Diabetes Diagnosis and Prediction using Health Indicators

CSS 590 A Su 24: Special Topics In Computing And Software Systems

MLBLearners
Poulami Das Ghosh
Dhurka Rohini M

Final Report

University of Washington
16th August, 2024

Table of Contents

List Of Figures

List Of Tables

Introduction

Diabetes is one of the most prevalent chronic illnesses in the United States, significantly impacting public health[1]. This condition arises when the body fails to produce sufficient insulin to manage blood sugar levels, leading to a range of serious health complications[2]. In 2018, an estimated 34.3 million individuals across all age groups in the U.S., accounting for 10.5% of the population, were living with diabetes[3]. This widespread prevalence heightens the risk of related health issues, including heart disease, cardiovascular incidents, microvascular damage, retinopathy causing blindness, kidney failure and even early mortality. The disease's extensive impact underscores its threat to individual health in the U.S. and the world. The Behavioural Risk Factor Surveillance System (BRFSS), the largest telephone-based health survey in the nation, conducts over 400,000 interviews annually[4]. This system's main objective is to gather data on health-related behaviours, chronic conditions like diabetes, and the utilization of preventive healthcare services among U.S. residents. The primary objective of this project is to provide data-driven insights by employing machine learning techniques to address the research question: “Which specific health, lifestyle, and socio-economic factors significantly contribute to the diagnosis of diabetes in individuals?”

This project involves the analysis of the 2015 BRFSS Diabetes Health Indicators dataset, which contains 21 health indicators potentially linked to diabetes. The motivation behind this research is to identify the indicators that are strongly correlated with diabetes and to develop predictive models that could aid in early diagnosis and intervention. The dataset includes 23,580 records from the 2015 BRFSS survey. Diabetes is generally classified into two primary types: type 1 and type 2, which represent 5% and 95% of cases, respectively[3]. The dataset used in this project does not distinguish between type 1 and type 2 diabetes. Given the overwhelming prevalence of type 2 diabetes, the findings are more likely to pertain to type 2. One limitation of this dataset is the imbalance between the non-diabetic and diabetic classes, with 82.7% of the data representing non-diabetic individuals and 17.3% representing those with diabetes. Within the diabetic class, 2% of the instances are prediabetic and 15.3% are diabetic. Due to the low number of prediabetic cases and for simplification, we have combined the prediabetic and diabetic instances into a single class. Nonetheless, the class imbalance remains a significant challenge in this dataset.

Understanding the factors that contribute to diabetes diagnosis is crucial for developing effective prevention and management strategies. This serves as a central motivation for investigating various health indicators. By leveraging advanced machine learning techniques on this comprehensive dataset, the project aims to offer valuable insights for healthcare providers, patients, and public health officials.

In this endeavour, we have developed and evaluated several predictive models, including Random Forest, Light Gradient Boosting Machine (LightGBM), K-Nearest Neighbours (KNN), and XGBoost. These models are built on training data that integrates correlation analysis, the Chi-square test for feature selection, and the SMOTE-ENN technique to address class imbalance. The models exhibit high precision and recall for both non-diabetic and diabetic classes comparable to state-of-art model of Ren [13] , demonstrating their effectiveness in accurately identifying instances of each class. Among them, the Random Forest model performs best, achieving an accuracy of 97.2% on the dataset.

The performance of these predictive classifiers is assessed using accuracy, recall, F1 score, and confusion matrix metrics on the training set, ensuring robust performance across both non-diabetic and diabetic classes. Special emphasis is given to recall over other matrix as this is medical data[23]. Furthermore, interpretability analysis highlights key factors such as obesity, general health, age, high cholesterol, high blood pressure, physical activity, and income as significant contributors to diabetes diagnosis.

Related Work

Numerous studies have explored the application of machine learning in the diagnosis and prediction of diabetes. For instance, Kavakiotis et al. [8] provided a comprehensive review of machine learning and data mining approaches in diabetes research, highlighting the potential of these techniques in improving diagnosis, management, and the prediction of diabetes-related complications. Similarly, Sisodia et al. [9] developed a predictive model using decision trees, achieving significant accuracy in diagnosing diabetes from patient data. In another study, Ali et al. [10] employed a combination of support vector machines (SVM) and logistic regression to classify diabetes risk, demonstrating the efficacy of hybrid models in enhancing predictive accuracy. More recently, Razzak et al. [11] utilized deep learning techniques, specifically convolutional neural networks (CNNs), to predict diabetes mellitus from electronic health records, showcasing advancements in leveraging complex neural networks for medical diagnosis.

Xie et al. [12] and Ren [13] both conducted studies using the US Diabetes Health Indicators Dataset. Xie et al. [12] implemented multiple models, including support vector machines and artificial neural networks, to investigate the associations between potential risk factors and type 2 diabetes. The neural network model achieved the highest accuracy (82.4%), specificity (90.2%), and AUC (0.7949). However, these models exhibited an imbalance in class predictions, indicating significant room for improvement in terms of accuracy and overall performance. Similarly, Ren [13] introduced an innovative feature selection method, applying the Chi-square test to explore the association between health indicators and diabetes. Several machine learning models were developed to predict the disease, with the CatBoost Classifier being selected for its 86.6% accuracy on the testing set achieving a state-of-art status. This study also employed the SMOTE technique to address data imbalance, resulting in better-balanced class predictions compared to Xie et al. [12]. Nevertheless, the overall accuracy of 86.6% still leaves considerable scope for enhancement.

Lamari et al. [14] proposed a novel approach for subset selection in the classification of imbalanced medical datasets. Their method combines an improved dynamic ensemble selection, called the META-DES framework, with a hybrid sampling technique known as SMOTE-ENN. Experimental results demonstrated the superiority of this ensemble learning system across three UCI datasets.

In our project, we leveraged the combination of SMOTE-ENN for class balancing and the Chi-square test for feature selection inspired by Lamari et al. [14] and Ren [13]. This approach led to the development of multiple models, achieving over 92% accuracy. Our best-performing model, a random forest classifier, attained an accuracy of 97.2% on this dataset. Additionally, we employed global and local interpretability methods to reinforce our findings, identifying the top contributors to diabetes in alignment with our research objectives.

About The Data

The US Diabetes Health Indicators Dataset was originally collected by the Behavioural Risk Factor Surveillance System (BRFSS) and later cleaned and published by Teboul on Kaggle[5] . This dataset consists of 23,580 records from adult respondents across the United States and contains a total of 22 variables, including 21 feature variables and 1 target variable indicating diabetes status. Originally, 0 indicated “non-diabetic” , 1 indicate “prediabetic” and 2 indicates “diabetic”. After merging, the binary target variable is coded as "1" for individuals who are either diabetic or prediabetic, and "0" for those who are non-diabetic. The dataset exhibits a class imbalance, with 82.7% of the records representing non-diabetic individuals and 17.3% representing diabetic or prediabetic individuals, as illustrated in Figure 1. The 21 features in the dataset include both numeric and categorical variables. The only numeric continuous variable is BMI, which has been converted into categorical values using the World Health Organization's (WHO) classification criteria[6]:

- **Below 18.5**: Underweight
- **18.5—24.9**: Healthy Weight
- **25.0—29.9**: Overweight
- **30.0 and Above**: Obesity

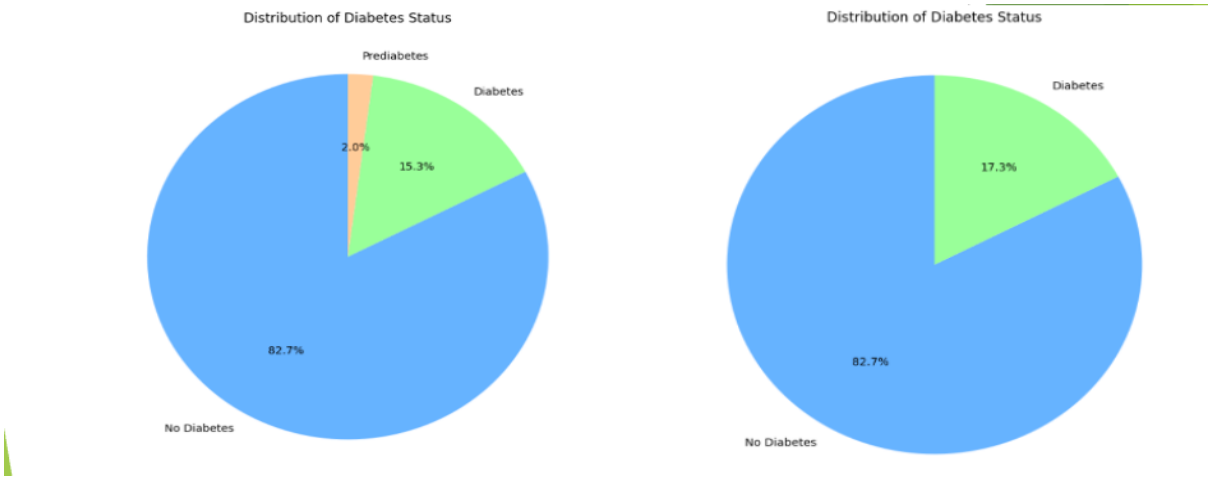

Figure 1: Distribution of classes before and after merging of diabetic and prediabetic group

Figure 2 illustrates the distribution of data for each BMI categories and figure 3 demonstrate the distribution of data for all the columns. The remaining 20 categorical features are grouped into four categories:

1. **Socio- Economic Information**: Age, gender, education level, income, access to healthcare services, and financial barriers to seeing a doctor.
2. **Physical Disease Indicators**: High blood pressure, high cholesterol, cholesterol check history, history of stroke or heart disease, and difficulty walking.
3. **Self-Assessed Health Status**: General health, mental health, and physical health.
4. **Personal Habits**: Physical activity, smoking status, fruit and vegetable consumption, and heavy alcohol consumption.

