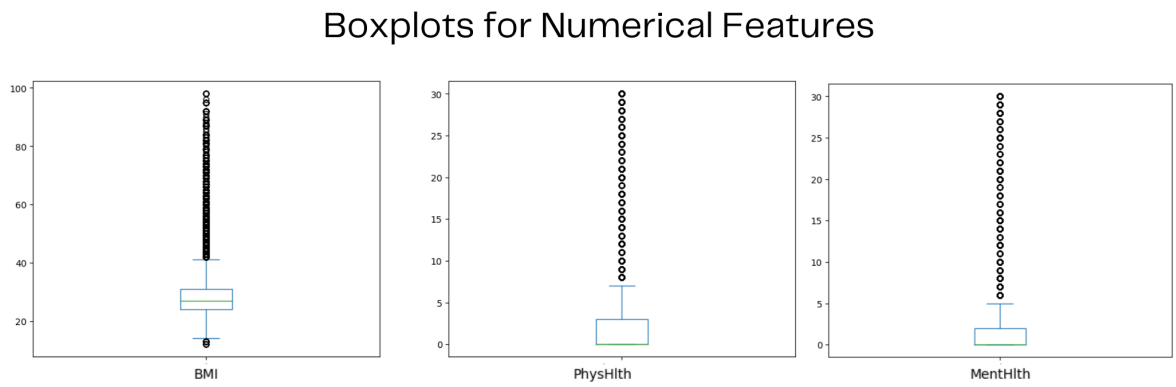


Figure 4: Boxplots for Numerical Features

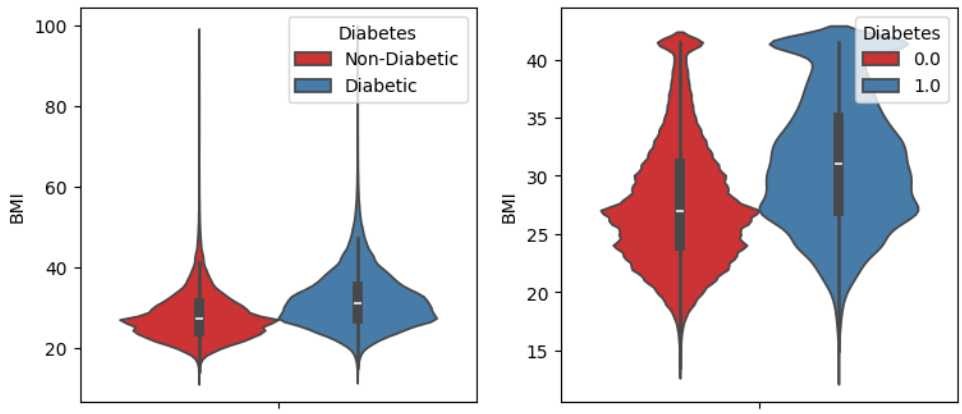


Figure 5: BMI by target variable. With outliers (left) and without outliers (right)

Data Preprocessing

Dealing with Imbalanced Data: The original dataset was imbalanced, meaning there were significantly more non-diabetic examples than diabetic ones. The ‘NearMiss’ algorithm was used to under-sample the majority class (non-diabetic), reducing its instances to be more balanced with the minority class (diabetic).

Feature Selection: A chi-square test was performed to see the correlation of categorical variables with the target variable. Features for the modelling section were chosen using the chi-square test score.

Feature Engineering

No additional feature creation was performed, but feature selection methods were employed to reduce multicollinearity and improve model performance.

4. Modelling

Model Selection Process

Several classification models were considered for predicting diabetes:

- **Logistic Regression**: A baseline model due to its interpretability.
- **Random Forest**: To improve performance by using an ensemble of decision trees.
- **XGBoost**: Applied for its strength in handling imbalanced datasets and providing high accuracy.
- **Model Stacking**: Used to combine the predictions of multiple models (such as Logistic Regression, Random Forest, and XGBoost) to improve the overall predictive performance by leveraging the strengths of each model.

