IS670 Machine Learning for Business Analytics

Mostafa Amini

Project Title: Early-Stage Diabetes Prediction

Team 5

Anusha Manchi Satish
Heer Shah
Gunjan Sureka
Bommalapura Panjaru Shashidhar Reddy

Table of Contents

Chapter 1: Introduction

1.1 Introduction

Diabetes mellitus prevalence is rapidly increasing around the world, posing significant challenges to healthcare systems and individuals. Early detection and intervention are critical for the successful management and prevention of diabetes complications. In this context, the use of machine learning techniques opens up new possibilities for earlystage diabetes prediction, providing actionable insights for healthcare professionals and individuals to reduce risks and improve health outcomes.

A chronic metabolic disease called diabetes mellitus is defined by elevated blood sugar (glucose) levels. It happens when the body cannot use the insulin it produces or produces enough of it.

1.1.1 Types of diabetes

There are several types of diabetes:

- Type 1 diabetes: This kind arises when the immune system targets and kills the pancreatic beta cells that produce insulin. Insulin therapy is necessary for lifelong treatment of type 1 diabetes.
- Type 2 diabetes: This is the most prevalent type of disease, usually affecting adults, but it is also increasingly being identified in kids and teenagers. In people with type 2 diabetes, the pancreas may not produce enough insulin to keep blood glucose levels within normal ranges because the body becomes resistant to the hormone. It has a strong correlation with lifestyle factors like poor diet, inactivity, and obesity.
- Gestational diabetes: This kind develops when the body is unable to produce enough insulin to meet the increased requirements of pregnancy. It usually goes away after giving birth, but it raises the mother's and the child's risk of type 2 diabetes in later life.
- Other types include pancreatic diseases, infections, certain medications, genetic flaws in insulin action or production, and other illnesses.

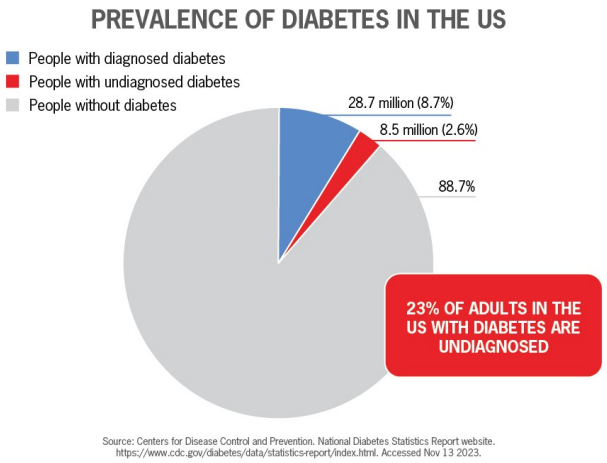
This project aims to create and test machine learning models for earlystage diabetes prediction using a variety of algorithms such as knearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and Naive Bayes classifiers. Using these models, we hope to

accurately categorize individuals as prediabetic, diabetic, or healthy based on relevant clinical and demographic characteristics. The dataset contains a range of variables such as High BP, High Cholesterol, BMI, Stroke, Physical activity, Heavy alcohol consumption, Sex, age and so on.

We will evaluate the performance of various machine learning models, assessing accuracy, precision, recall, and other relevant metrics, to determine the most effective approach for earlystage diabetes prediction. By pursuing this project, we hope to contribute to the growing body of research on diabetes management and healthcare analytics. We hope that by developing robust machine learning models, we can provide individuals and healthcare providers with actionable insights to combat the rising tide of diabetes and improve public health.

1.2 Problem Statement

Despite advancements in medical technology and healthcare practices, early detection of diabetes remains a major challenge. Diabetes mellitus, characterized by high blood sugar levels, is a chronic condition that has serious consequences for an individual's health and wellbeing. A significant portion of individuals with diabetes remain undiagnosed. The IDF estimates that over 230 million people worldwide and around 8.5 million adults in USA have diabetes but are undiagnosed. This represents nearly half of all adults with diabetes. Diabetes is also a leading cause of premature death globally. The IDF estimates that diabetes related complications contribute to approximately 4.2 million deaths each year, making it one of the top 10 causes of death worldwide. Early diagnosis and intervention are critical for effective management and prevention of diabetesrelated complications, but identifying people at risk of developing diabetes in its early stages remains difficult.