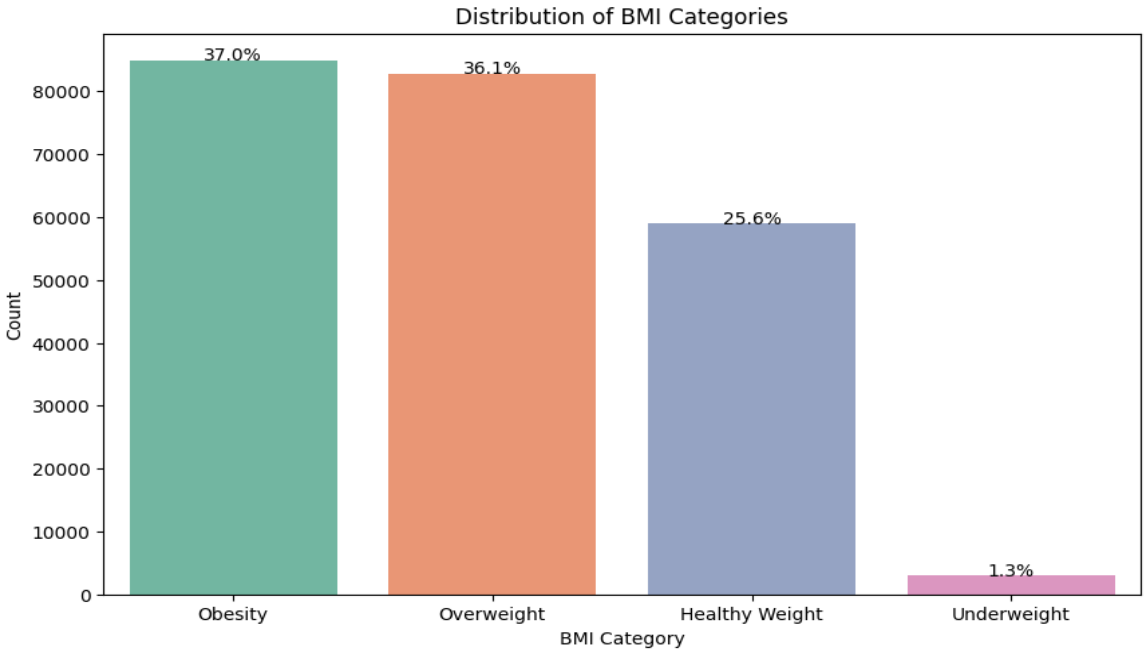

Figure 2:Distribution Of data across BMI Category

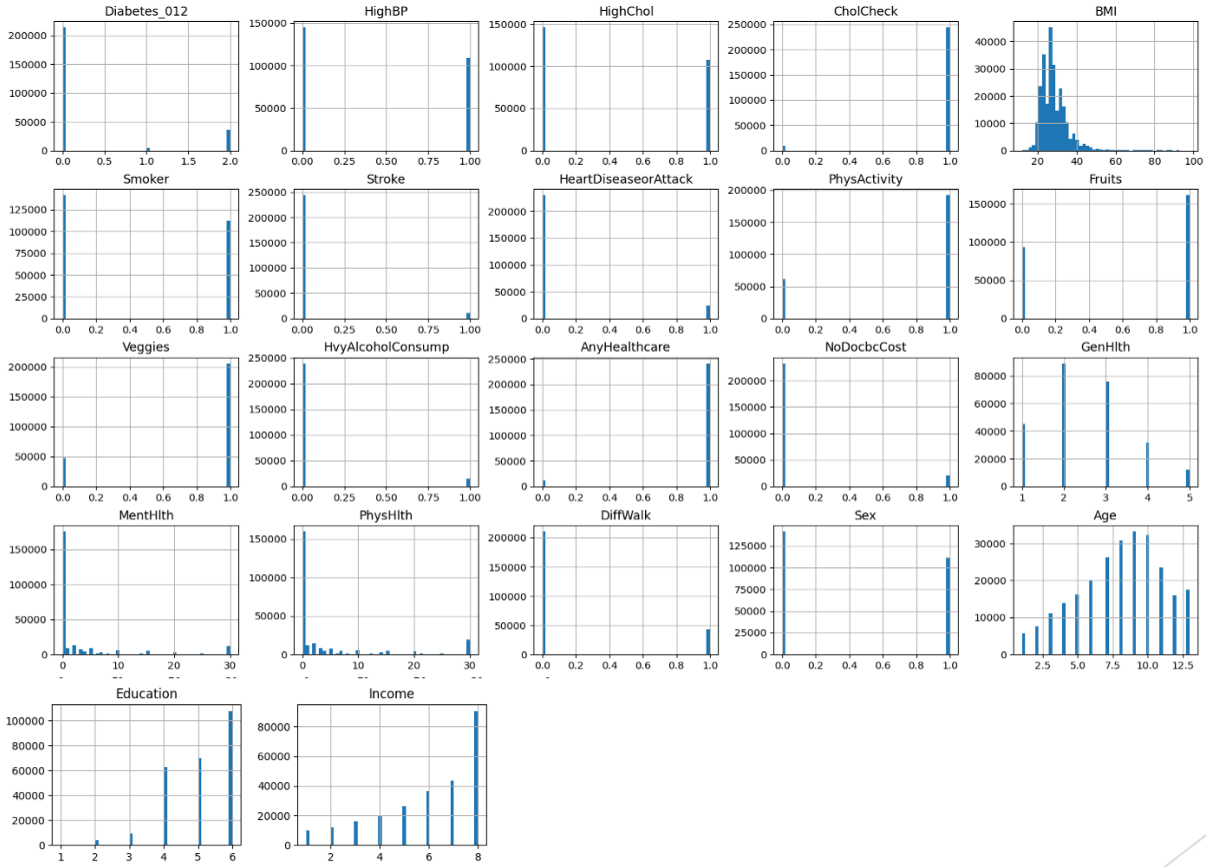

Figure 3:Data distribution across feature set

Methodology

Architecture & Technologies

For this project, we predominantly used Python and Google Colab, with report documentation managed in Microsoft Excel. GitHub was employed for code management. During the initial phase, we leveraged the GPU capabilities of Google Colab to run experiments. However, the final implementation does not require GPU, ensuring that the project remains accessible and efficient, even with limited resources. This consideration is particularly important for potential future integration with web or mobile applications.

This architecture outlines the machine learning approach to developing a predictive model for diagnosing diabetic patients based on their health, lifestyle, and socio-economic conditions. Minimal effort was required for data collection, as the US Diabetes Health Indicators Dataset was originally collected by the Behavioral Risk Factor Surveillance System (BRFSS) and later cleaned and published by Teboul on Kaggle [5]. For this project, we structured our efforts into four major areas: Data Cleansing & Exploratory Data Analysis (EDA), Feature selection, feature engineering, class balancing, model building, Evaluate and Interpretion.

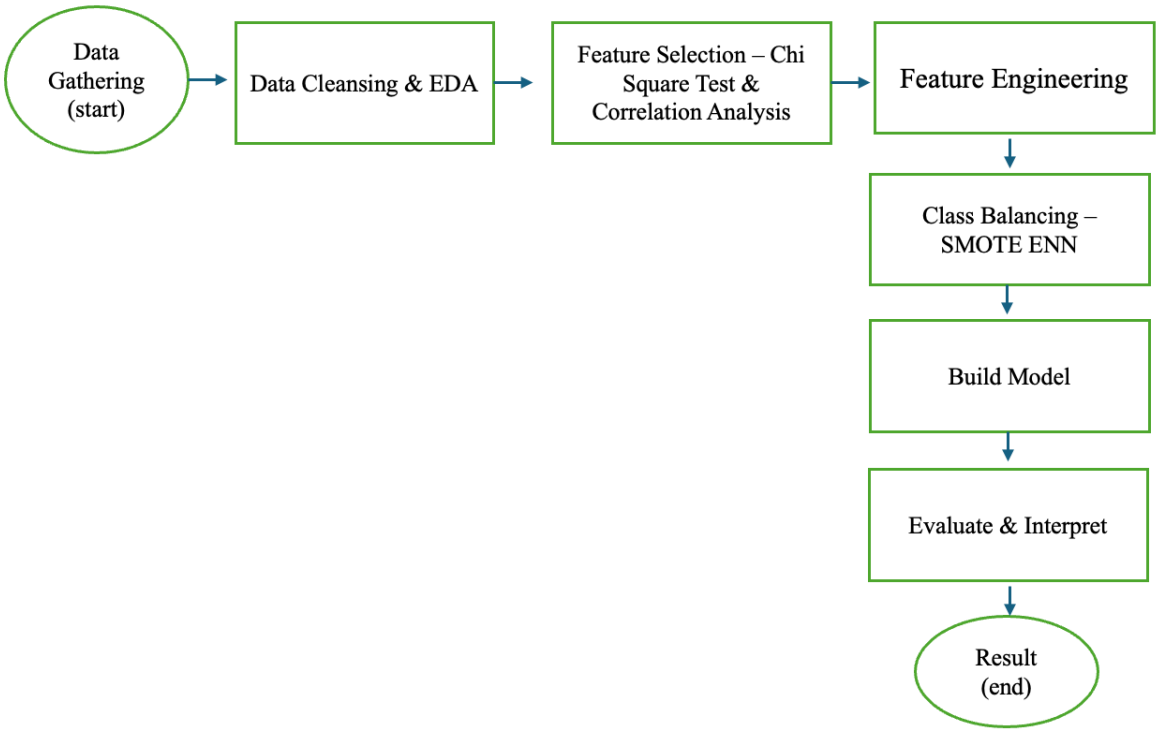

Figure 4:Process flow of the project

Data Cleansing & Exploratory Data Analysis (EDA)

Data Cleansing & EDA

The first step in our data cleansing process was de-duplication, during which we removed 23,899 duplicate records from the US Diabetes Health Indicators Dataset. We then conducted a null value analysis and were fortunate to find that the dataset contained no missing values, eliminating the need for any imputation techniques.

Following this, we proceeded with Exploratory Data Analysis (EDA) with a dual focus on feature selection and data interpretability. For the initial EDA, we utilized the ydata-profiling library's out-of-the-box methods to generate a comprehensive report on the dataset. This report provided a general overview, including statistical information, visualizations, missing and null value analysis, cardinality analysis, correlation analysis, and guardrail alerts. This approach was highly advantageous in gaining a complete understanding of the dataset, which helped us formulate our strategy for efficiently handling of the data.

Feature Selection

The next step involved feature selection, which was carried out by integrating the results from the Chi-Square test, correlation matrix analysis, and Y-Data guardrail alerts. The Chi-Square test [15] is a statistical method used to assess whether there is a significant association between two categorical variables. It compares the observed frequencies in a contingency table with the expected frequencies, assuming the variables are independent.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency in the ith category
- E_i = Expected frequency in the ith category

Figure 5 Chi Square Test formula

Figure 6 presents the results of the Chi-square test applied to this dataset. Features with low Chi-square scores indicate a lack of association with the dependent variable (Diabetes_012), implying that they do not significantly contribute to diabetes prediction. We excluded all features with a Chi-square score lower than 200 [13]. Additionally, features were further removed based on high correlation values with one another. A correlation matrix, which displays the correlation coefficients between multiple variables, was utilized to identify the strength and direction of linear relationships. Correlation coefficients range from +1 (perfect positive correlation) to -1 (perfect negative correlation), with 0 indicating no correlation [22].