5. Model Evaluation Metrics

Model Performance Comparison

Model	Test Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.8%	91%	82%	86%
Logistic Regression (after tuning)	86.8%	91.2%	81.6%	86.8%
Random Forest	88.2%	94%	82%	87%
Random Forest (after tuning)	88.4%	92.7%	83.4%	88.4%
XGBoost	88.8%	94%	83%	88%
XGBoost (after tuning)	88.9%	93.4%	83.7%	88.9%
Model Stacking	88.9%	92.2%	85%	88.9%

Figure 6: Model Performance

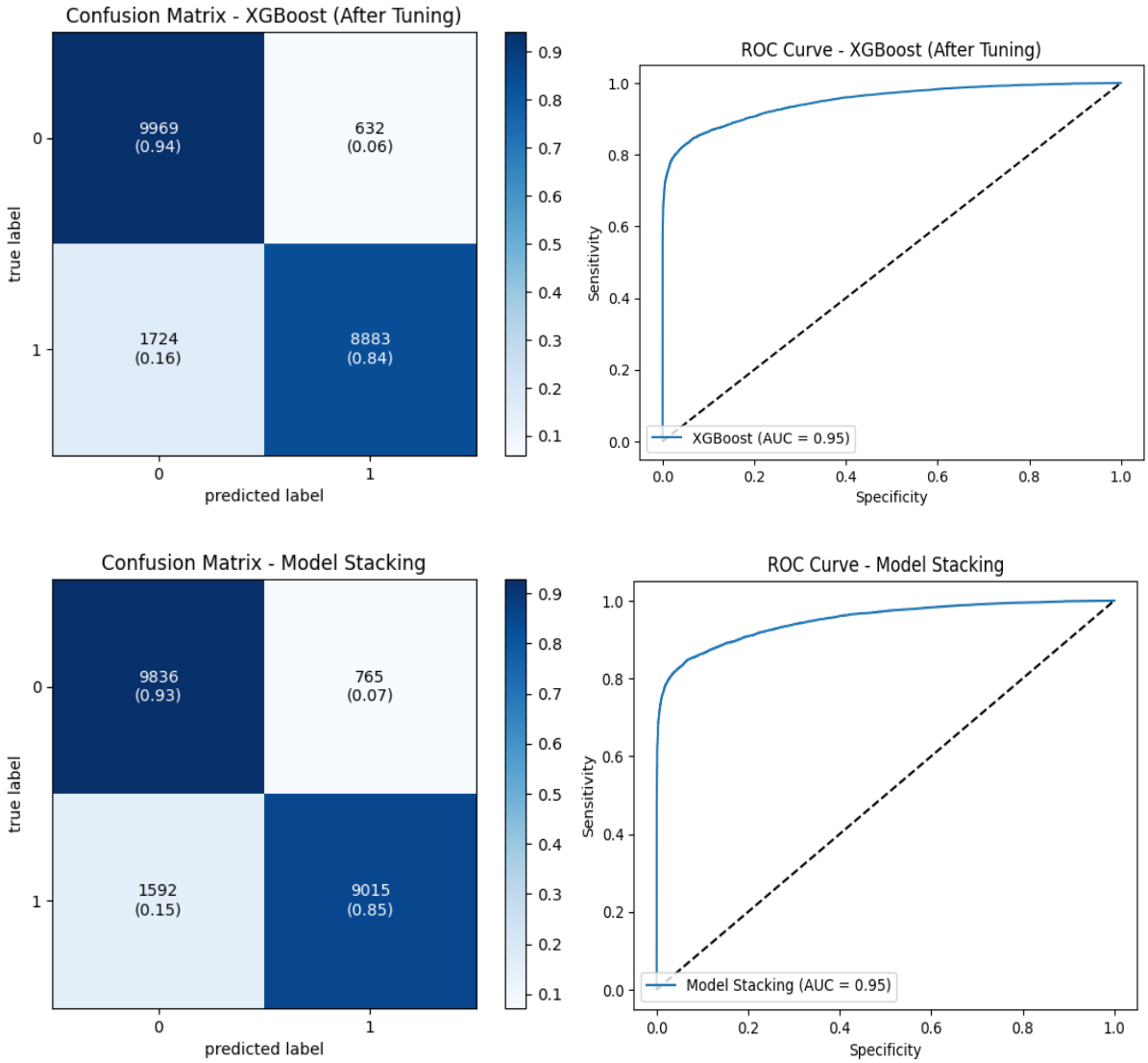


Figure 7: Top two models' performance

6. Evaluation

a. Key Insights

- **Significant Features**: Variables such as BMI, high blood pressure, and cholesterol levels stood out as the most important predictors of diabetes.
- **Model Strength**: Ensemble models, particularly Random Forest and XGBoost, consistently outperformed simpler models like Logistic Regression. Notably, these models excelled in **recall**, a critical metric for identifying individuals at risk of diabetes. Model Stacking also provided strong performance, especially in recall, making it competitive with ensemble models.

b. Business Impact and Benefits

- **Early Detection**: By accurately predicting diabetes, these models enable healthcare providers to initiate early interventions, helping to minimise long-term treatment costs and enhance patient health outcomes.
- **Risk Stratification**: The models can be employed for risk stratification, allowing healthcare systems to focus on high-risk individuals for more intensive screening and preventive actions.

c. Quantified Improvements

- The XGBoost model demonstrated a 3% improvement in recall over Logistic Regression, meaning more individuals with diabetes were correctly identified.
- The overall prediction accuracy increased by 1-2% after model tuning, contributing to better overall performance, particularly for ensemble methods like Random Forest and XGBoost.

7. Conclusion