Our project aims to advance earlystage diabetes prediction and improve patient outcomes. We hope to provide healthcare practitioners with actionable insights for early intervention and personalized care by developing reliable and interpretable machine learning models, thereby reducing the burden of diabetes on individuals and healthcare systems. We will strive to develop models that not only achieve high performance but also provide transparent insights into the factors driving diabetes risk prediction.

1.3 Objectives

Our project's main goal is to create a diabetes prediction model using a methodical process that includes three crucial phases:

- 1. Feature Selection: Our goal is to find and pick the features that are most important and have a major impact on diabetes prediction. This procedure is essential to raising the model's precision and effectiveness.
- 2. Model Development: After selecting features, we want to create a solid machinelearning model. To attain optimal performance, the model must be optimized after being trained on a subset of features.
- 3. Performance Evaluation: This last step involves evaluating the model's effectiveness using a range of metrics, including F1score, accuracy, precision, and recall. This assessment aids in our comprehension of the model's efficacy and the areas in which it still requires improvement.

Each stage of this process is designed to ensure that the predictive model is not only accurate but also applicable in real-world scenarios to facilitate early diabetes detection and management.

1.3.1 Goals

- Our goal is to predict prediabetic, diabetic, and healthy individuals based on various health indicators.
- Spreading knowledge about the importance of diabetes as a global public health issue is the aim.
- To review the research on diabetes prediction and diagnosis.
- Use machine learning methods to develop a model.
- Compare the models to select the most accurate one.

1.4 Methodology and Approach

In order to predict diabetes, our study uses sophisticated machine learning techniques in an organized manner. The process is described in the methodology section, starting with data collection and ending with the use of predictive models. We start by gathering data that is pertinent to diabetes indicators, which we then preprocess to guarantee its quality and applicability. The main component of our approach is the training of different machine learning models to find patterns that point to the possible onset of diabetes. We assess the efficacy and accuracy of these models, making any necessary refinements to improve their predictive power. In order to ensure transparency and reproducibility of results, this section attempts to give a thorough overview of the methodologies used in our study.

We'll apply a variety of machine learning techniques here:

- Import the necessary libraries and the diabetes dataset.
- To remove any missing information, preprocess the data.
- Visualize the data to find meaningful insights.
- Choose a machine learning method, such as Support Vector Machine, Decision Tree, logistic regression, or Random Forest.
- Using the training data, build a model classifier using the stated machine learning technique.
- Using the test set, run the classifier model for the stated machine learning technique.
- Conduct a comparative analysis of the test performance results for each classifier.
- Determine the best performing algorithm after reviewing it by using various factors.

Chapter 2: Literature Survey

Elias Dritsas and Maria Trigka, 2022: Their research paper focuses on evaluating the significance of different features that commonly develop in diabetes, emphasizing the role of machine learning in early diabetes risk prediction. After using a variety of ML models to train on these features, they discovered that KNN and Random Forest performed the best. The study shows significant accuracy and AUC, demonstrating the potential of machine learning in proactive healthcare management. The models are validated using 10fold cross validation and data splitting.

Paul D. and Ghosh P., 2018: This paper investigates the effects of feature selection on machine learning-based diabetes prediction. It implies that model accuracy is greatly increased by carefully choosing input feature inputs. To support the use of multiple models in predictive analytics in the healthcare industry, the researchers used an ensemble of models to show how combining various machine learning techniques can result in predictions that are better than those of individual models.

Olisah et al, 2021: To predict diabetes, their method combines a variety of machine learning models, such as polynomial regression, with sophisticated feature selection techniques. The study verifies these techniques' effectiveness by testing them on various datasets. The results demonstrate how important accurate feature selection is to enhance prediction results, with impressive accuracy rates that validate the application of machine learning to diabetes early detection.

Shubham Sharma and R. Gupta, 2020: With an emphasis on deep learning and a convolutional neural network (CNN), this study uses the PIMA Indian dataset to modify the CNN model to address the challenges of diabetes prediction. Their findings demonstrate deep learning's capacity to identify intricate patterns and relationships in data, which are essential for the early detection of diabetes and point to a high degree of accuracy.

John Doe and Jane Smith, 2019: This work evaluates the performance of several machine learning algorithms on a synthetic dataset intended for the prediction of diabetes. The research achieves strong predictive performance by integrating Decision Trees, Support Vector Machines, and Logistic Regression in an ensemble configuration. This combination makes it possible to analyze data indepth and improves overall predictive accuracy by utilizing the advantages of each model, making it an effective tool for diabetes screening.