In the exploratory data analysis (EDA) phase, correlation matrices were instrumental in uncovering relationships between variables, detecting multicollinearity, and guiding feature selection by identifying highly correlated variables that could contribute redundant information to the model. This refinement process enhances both the efficiency and accuracy of the modelling. For example, as depicted in Figure 7, Education and Income exhibited high

correlation; hence, retaining only one of these features sufficed for maintaining predictive accuracy.

Moreover, based on guardrail alerts indicating that certain columns had over 90% zero entries, we excluded non-significant columns, such as CholCheck, HvyAlcoholConsump, MentHlth, and PhysHlth, as they were likely to be irrelevant or comprised of poor-quality data. Additionally, we removed columns like Stroke, HeartAttack, and DiffWalk, as these conditions are typically consequences of diabetes rather than causative factors [21]. Through this methodical approach, we successfully reduced the feature set from 21 to 11 features, streamlining the data for more effective model training and analysis.

Chi-Square Test Results:

Feature	Chi-Square Score
GenHlth	18893.5
HighBP	15573.2
DiffWalk	9951.88
HighChol	9600.74
Age	8877.51
bmi_Obesity	8686.56
HeartDiseaseorAttack	6421.03
PhysHlth	6250.03
Income	5115.86
bmi_Healthy Weight	4984.12
Education	2832.88
PhysActivity	2408.39
Stroke	2191.4
CholCheck	1332.84
MentHlth	1254.43
HvyAlcoholConsump	1003.56
bmi_Overweight	659.759
Smoker	491.535
Veggies	431.299
bmi_Underweight	253.298
Sex	217.779
NoDocbcCost	154.38
Fruits	143.108
AnyHealthcare	129.396

Figure 6 Chi Square Test Result

Alerts

`bmi_Underweight` is highly imbalanced (89.8%)	Imbalance
`Diabetes_012` has 190055 (82.7%) zeros	Zeros
`HighBP` has 125359 (54.6%) zeros	Zeros
`HighChol` has 128273 (55.8%) zeros	Zeros
`CholCheck` has 9298 (4.0%) zeros	Zeros
`Smoker` has 122781 (53.4%) zeros	Zeros
`Stroke` has 219497 (95.5%) zeros	Zeros
`HeartDiseaseorAttack` has 206064 (89.7%) zeros	Zeros
`PhysActivity` has 61270 (26.7%) zeros	Zeros
`Fruits` has 88933 (38.7%) zeros	Zeros
`Veggies` has 47148 (20.5%) zeros	Zeros
`HvyAlcoholConsump` has 215831 (93.9%) zeros	Zeros
`AnyHealthcare` has 12391 (5.4%) zeros	Zeros
`NoDocbcCost` has 208455 (90.7%) zeros	Zeros
`MentHlth` has 152623 (66.4%) zeros	Zeros
`PhysHlth` has 136877 (59.6%) zeros	Zeros
`DiffWalk` has 187155 (81.4%) zeros	Zeros
`Sex` has 128854 (56.1%) zeros	Zeros

Figure 7:Y-Data Guardrails

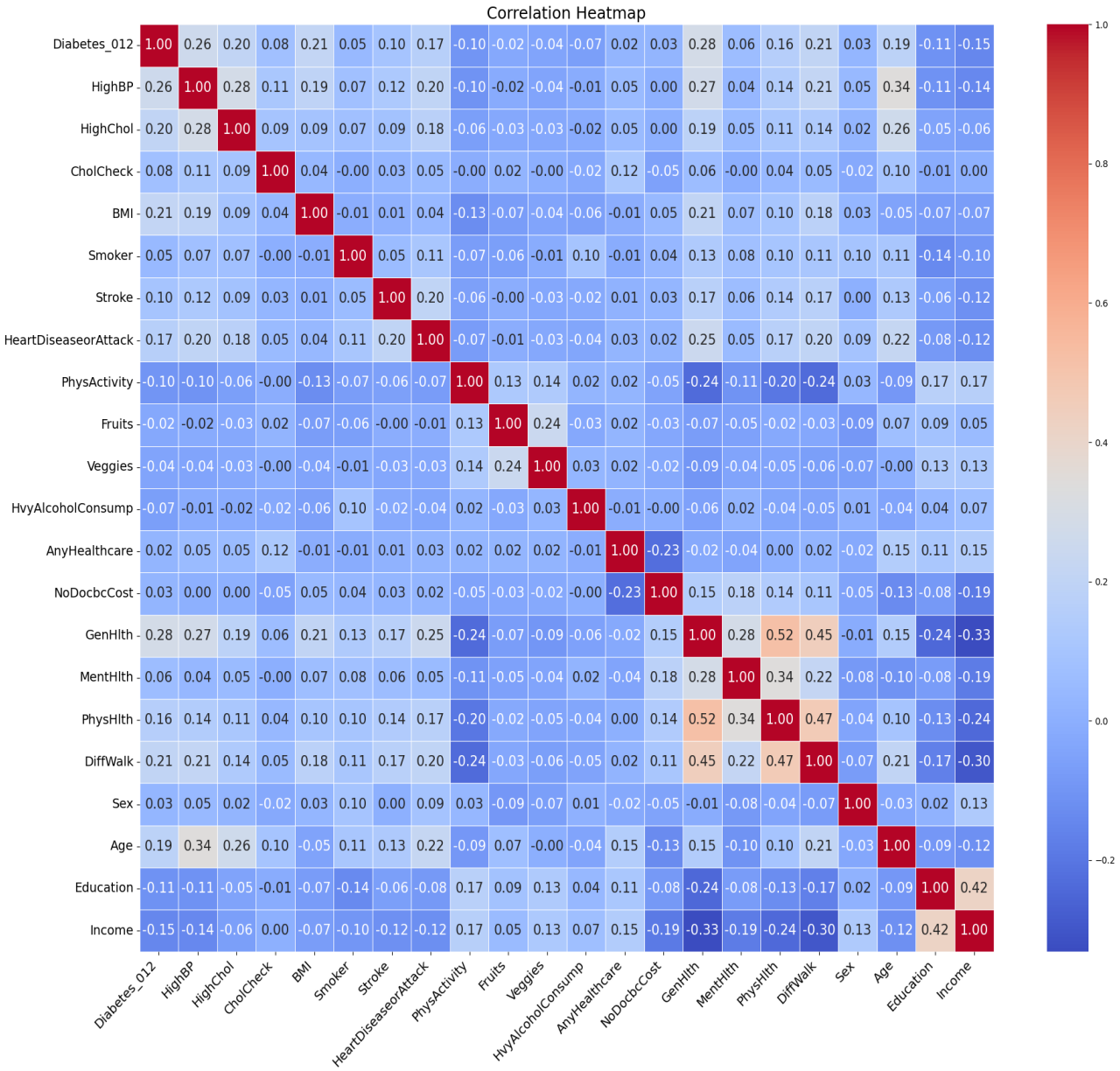

Figure 8 Corelation Heat Map

EDA

The correlation heat map(figure 8) provides valuable insights into the relationships between various features in the dataset. Notably, Education and Income exhibit a strong positive correlation, which is expected since income generally increases with higher levels of education. Additionally, General Health, Mental Health, and Physical Health all show negative correlations with Income and Education. This suggests that as income and education levels rise, overall health improves. It is essential to note that within this dataset, lower values for General, Mental, and Physical Health (before feature engineering) represent better health, aligning with the expectation that higher income and education levels lead to greater health awareness and access to superior healthcare services. Furthermore, Age is highly correlated with both High Blood Pressure and High Cholesterol, which is consistent with the fact that these conditions tend to become more prevalent with increasing age.

Some of the other insights based on further data analysis are as follows:

- People with high BP and high Cholesterol are more likely to be diabetic as illustrated in Figure 9
- General Health encompasses overall well-being, including factors such as stress and sleep. The scale for General Health ranges from 1 to 5, where a value of 1 represents poor health and a value of 5 represents excellent health (reversed to make it more intuitive during feature engineering). Therefore, as the scale increases, general health improves. In Figure 10, a clear trend is observed: as General Health decreases, the occurrence of diabetes increases.
- Figure 10 also illustrates that as income increases, the likelihood of diabetes decreases. This trend may be attributed to a higher standard of living, access to better healthcare facilities, and improved dietary habits, such as avoiding cheaper, less nutritious fast food.
- Diabetes occurrences increase with age [16], and this trend is also supported by our dataset. Figure 11 highlights how the proportion of diabetic individuals rises with advancing age.
- The effect of weight (BMI) on blood sugar is well documented in multiple studies [17]. And as expected out data also support this hypothesis in Figure 11.