This analysis confirms that predictive models, especially XGBoost and Random Forest, are highly effective at identifying individuals at risk of diabetes based on simple health indicators. XGBoost achieved the highest F1 score, making it the most reliable model for balancing precision and recall. These models can serve as valuable tools for healthcare professionals to facilitate early diagnoses, ultimately improving patient outcomes and reducing healthcare costs associated with late-stage diabetes treatments. Obtaining additional variables, such as Having anyone in the Family History having Diabetes or racial background, can also play an essential role in developing more accurate models.

References

American Diabetes Association. (2023). Standards of medical care in diabetes—2023. *Diabetes Care*, 46(Supplement 1), S1-S26. https://doi.org/10.2337/dc23-Sintroduction

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Centers for disease control and prevention, & Dane, S. (n.d.). *Behavioral risk factor surveillance system* [Dataset]. Kaggle. Retrieved 22 September 2024, from https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Teboul, A. (n.d.). *Diabetes health indicators dataset* [Dataset]. Kaggle. Retrieved 22 September 2024, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22-30. https://doi.org/10.1109/MCSE.2011.37/

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

World Health Organization. (n.d.). *Diabetes*. Retrieved 20 October 2024, from https://www.who.int/health-topics/diabetes

Appendix

1. Chi-square Test results

	HighBP	HighChol
HighBP	-	Chi2: 22556.72, p: 0.000
HighChol	Chi2: 22556.72, p: 0.000	-
CholCheck	Chi2: 2460.63, p: 0.000	Chi2: 1859.73, p: 0.000
Smoker	Chi2: 2386.06, p: 0.000	Chi2: 2114.20, p: 0.000
Stroke	Chi2: 4257.87, p: 0.000	Chi2: 2175.24, p: 0.000
HeartDiseaseorAttack	Chi2: 11117.88, p: 0.000	Chi2: 8288.02, p: 0.000
PhysActivity	Chi2: 3980.10, p: 0.000	Chi2: 1544.85, p: 0.000
Fruits	Chi2: 417.05, p: 0.000	Chi2: 423.34, p: 0.000
Veggies	Chi2: 951.88, p: 0.000	Chi2: 403.12, p: 0.000
NoDocbcCost	Chi2: 76.31, p: 0.000	Chi2: 44.85, p: 0.000
GenHlth	Chi2: 23417.26, p: 0.000	Chi2: 11184.25, p: 0.000
DiffWalk	Chi2: 12684.12, p: 0.000	Chi2: 5308.70, p: 0.000
Sex	Chi2: 691.21, p: 0.000	Chi2: 246.90, p: 0.000
Education	Chi2: 5189.24, p: 0.000	Chi2: 1276.01, p: 0.000
Diabetes_binary	Chi2: 17562.45, p: 0.000	Chi2: 10174.07, p: 0.000