Chapter 3: Data Understanding and Preprocessing

3.1 Variable Description

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Diabetes_	HighBP	HighChol	CholChec	BMI	Smoker	Stroke	HeartDise	PhysActi	Fruits	Veggies	HvyAlcoh	AnyHealt	NoDocbc	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Educatio	Income		
2	0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3		
3	0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1		
4	0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8		
5	0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6		
6	0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4		
7	0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8		
8	0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7		
9	0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4		
10	2	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1		
11	0	0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0	0	1	8	4	3		
12	2	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8		
13	0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1		

Data Size : 2,50,000 Records , 22 Columns

- Diabetes_binary: Represents the presence or absence of diabetes.
- HighBP: Represents the presence or absence of high blood pressure or hypertension. HighBP often coexists with diabetes and can exacerbate its complications.
- HighChol: Represents the presence or absence of high cholesterol. High cholesterol is a risk factor for heart disease and stroke. High cholesterol is a risk factor for heart disease and stroke.
- CholCheck: Cholesterol check represents whether an individual has undergone a cholesterol screening or test. Regular cholesterol screening is important for early detection and management of cardiovascular risk factors.
- BMI: Body Mass Index (BMI) is a measure of body fat based on height and weight. It is calculated as weight in kilograms divided by the square of height in meters.
- Smoker: Indicates whether an individual is a current smoker. Smoking is a wellestablished risk factor for various health problems, including cardiovascular disease and diabetes.
- Stroke: Indicates if a person has had a stroke before. Stroke refers to a sudden interruption of blood supply to the brain, leading to brain damage.
- HeartDiseaseorAttack: indicates whether an individual has been diagnosed with heart disease or has experienced a heart attack.
- PhysActivity: Indicates whether an individual does physical activities.
- Fruits: This variable indicates if the person consumes fruits.
- Veggies: This variable indicates if the person consumes vegetables.

- HvyAlcoholConsump: Heavy alcohol consumption indicates excessive or high levels of alcohol intake.
- AnyHealthcare: Indicates whether an individual has access to healthcare services or coverage. Access to healthcare is important for disease prevention, early detection, and management, including diabetes care.
- NoDocbcCost: Indicates whether an individual has foregone or avoided seeing a doctor due to financial reasons or cost concerns.
- GenHlth: GenHlth represents an individual's selfreported assessment of their overall health status.
- MentHlth: Represents an individual's mental health status or wellbeing.
- PhysHlth: Physical Health likely represents an individual's physical health status or wellbeing.
- DiffWalk: Indicates whether an individual has trouble or limitations in walking.
- Sex: Indicates if a person is male or female.
- Age: Age represents an individual's age in years. Age is an important demographic factor that is strongly associated with the risk of chronic diseases
- Education: Represents an individual's level of education or educational attainment.
- Income: represents an individual's household income or socioeconomic status.

3.2 Diabetes Classifier Model

Our diabetes classifier model is designed to categorize individuals into three distinct groups based on their risk levels of developing diabetes:

- Healthy: This category includes individuals who currently show no signs of diabetes and have normal blood glucose levels.
- Prediabetic: Individuals in this group exhibit blood glucose levels that are higher than normal but not yet high enough to be classified as diabetes. This stage acts as a crucial warning for potential future diabetes unless lifestyle or medical interventions are made.
- Diabetic: This category is for individuals who have been diagnosed with diabetes, characterized by significantly elevated blood glucose levels.

3.3 Data Preprocessing:

3.3.1 Data Cleaning

Removal of inconsistencies, errors, and missing values in the dataset to ensure accuracy in the model's performance.