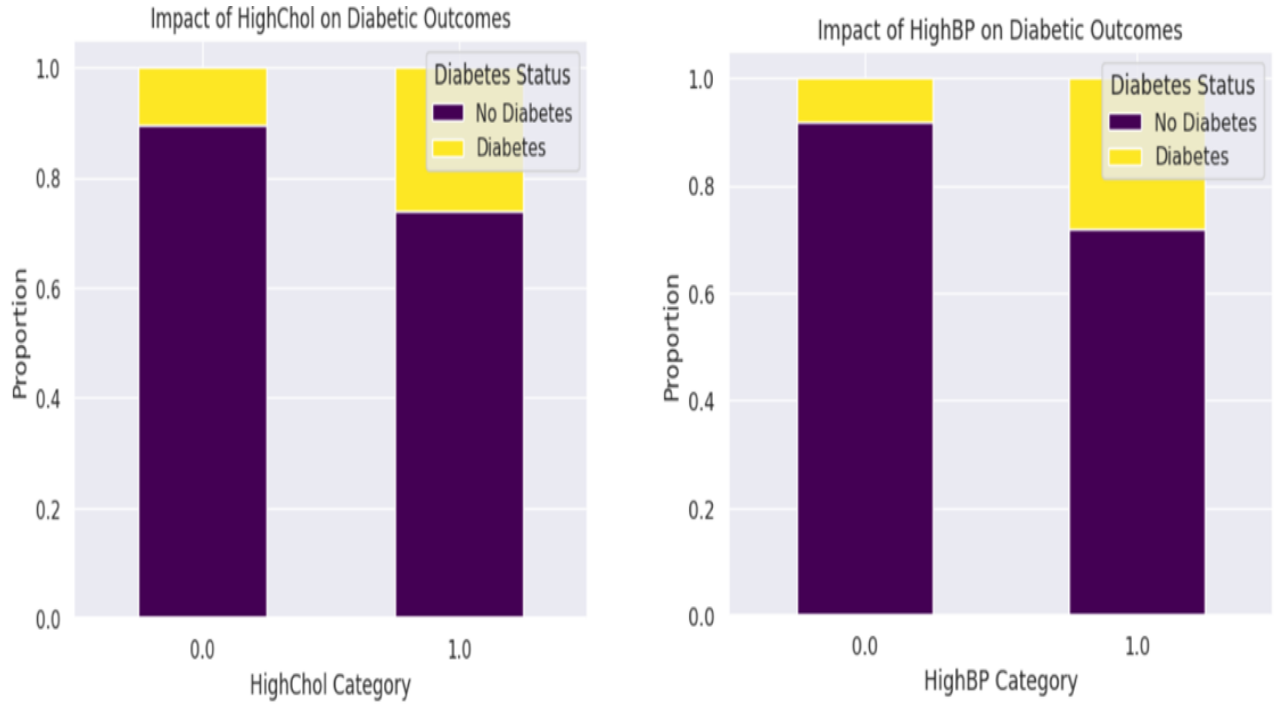

Figure 9: Data distribution based on High BP and cholesterol

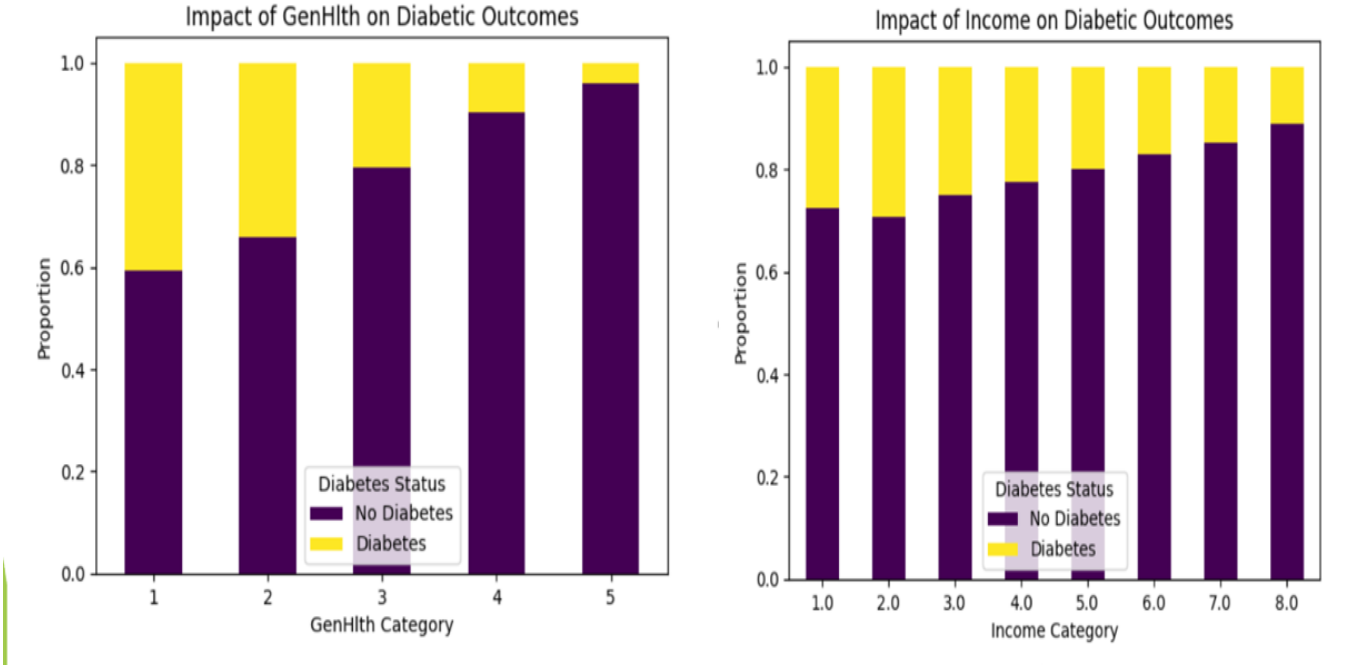

Figure 10:Data distribution based on General health and Income

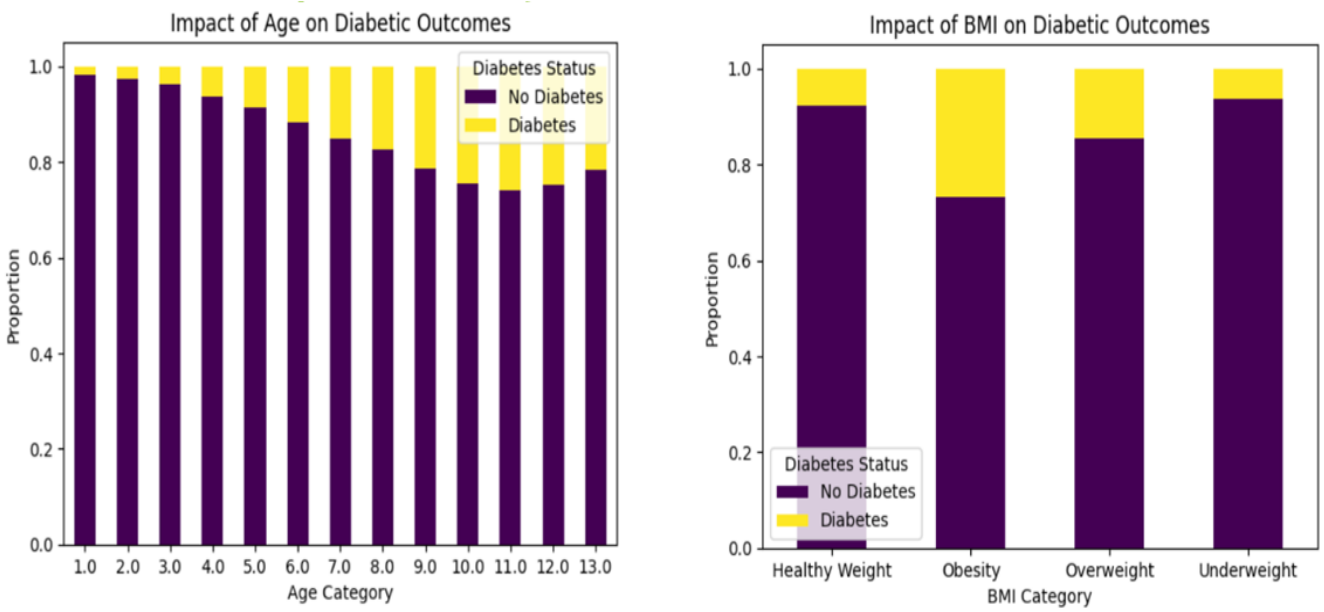

Figure 11:Data distribution based on Age and BMI

Feature Engineering & Class Balancing

The data pre-processing phase we perform two main steps: feature engineering & class balancing.

In the feature engineering phase, we categorized BMI values based on the World Health Organization's (WHO) classification criteria [6]. BMI values were converted into categorical classes as follows: Below 18.5 is classified as Underweight, 18.5—24.9 as Healthy Weight,

25.0—29.9 as Overweight, and 30.0 and Above as Obesity. This categorization helps in better understanding and interpreting the impact of BMI on diabetes.

We also applied One-Hot Encoding to features such as BMI. One-Hot Encoding is a technique used to convert categorical variables into a numerical format suitable for machine learning algorithms[22]. It involves creating binary columns for each category of a feature, where each column represents a specific category and is marked as 1 if the category is present and 0 otherwise.

To make the General Health scale more intuitive, we reversed its scale so that higher values represent better health. This adjustment aids in clearer interpretation and analysis of the data. Additionally, we merged the prediabetic and diabetic instances into a single class. Within the dataset, 2% of the instances are prediabetic, and 15.3% are diabetic. Given the low number of prediabetic cases, combining them with the diabetic class simplifies the classification process. Despite this merger, class imbalance remains a significant challenge in the dataset.

Before commencing the class balancing phase, we split the data in the ratio of 80:20 and set 20% of the data aside for testing. The data augmentation is then performed on the train set to address class imbalance. We used SMOTE-ENN in our project (Lamari et al. [14]) as it offers a more robust solution to class imbalance, leading to better model performance and more reliable predictions. SMOTE can sometimes create overlapping samples or introduce noise, which can degrade model performance. SMOTE-ENN improves upon this by integrating SMOTE with Edited Nearest Neighbours (ENN) [22, 14]. ENN is a cleaning method that refines the dataset by removing noisy or misclassified samples in the vicinity of each instance, thereby enhancing the quality of the data. By combining these techniques, SMOTE-ENN not only increases the number of minority class samples but also ensures that the dataset is cleaner and more balanced.

Model Building & Evaluation

Model Building and Training Steps

In this project, we built several machine learning models, including Random Forest, XGBoost, LightGBM, and K-Nearest Neighbours (KNN), with minimal hyperparameter tuning. Each model was selected for its unique strengths and suitability for classification tasks along with cross validation with Lazy Predict trend on this large dataset. Table 1 illustrates the performance of the models in terms of accuracy and recall. Here are the four top models namely:

Random Forest [22] is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It operates by aggregating the results of many decision trees to improve accuracy and control overfitting.

- **Advantages:** It is robust to overfitting, handles large datasets well, and provides feature importance scores.

XGBoost (Extreme Gradient Boosting) [22] is an optimized gradient boosting algorithm designed to be highly efficient, flexible, and portable. It builds models in a sequential manner, where each new model corrects the errors of the previous one.

- **Advantages:** XGBoost often achieves better performance and accuracy due to its boosting approach, regularization techniques to prevent overfitting, and efficient handling of large datasets.

LightGBM (Light Gradient Boosting Machine)[22] is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with lower memory usage and faster training speed compared to traditional gradient boosting methods.

- **Advantages:** LightGBM is known for its speed, efficiency, and ability to handle large-scale data and high-dimensional features effectively.

K-Nearest Neighbours (KNN) [22] is a simple, instance-based learning algorithm that classifies new instances based on the majority class among its K-nearest neighbours in the feature space.

- **Advantages:** KNN is easy to understand, requires no training phase (as it's a lazy learner), and can handle multi-class classification tasks effectively.

Training Steps: We trained all the models using their default parameters on the SMOTE-ENN balanced training data, without engaging in hyperparameter tuning due to resource limitations during training. Despite this, all the models performed exceptionally well on the test set, demonstrating robust predictive capabilities even without the need for fine-tuning. This approach allowed us to efficiently evaluate the models' performance while maintaining a focus on the primary objectives of accuracy, precision, recall and F1 score in predicting diabetes. After assessing these metrics, interpretable methods were employed to address the research question effectively, ensuring that the results were both actionable and comprehensible.