	CholCheck	Smoker
HighBP	Chi2: 2460.63, p: 0.000	Chi2: 2386.06, p: 0.000
HighChol	Chi2: 1859.73, p: 0.000	Chi2: 2114.20, p: 0.000
CholCheck	-	Chi2: 24.90, p: 0.000
Smoker	Chi2: 24.90, p: 0.000	-
Stroke	Chi2: 147.40, p: 0.000	Chi2: 948.67, p: 0.000
HeartDiseaseorAttack	Chi2: 494.93, p: 0.000	Chi2: 3321.61, p: 0.000
PhysActivity	Chi2: 4.40, p: 0.036	Chi2: 1937.44, p: 0.000
Fruits	Chi2: 144.03, p: 0.000	Chi2: 1529.87, p: 0.000
Veggies	Chi2: 9.42, p: 0.002	Chi2: 238.59, p: 0.000
NoDocbcCost	Chi2: 859.80, p: 0.000	Chi2: 607.38, p: 0.000
GenHlth	Chi2: 561.96, p: 0.000	Chi2: 6777.15, p: 0.000
DiffWalk	Chi2: 417.28, p: 0.000	Chi2: 3803.84, p: 0.000
Sex	Chi2: 123.83, p: 0.000	Chi2: 2225.06, p: 0.000
Education	Chi2: 42.83, p: 0.000	Chi2: 8254.65, p: 0.000
Diabetes_binary	Chi2: 1062.94, p: 0.000	Chi2: 937.06, p: 0.000

	Stroke	HeartDiseaseorAttack	\
HighBP	Chi2: 4257.87, p: 0.000	Chi2: 11117.88, p: 0.000	
HighChol	Chi2: 2175.24, p: 0.000	Chi2: 8288.02, p: 0.000	
CholCheck	Chi2: 147.40, p: 0.000	Chi2: 494.93, p: 0.000	
Smoker	Chi2: 948.67, p: 0.000	Chi2: 3321.61, p: 0.000	
Stroke	-	Chi2: 10450.58, p: 0.000	
HeartDiseaseorAttack	Chi2: 10450.58, p: 0.000	-	
PhysActivity	Chi2: 1212.26, p: 0.000	Chi2: 1932.63, p: 0.000	
Fruits	Chi2: 45.34, p: 0.000	Chi2: 99.22, p: 0.000	
Veggies	Chi2: 428.49, p: 0.000	Chi2: 388.82, p: 0.000	
NoDocbcCost	Chi2: 306.65, p: 0.000	Chi2: 243.40, p: 0.000	
GenHlth	Chi2: 9596.37, p: 0.000	Chi2: 19008.16, p: 0.000	
DiffWalk	Chi2: 7906.30, p: 0.000	Chi2: 11475.80, p: 0.000	
Sex	Chi2: 2.22, p: 0.136	Chi2: 1879.79, p: 0.000	
Education	Chi2: 1557.70, p: 0.000	Chi2: 2589.79, p: 0.000	
Diabetes_binary	Chi2: 2838.92, p: 0.000	Chi2: 7971.16, p: 0.000	