Model Evaluation

To assess the performance of our machine learning models, we utilized several key metrics: accuracy, precision, recall, F1 score and confusion matrix [18, 22].

Accuracy: Accuracy is the ratio of correctly predicted instances to the total number of instances in the dataset. It provides an overall measure of how well the model performs. The formula for accuracy is:

Accuracy = Total Number of Predictions / Number of Correct Predictions

Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is defined as:

Precision = True Positives / (True Positive + False Positive)

Precision indicates how many of the predicted positive cases are actually diabetic for our data.

Recall: Recall, also known as Sensitivity or True Positive Rate, measures the proportion of actual positive instances that were correctly identified by the model. It is calculated as:

Recall = True Positives / (True Positives + False Negatives)

Recall is crucial for this dataset because it reflects the model's ability to identify all potential cases of diabetes. High recall is important in medical diagnosis to ensure that as many cases as possible are detected, even if it means sacrificing some precision[23]. This helps in minimizing the risk of missing out on diabetic patients who might need early intervention.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is defined as:

F1-Score = 2 * ((Precision * Recall) / (Precision + Recall))

The F1-score is particularly useful when dealing with imbalanced datasets, as it takes both precision and recall into account.

Confusion Matrix: The confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This matrix helps in understanding the distribution of predictions and errors, providing insights into where the model might be making mistakes.

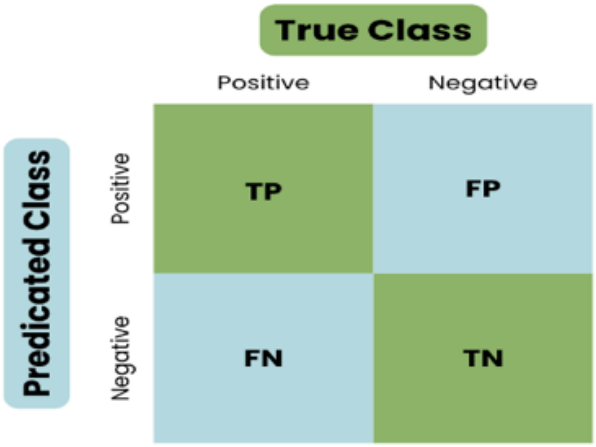

Figure 12: Confusion Matrix[18]

Table 1: Performance Evaluation

Classifier	Accuracy	Recall (Diabetic)	Recall (Non-Diabetic)
Random Forest	97.29%	0.95	0.99
XGBoost	94.6%	0.91	0.97
LightGBM	93.68%	0.91	0.96
KNN	95.35%	0.94	0.96

The table1 presents the performance metrics of four classifiers for diagnosing diabetes. The Random Forest model achieved the highest accuracy at 97.29%, with a recall of 0.95 for diabetic cases and 0.99 for non-diabetic cases, indicating a strong ability to detect both

classes effectively. The XGBoost model followed with an accuracy of 94.6%, showing a recall of 0.91 for diabetic cases and 0.97 for non-diabetic cases, reflecting good performance but slightly lower recall for diabetes detection. LightGBM had an accuracy of 93.68%, with recall values of 0.91 for diabetic and 0.96 for non-diabetic cases, similar to XGBoost but with slightly lower overall accuracy. The KNN classifier, while having an accuracy of 95.35%, demonstrated a recall of 0.94 for diabetic cases and 0.96 for non-diabetic cases, striking a balance between detecting both classes effectively. In this context, recall is a critical metric because it measures the model's ability to correctly identify diabetic cases [23]. High recall for the diabetic class ensures that most true cases of diabetes are detected, which is essential for early diagnosis and effective treatment. Prioritizing recall helps in minimizing the risk of missing out on individuals who are diabetic, thus improving patient outcomes and facilitating timely medical intervention.

Interpretability

In the context of the Diabetes Health Indicators Dataset, interpretability is crucial for effectively leveraging machine learning models in diabetes management and prediction. This dataset plays a vital role in identifying the factors associated with diabetes, making it essential for healthcare professionals to understand how models arrive at their predictions. Interpretability builds trust in the model's results and supports informed clinical decision-making, ensuring that insights are actionable and relevant. This facilitates personalized treatment plans and improves patient outcomes while adhering to regulatory standards by providing transparent explanations of the prediction processes. Permutation Feature Importance (PFI) and SHAP (Shapley Additive Explanations) are powerful interpretability tools used in machine learning[19]. These are model agnostic and works with both black box and white box models.

Permutation Feature Importance (PFI) [19] measures the impact of each feature on the model's performance by randomly shuffling the values of a feature and observing the change in the model's accuracy. If shuffling a feature's values significantly reduces the model's accuracy, that feature is considered important. PFI provides a global perspective on feature importance, helping to identify which variables most influence the model's predictions.

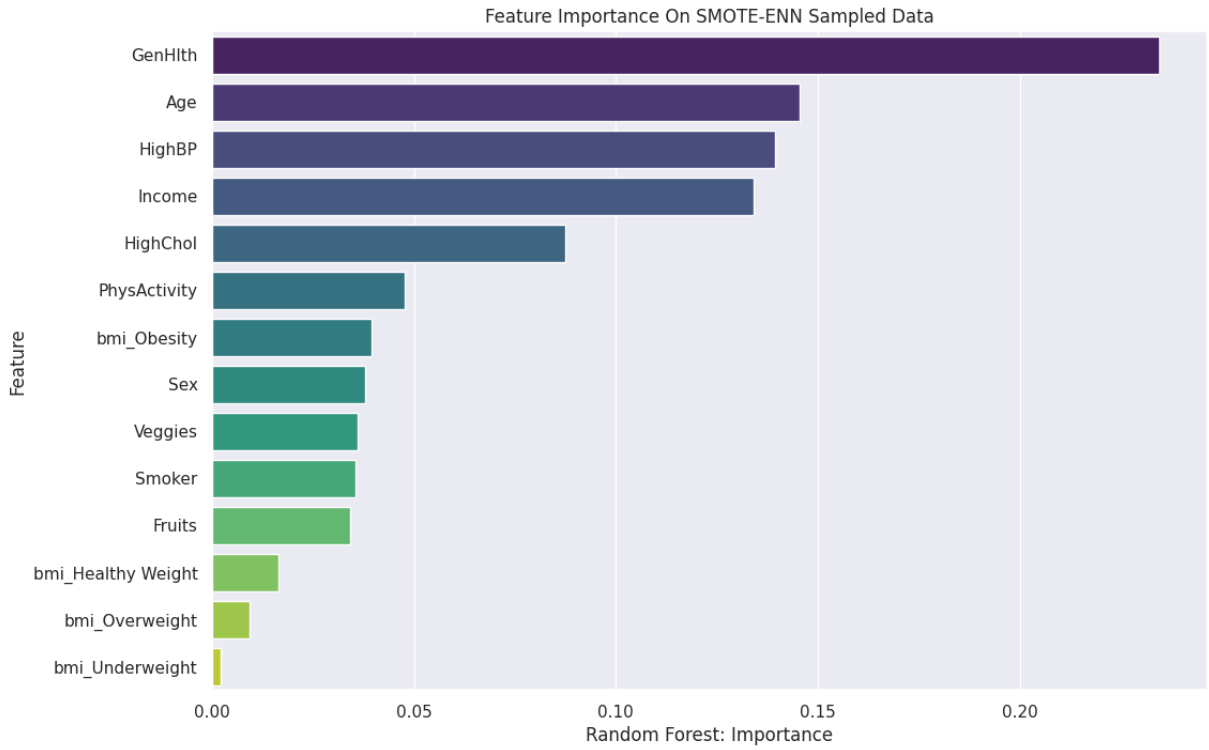

Figure 13:PFI bar plot for Random Forest

SHAP (Shapley Additive Explanations) [19] , on the other hand, is a game-theory-based approach that explains the contribution of each feature to the final prediction. SHAP values provide insights into how much each feature contributes to increasing or decreasing the prediction. Bar plots in SHAP show the average contribution of each feature across all predictions, while Beeswarm plots visualize the distribution of SHAP values for all instances, offering a more detailed view of feature impacts.

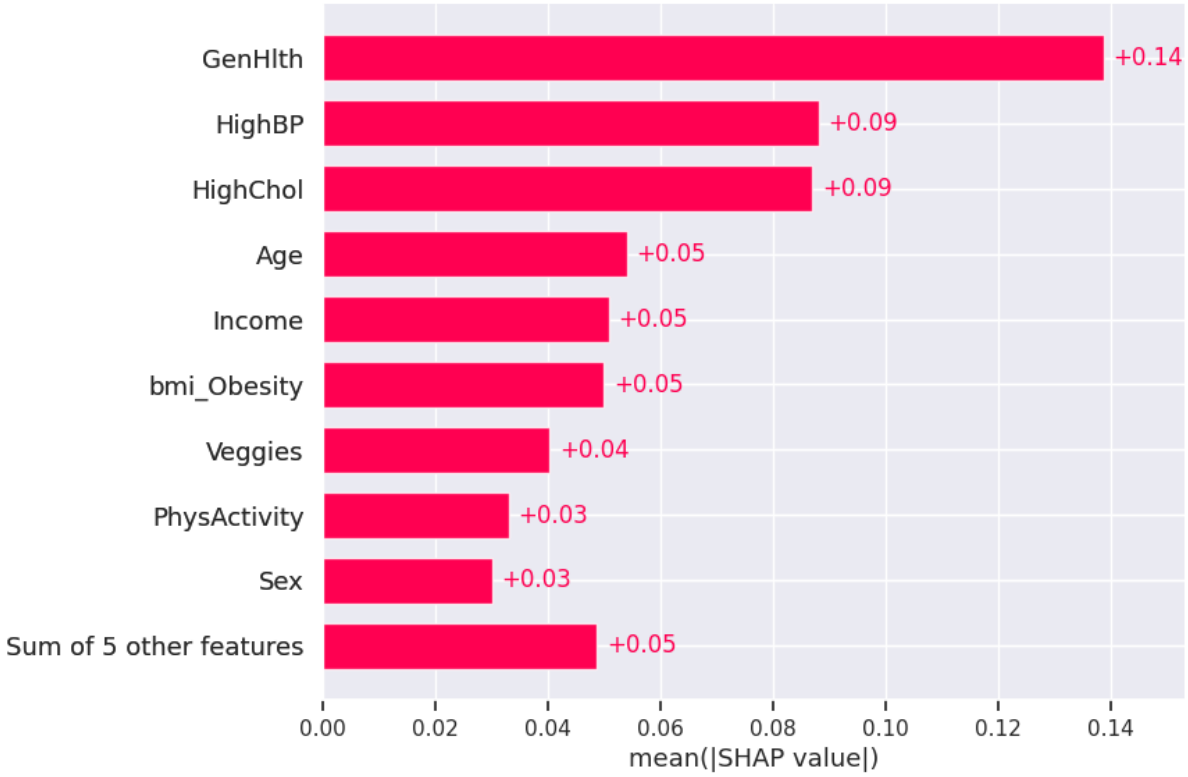

Figure 14:SHAP Values for Random Forest

In Figure 13, the bar plot illustrates the impact of each feature on the random forest model's prediction, as determined by the Permutation Feature Importance (PFI) method. The analysis highlights General Health, Age, High Blood Pressure, Income, High Cholesterol, Physical Activity, and Obesity as the most significant factors in diagnosing diabetes. These findings align with established medical understanding of the disease. Notably, a similar trend is observed in the global SHAP analysis (figure 14) for the all models, as well as in the PFI results of other models, with minor variations in feature ranking.