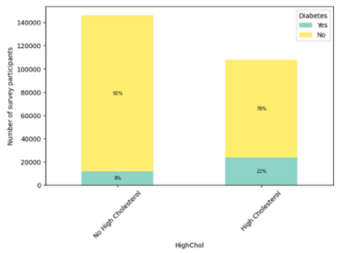
	PhysActivity	Fruits	\
HighBP	Chi2: 3980.10, p: 0.000	Chi2: 417.05, p: 0.000	
HighChol	Chi2: 1544.85, p: 0.000	Chi2: 423.34, p: 0.000	
CholCheck	Chi2: 4.40, p: 0.036	Chi2: 144.03, p: 0.000	
Smoker	Chi2: 1937.44, p: 0.000	Chi2: 1529.87, p: 0.000	
Stroke	Chi2: 1212.26, p: 0.000	Chi2: 45.34, p: 0.000	
HeartDiseaseorAttack	Chi2: 1932.63, p: 0.000	Chi2: 99.22, p: 0.000	
PhysActivity	-	Chi2: 5169.11, p: 0.000	
Fruits	Chi2: 5169.11, p: 0.000	-	
Veggies	Chi2: 5949.10, p: 0.000	Chi2: 16409.20, p: 0.000	
NoDocbcCost	Chi2: 963.29, p: 0.000	Chi2: 496.23, p: 0.000	
GenHlth	Chi2: 18742.97, p: 0.000	Chi2: 2840.85, p: 0.000	
DiffWalk	Chi2: 16258.57, p: 0.000	Chi2: 592.81, p: 0.000	
Sex	Chi2: 267.50, p: 0.000	Chi2: 2108.42, p: 0.000	
Education	Chi2: 10333.65, p: 0.000	Chi2: 3398.93, p: 0.000	
Diabetes_binary	Chi2: 3539.42, p: 0.000	Chi2: 421.61, p: 0.000	

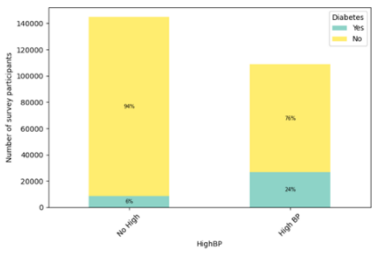
	Veggies	NoDocbcCost	\
HighBP	Chi2: 951.88, p: 0.000	Chi2: 76.31, p: 0.000	
HighChol	Chi2: 403.12, p: 0.000	Chi2: 44.85, p: 0.000	
CholCheck	Chi2: 9.42, p: 0.002	Chi2: 859.80, p: 0.000	
Smoker	Chi2: 238.59, p: 0.000	Chi2: 607.38, p: 0.000	
Stroke	Chi2: 428.49, p: 0.000	Chi2: 306.65, p: 0.000	
HeartDiseaseorAttack	Chi2: 388.82, p: 0.000	Chi2: 243.40, p: 0.000	
PhysActivity	Chi2: 5949.10, p: 0.000	Chi2: 963.29, p: 0.000	

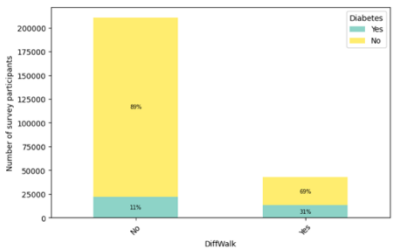
	Sex	Education	\
HighBP	Chi2: 691.21, p: 0.000	Chi2: 5189.24, p: 0.000	
HighChol	Chi2: 246.90, p: 0.000	Chi2: 1276.01, p: 0.000	
CholCheck	Chi2: 123.83, p: 0.000	Chi2: 42.83, p: 0.000	
Smoker	Chi2: 2225.06, p: 0.000	Chi2: 8254.65, p: 0.000	
Stroke	Chi2: 2.22, p: 0.136	Chi2: 1557.70, p: 0.000	
HeartDiseaseorAttack	Chi2: 1879.79, p: 0.000	Chi2: 2589.79, p: 0.000	
PhysActivity	Chi2: 267.50, p: 0.000	Chi2: 10333.65, p: 0.000	
Fruits	Chi2: 2108.42, p: 0.000	Chi2: 3398.93, p: 0.000	
Veggies	Chi2: 1063.73, p: 0.000	Chi2: 6242.90, p: 0.000	
NoDocbcCost	Chi2: 511.81, p: 0.000	Chi2: 2959.08, p: 0.000	
GenHlth	Chi2: 125.76, p: 0.000	Chi2: 22965.53, p: 0.000	
DiffWalk	Chi2: 1253.29, p: 0.000	Chi2: 9862.88, p: 0.000	
Sex	-	Chi2: 468.44, p: 0.000	
Education	Chi2: 468.44, p: 0.000	-	
Diabetes_binary	Chi2: 250.41, p: 0.000	Chi2: 4027.11, p: 0.000	