Local SHAP [19] provides explanations by attributing a portion of the model's prediction to each feature, offering a clear insight into how specific features contribute to a particular prediction. It is based on Shapley values from cooperative game theory, ensuring that the contributions are fairly distributed among features.

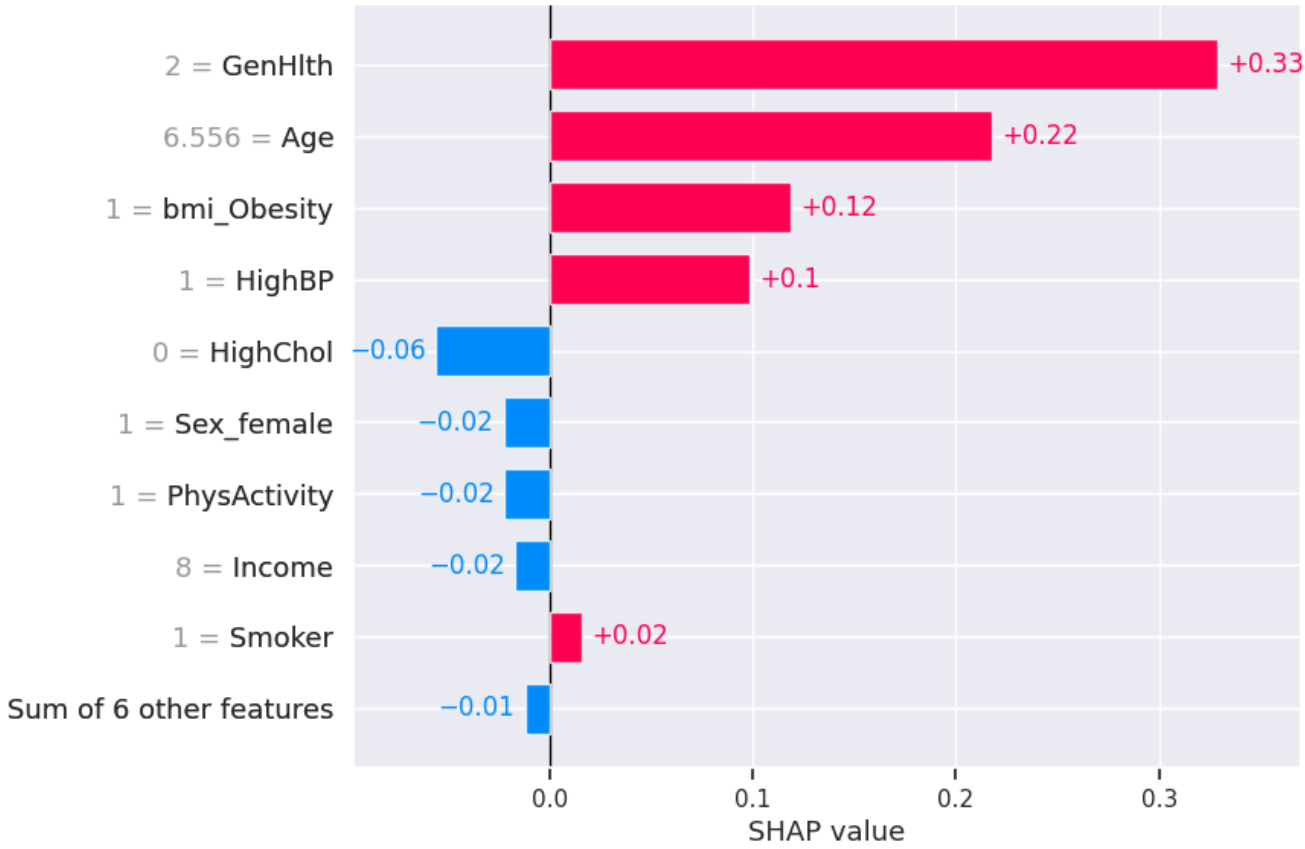

Figure 15:Local Shap for Random Forest

Figure 15 illustrates the decision-making process for a specific diabetic instance using local SHAP. In this case, the individual exhibits several high-risk factors: lower general health, advanced age (between 60 and 65 years), obesity, high blood pressure, and smoking. These factors collectively increase the likelihood of diabetes, contributing positively to the prediction of diabetes. Conversely, the absence of high cholesterol, physical activity, and a higher income bracket act as mitigating factors, reducing the likelihood of diabetes. Despite these negative factors, the positive risk factors outweigh the mitigating ones, leading to the final classification of the individual as diabetic.

LIME (Local interpretable model-agnostic explanations) [19], on the other hand, approximates the model locally around the prediction of interest using a simpler, interpretable model, like a linear regression. It helps in understanding why the model made a specific prediction by highlighting the most influential features.

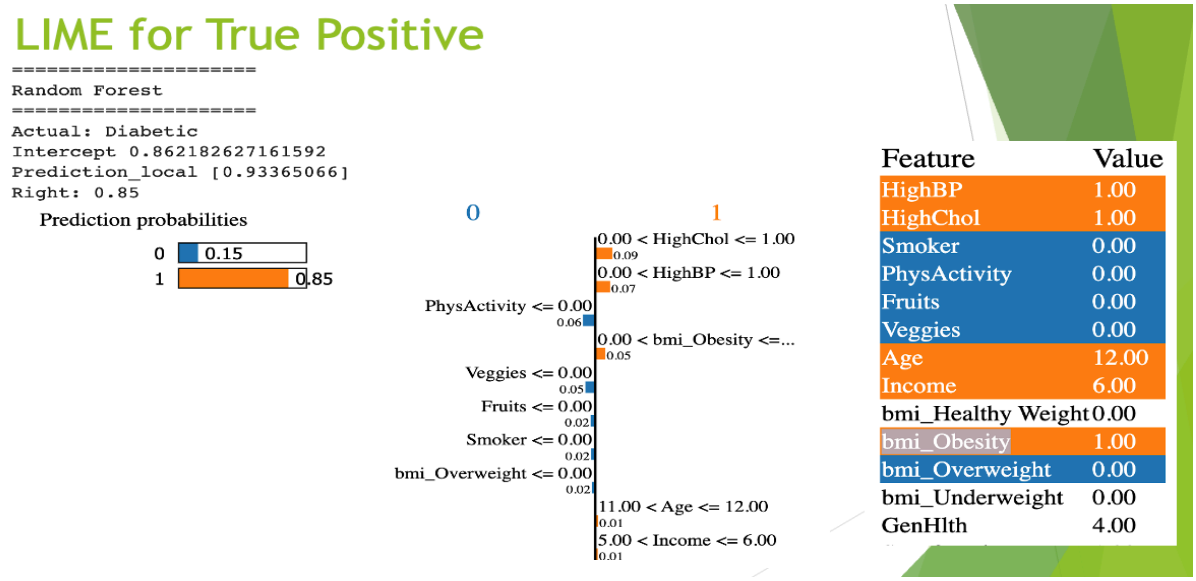

Figure 16:LIME for true positive case in random forest

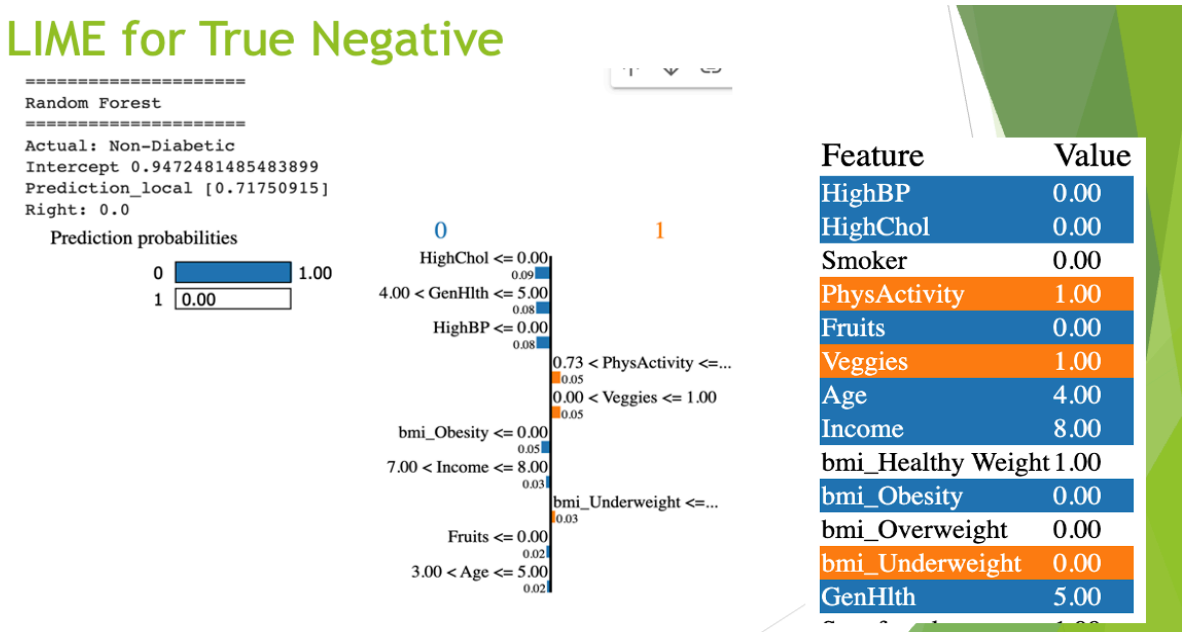

Figure 17:True Negative prediction by LIME

Figure 16 illustrates the local interpretability of a diabetic instance using the LIME method with the random forest model. The model shows 85% confidence in classifying this instance as diabetic. The key factors influencing this diagnosis include high blood pressure, high cholesterol, advanced age (between 60 and 65 years), and obesity, all of which significantly contribute to the positive diagnosis of diabetes. Conversely, Figure 17 presents the local interpretability of a non-diabetic instance using the same LIME method. Here, the model demonstrates 100% confidence in classifying the individual as non-diabetic. The absence of high blood pressure, high cholesterol, and obesity, along with overall excellent general health, is responsible for the model's prediction of non-diabetes.

For ease of computation, all interpretability analyses were conducted on a subset of the test data. We employed Permutation Feature Importance (PFI) and SHAP (Shapley Additive Explanations) with bar and Beeswarm plots for global interpretability, as well as SHAP and LIME for local interpretability. Local interpretability is particularly valuable in identifying factors that contribute to an individual patient's diabetes risk, enabling personalized treatment and more informed clinical decisions—especially if the model is integrated into a mobile or web application for healthcare professionals. These methods help us understand both the overall and individual contributions of features to model predictions. Our analysis identified obesity, general health, age, high cholesterol, high blood pressure, and income as key factors influencing diabetes, providing concrete answer to our original research question.

Result

The Random Forest model emerged as the top performer across all evaluation metrics, as shown in Figures 18, 19, and 20. The model achieved an impressive recall of 0.99 for class 0 (non-diabetic), correctly identifying 99% of non-diabetic instances, and a recall of 0.95 for class 1 (diabetic), accurately detecting 95% of diabetic cases. The F1-scores were equally strong, with 0.98 for class 0, indicating a well-balanced precision and recall, and 0.97 for class 1, underscoring the model's effectiveness in identifying diabetic cases. Overall accuracy reached 97.29% on the test set, reflecting a high level of correct classifications. Precision was also noteworthy, with 0.98 for class 0, meaning 98% of non-diabetic predictions were true positives, and 0.97 for class 1, signifying that 97% of diabetic predictions were accurate. The macro and weighted average precision, recall, and F1-scores were all 0.97, demonstrating the model's balanced performance across both classes.

The confusion matrix further highlighted the model's robustness, revealing only 594 false positives and 294 false negatives. The low number of false positives suggests that non-diabetic individuals were rarely misclassified as diabetic, while the low number of false negatives indicates that the model effectively identified most diabetic cases. This balance in minimizing misclassifications underscores the Random Forest model's reliability in predicting diabetes. The model can further benefit from hyperparameter tuning in future endeavours.

Table 2: Result Table with focus on Recall and Accuracy

Classifier	**Accuracy**	**Recall (Diabetic)**	**Recall (Non-Diabetic)**
Random Forest	97.29%	0.95	0.99
XGBoost	94.6%	0.91	0.97
LightGBM	93.68%	0.91	0.96
KNN	95.35%	0.94	0.96

Moreover, models like LightGBM, XGBoost, and KNN also performed admirably, though their recall, accuracy and recall were slightly lower than those of the Random Forest model (Table 2). These strong results can be attributed in part to the SMOTE-ENN class balancing technique used for data augmentation, which enhances traditional SMOTE by incorporating Edited Nearest Neighbours (ENN) to reduce noise and refine synthetic samples. Additionally, the feature selection strategy, including Chi-Square tests, correlation analysis, and guardrail alerts, further boosted model performance.

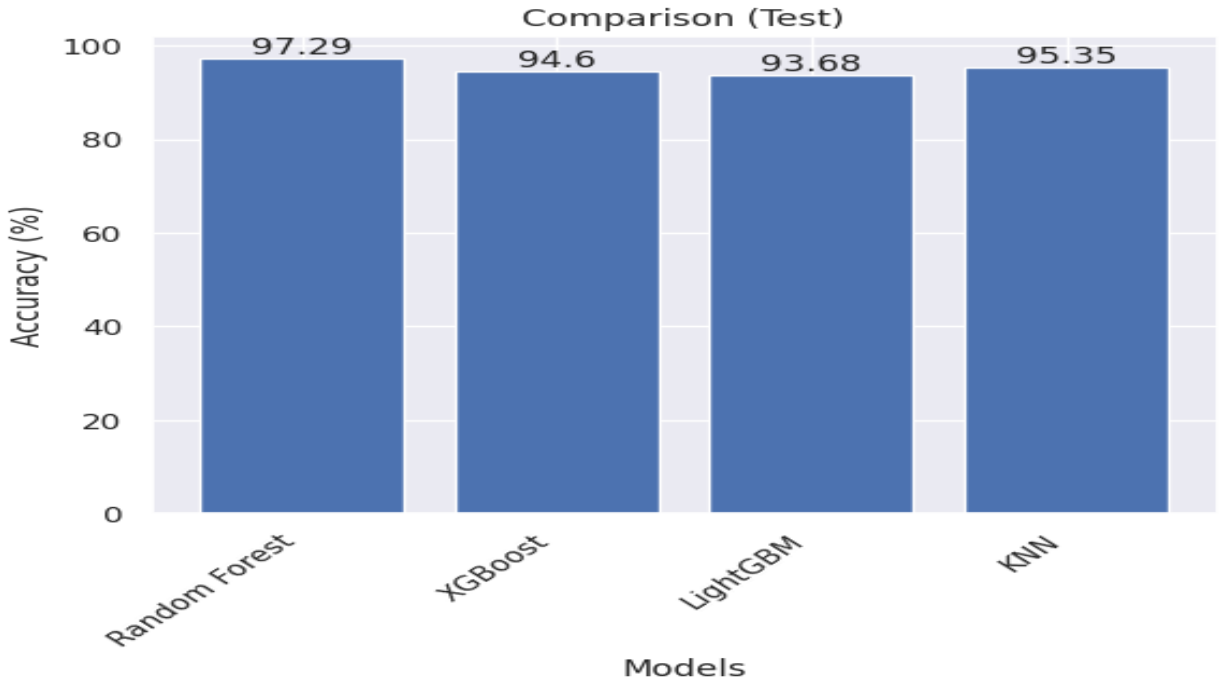

Figure 18:Accuracy Comparison

Predicted Label

Test Set:

	precision	recall	f1-score	support
0.0	0.97	0.99	0.98	19868
1.0	0.98	0.95	0.97	12847
accuracy			0.97	32715
macro avg	0.97	0.97	0.97	32715
weighted avg	0.97	0.97	0.97	32715

Test Set Accuracy: 97.28564878496103

Figure 19:Classification Matrix of Random Forest

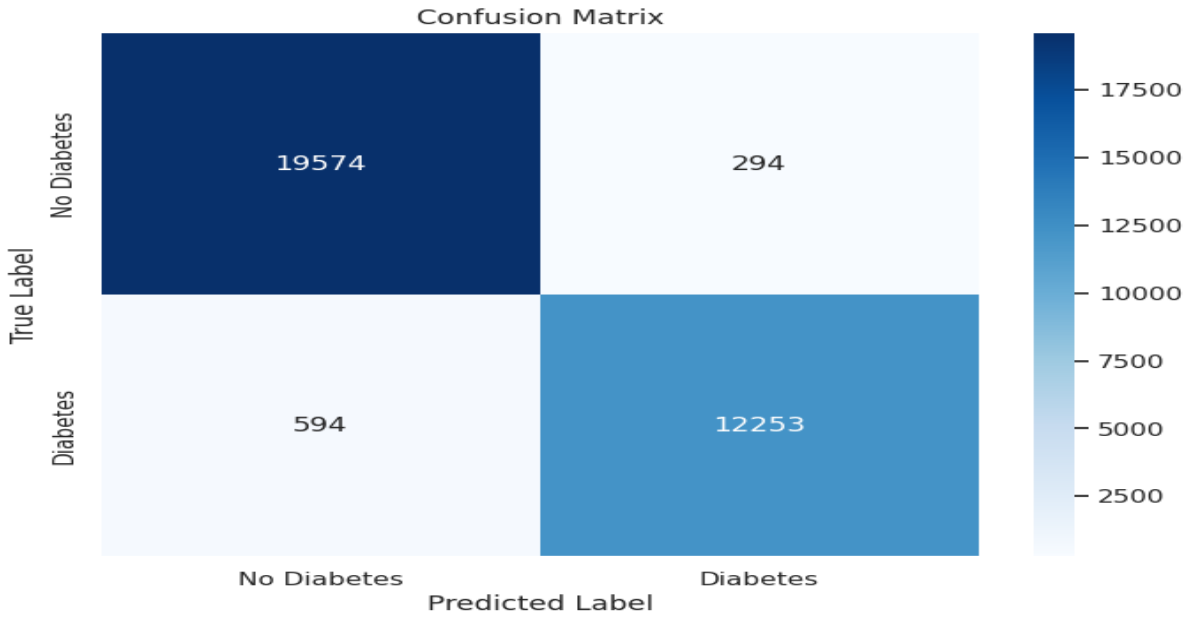

Figure 20:Confusion Matrix of Random Forest

Model interpretability is a central focus of this project, particularly in addressing the research question: “Which specific health, lifestyle, and socio-economic factors play a significant role in the diagnosis of diabetes in individuals?” Our analysis, supported by SHAP and Permutation Feature Importance (PFI), consistently identifies General Health, Age, High Blood Pressure, Income, High Cholesterol, Physical Activity, and Obesity as the most influential factors in diagnosing diabetes, as shown in Figures 14 and 21. Interestingly, these factors have remained consistent across models and interpretability techniques, with only minor variations in feature rankings. These findings align closely with the insights gained during the Exploratory Data Analysis (EDA) phase, reinforcing their significance and efficacy.

Furthermore, the Beeswarm plot (Figure 20) offers a comprehensive overview of how the top features influence the model's predictions, further deepening our understanding of its behaviour. Each dot on the plot represents an individual instance's contribution to a specific feature. The x-axis position of each dot corresponds to the SHAP value, while the density of dots indicates the distribution of these values. The color of each dot reflects the original value of the feature, enhancing the interpretability of the model’s predictions. For instance, as shown in Figure 22, a high concentration of higher General Health values correlates with a lower likelihood of being diabetic, while high cholesterol and obesity are associated with an increased likelihood of diabetes.

Additionally, the local SHAP and LIME interpretation techniques discussed in the previous section are particularly valuable for identifying factors that contribute to an individual patient's diabetes risk. This capability enables personalized treatment and more informed clinical decisions, especially if the model is integrated into a mobile or web application for healthcare professionals. Overall, the combination of local and global interpretability methods strengthens our confidence in understanding the key factors influencing diabetes.