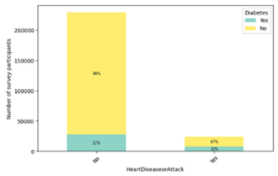
	Diabetes_binary
HighBP	Chi2: 17562.45, p: 0.000
HighChol	Chi2: 10174.07, p: 0.000
CholCheck	Chi2: 1062.94, p: 0.000
Smoker	Chi2: 937.06, p: 0.000
Stroke	Chi2: 2838.92, p: 0.000
HeartDiseaseorAttack	Chi2: 7971.16, p: 0.000
PhysActivity	Chi2: 3539.42, p: 0.000
Fruits	Chi2: 421.61, p: 0.000
Veggies	Chi2: 811.81, p: 0.000
NoDocbcCost	Chi2: 250.31, p: 0.000
GenHlth	Chi2: 22728.07, p: 0.000
DiffWalk	Chi2: 12092.32, p: 0.000
Sex	Chi2: 250.41, p: 0.000
Education	Chi2: 4027.11, p: 0.000
Diabetes_binary	-

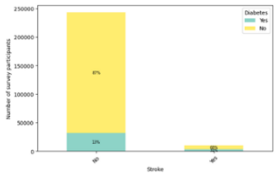
2. All Categorical bar charts divided into diabetes in percentages

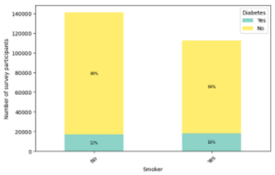


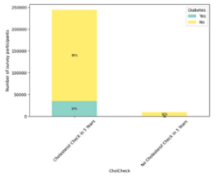


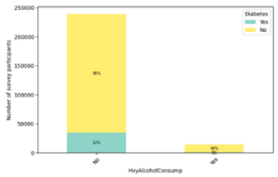


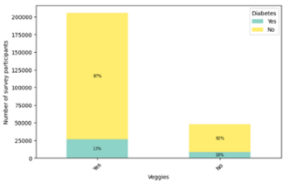


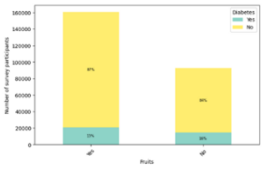


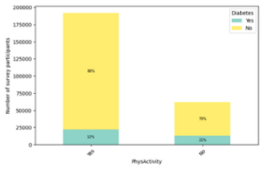


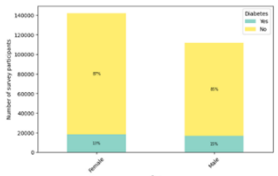


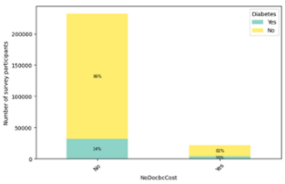


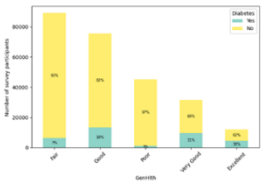


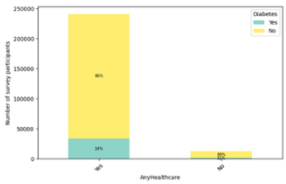


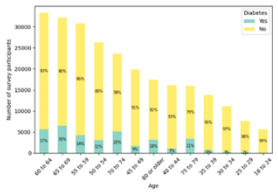


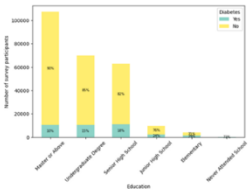


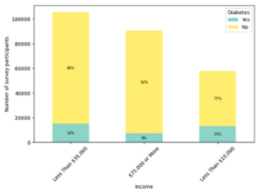


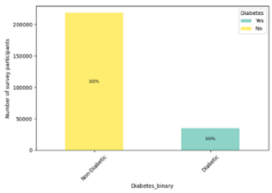












3. Model Performances