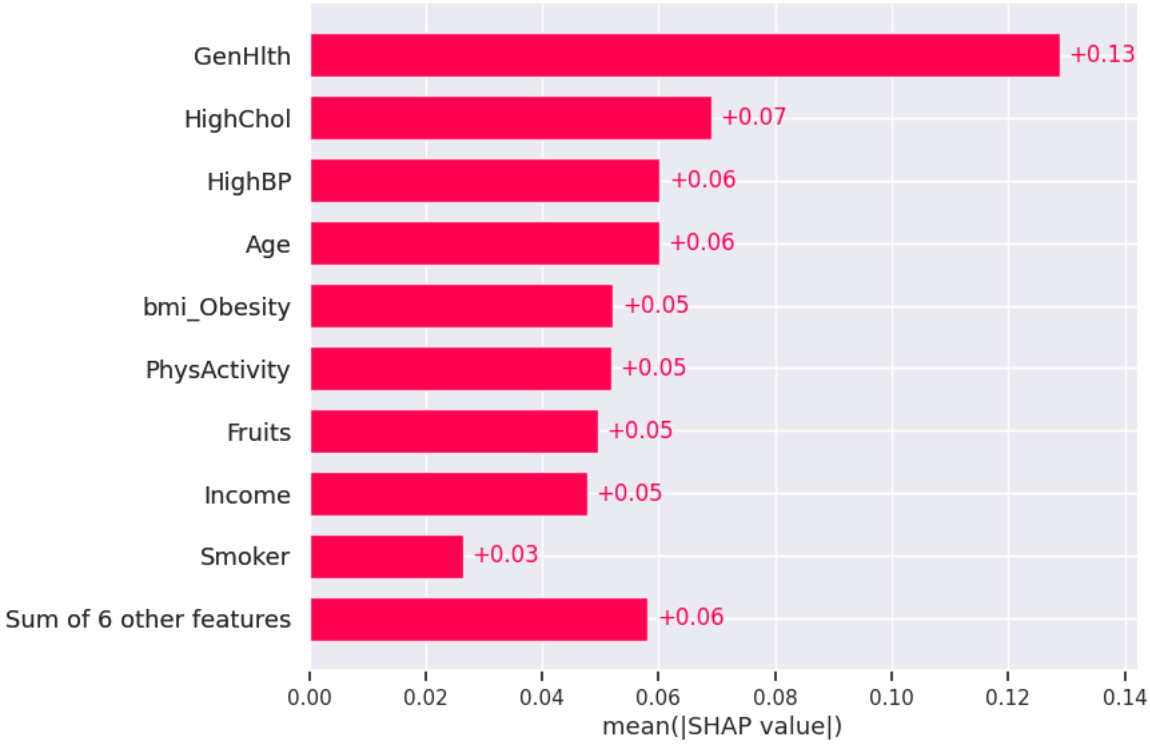

Figure 21:Bar plot for Global Shap (LGBM model)

Figure 22: Beeswarm Plot (Random Forest)

Conclusion & Future Work

Summary

Diabetes is a chronic condition that affects millions of people worldwide, leading to severe health complications if not managed effectively[2]. The growing prevalence of diabetes, coupled with its associated healthcare costs and impact on quality of life, underscores the urgent need for improved diagnostic and preventative tools. This project was motivated by the desire to enhance the accuracy and interpretability of diabetes diagnosis using machine learning techniques. The research focused on identifying the key health, lifestyle, and socio-economic factors that significantly contribute to diabetes risk, with the ultimate goal of informing more effective prevention and treatment strategies.
Using a comprehensive dataset Diabetes Health Indicators containing diverse features, we applied several machine learning models, including Random Forest, LightGBM, XGBoost, and KNN, to classify individuals as diabetic or non-diabetic. These models are built on training data created using a combination of correlation analysis and the Chi-square test for feature selection, along with the SMOTE-ENN technique to address class imbalance. By carefully selecting features and addressing class imbalances, the models were better equipped to learn from the data, leading to more accurate predictions and meaningful insights. The project emphasized not only high predictive accuracy and recall but also the interpretability of the models through methods such as SHAP, Permutation Feature Importance (PFI) etc, allowing for a deeper understanding of the factors driving diabetes onset.

Result Summary

The Random Forest model emerged as the top performer, achieving an impressive accuracy of 97.29% on the test set. The model achieved an impressive recall of 0.99 for class 0 (non-diabetic), correctly identifying 99% of non-diabetic instances, and a recall of 0.95 for class 1 (diabetic), accurately detecting 95% of diabetic cases. It also demonstrated strong precision, and F1-scores for both diabetic and non-diabetic classes, with minimal misclassification rates. Key factors such as General Health, Age, High Blood Pressure, Income, High Cholesterol, Physical Activity, and Obesity were consistently identified as the most influential in diagnosing diabetes. These findings were supported by SHAP, Permutation Feature Importance (PFI), LIME methods, which provided both global and local interpretability insights. The alignment of these results with the initial Exploratory Data Analysis (EDA) further validated the robustness of the model.

Limitations

While the models showed strong performance, there are several limitations to acknowledge. The dataset, though extensive, may not capture all relevant variables, particularly those related to genetic or environmental influences on diabetes. The project involved minimal hyperparameter tuning to expedite the process, which may have constrained the models' performance. The use of SMOTE-ENN for class balancing, while effective and better than SMOTE, might have introduced biases by oversampling minority classes, potentially affecting the generalizability of the models. Additionally, the models were tested on a

specific dataset, and their performance might differ when applied to diverse populations or real-world settings.

Future Work

Future work could focus on more extensive hyperparameter tuning and exploring advanced techniques like ensemble models or deep learning models to further enhance performance. Expanding the dataset to include more diverse populations and additional features, such as genetic markers or detailed dietary habits, could improve the models' accuracy and applicability. Integrating these models into real-time clinical decision support systems could enable healthcare professionals to make more informed and personalized treatment decisions. Additionally, investigating the potential of transfer learning could allow the models to adapt better to different populations or regions.

In this project, we merged the diabetic and prediabetic classes due to the lack of sufficient data in the prediabetic category (which constituted only 2% of the dataset). Future research could focus on effectively identifying this prediabetic class, thereby extending our current line of work. Moreover, integrating the model into an easy-to-use application for healthcare workers or patients could be a significant step towards introducing this concept into real-life scenarios, improving accessibility and practical utility.

Conclusion

In conclusion, this project successfully identified and validated key factors influencing diabetes diagnosis through interpretable machine learning models. The Random Forest model, in particular, demonstrated high accuracy and reliability, making it a strong candidate for integration into clinical decision-making tools. Despite certain limitations, the insights gained provide a solid foundation for future research and practical applications in diabetes management. Continued refinement of these models and expansion of the dataset could lead to more accurate, personalized, and accessible tools for diagnosing and preventing diabetes, ultimately improving patient outcomes and quality of life.

References:

[1] V. A. Kumari and R. Chitra, "Classification of Diabetes Disease Using Support Vector Machine," Int. J. Eng. Res. Appl. (IJERA), vol. 3, pp. 1797-1801, 2013.

[2] Cleveland Clinic, "Diabetes," available at: https://my.clevelandclinic.org/health/diseases/7104-diabetes.

[3] U.S. Department of Health and Human Services Centers for Disease Control and Prevention, "National Diabetes Statistics Report Estimates of Diabetes and Its Burden in the United States," 2020. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

[4] National Center for Chronic Disease Prevention and Health Promotion. Division of Population Health, "Behavioral Risk Factor Surveillance System (BRFSS)," available at: https://www.cdc.gov/brfss/index.html.

[5] A. Teboul, "Diabetes Health Indicators Dataset," 2021. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

[6] C. B. Weir and A. Jan, "BMI Classification Percentile And Cut Off Points," in StatPearls [Internet], Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK541070/.

[7] UC Irvine Machine Learning Repository, "CDC Diabetes Health Indicators Dataset," available at: https://www.archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators.

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104-116, 2017.

[9] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, pp. 1578-1585, 2018.

[10] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Diabetes Mellitus Based on Hybrid Features Extraction and an Ensemble of Machine Learning Classifiers," Healthc. Inform. Res., vol. 25, no. 4, pp. 251-261, 2019.

[11] M. I. Razzak, S. Naz, and A. Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges, and the Future. Classification and Diagnosis of Diabetes Mellitus Using Deep Learning Techniques," 2020.

[12] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," Preventing Chronic Disease, vol. 16, 2019.

[13] X. Ren, "Predictions of diabetes through machine learning models based on the health indicators dataset," Appl. Comput. Eng., vol. 32, pp. 216-222, 2024.

[14] M. Lamari, N. Azizi, N. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, and N. E. Benzebouchi, "SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification," 2021.

[15] A. Anishnama, "Understanding the Chi-Square Test: An Introduction to Its Concept and Applications," Medium. [Online]. Available: https://medium.com/@anishnama20/understanding-the-chi-square-test-an-introduction-to-its-concept-and-applications-9c9009ddb38.

[16] L. Severson, "Life with diabetes: What happens as we age?," Mayo Clinic Health System. [Online]. Available: https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/life-with-diabetes-what-happens-as-we-age.

[17] Y. K. Yashi and S. F. Daley, "Obesity and Type 2 Diabetes," in StatPearls [Internet], Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK592412/.

[18] "What is a Confusion Matrix in Machine Learning?," DataCamp. [Online]. Available: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning.

[19] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2020.

[20] "Beeswarm plot example," SHAP. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html.

[21] "Diabetes and Your Heart," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-your-heart.html.

[22] A. Müller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly Media, 2016.

[23] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," Sci Rep., vol. 12, no. 1, p. 5979, Apr. 2022, doi: 10.1038/s41598-022-09954-8.

Feature Codebook

Here is the definition of all the features in the Diabetes Health Indicators dataset

Table 3: Feature Description

Variable Name	Description
Diabetes_binary	0 = no diabetes 1 = prediabetes or diabetes
HighBP	0 = no high BP 1 = high BP
HighChol	0 = no high cholesterol 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI	Body Mass Index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
Stroke	(Ever told) you had a stroke. 0 = no 1 = yes
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity	physical activity in past 30 days - not including job 0 = no 1 = yes
Fruits	Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies	Consume Vegetables 1 or more times per day 0 = no 1 = yes

HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor. This includes general wellbeing, stress level and sleep hygiene.
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Sex	0 = female 1 = male
Age	13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more