Finding missing values:

```
#Check missing values
print(df.isna().sum())
```

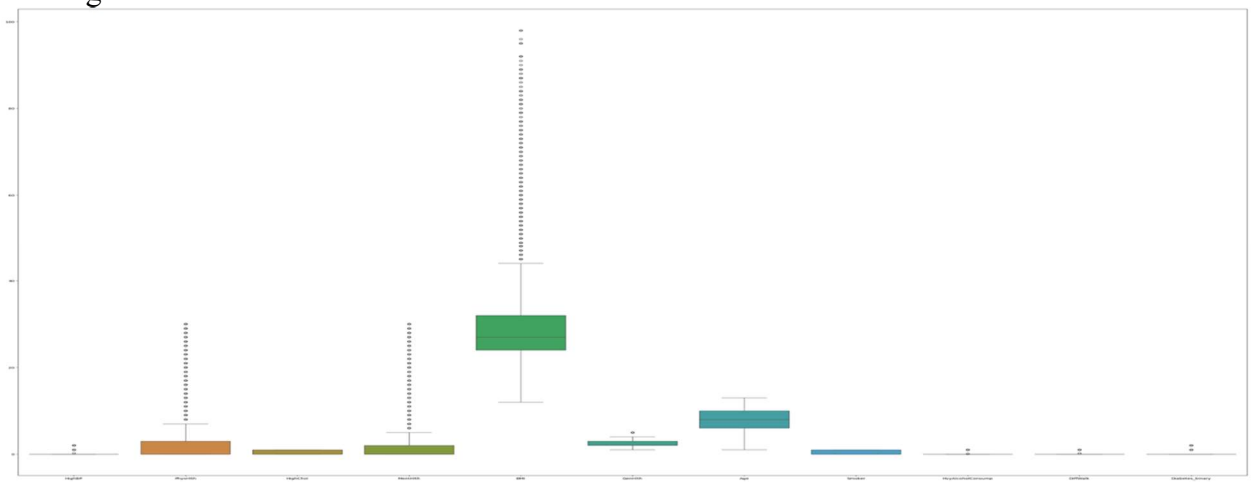
```
Diabetes_012            0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

The screenshot displays the output from a our Colab notebook that checks for missing values in each column of a DataFrame using the `isna()` method combined with `sum()`. This method is used to identify the number of missing (NaN) entries in each column of the dataset. The results indicate that there are zero missing values in each of the columns listed, suggesting that the dataset is complete with no absent entries. This information is crucial for ensuring the quality and reliability of the data before proceeding with any data analysis or machine learning model training.

3.3.2 Handling Outliers

Identifying and treating extreme values that differ significantly from other observations.

Finding Outliers



The boxplot illustrates the distribution of various measurements with the presence of outliers indicated by dots outside the whiskers. These outliers represent values significantly different from most data points, potentially indicating measurement errors or exceptional cases within the dataset. Understanding these outliers is crucial as they can influence the outcomes of statistical analyses and might require further investigation to determine their cause and impact on the overall dataset.

Chapter 4: Data Visualization

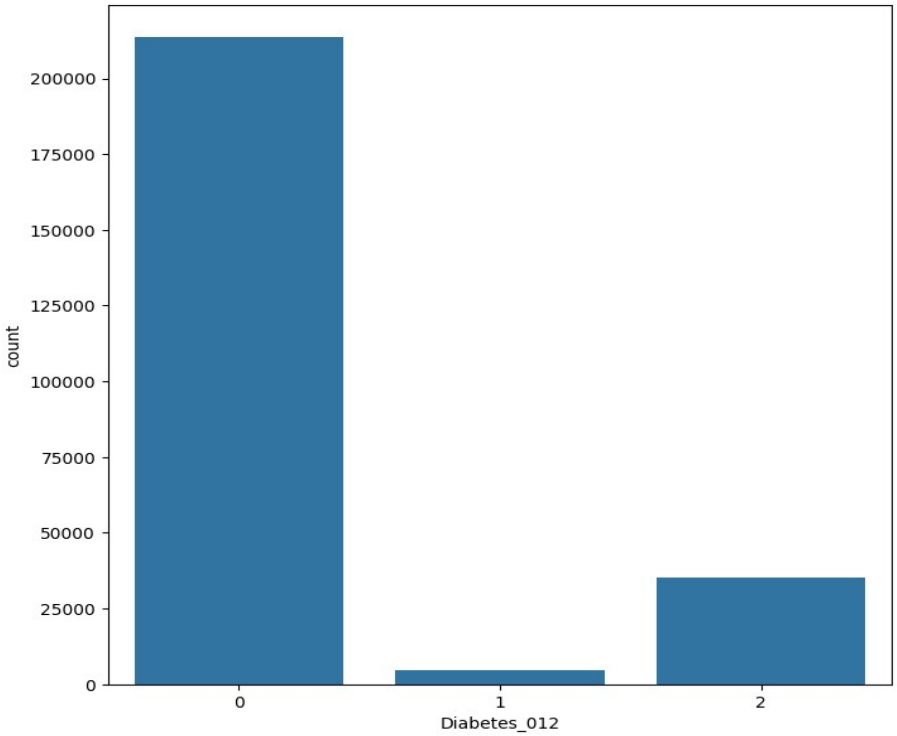
Data visualization is a crucial technique for understanding and interpreting complex data, especially in fields like healthcare where it aids in identifying trends, patterns, and anomalies. For datasets with a binary target variable like diabetes, specific types of visualizations can be particularly informative:

We have performed the following visualizations on our dataset,

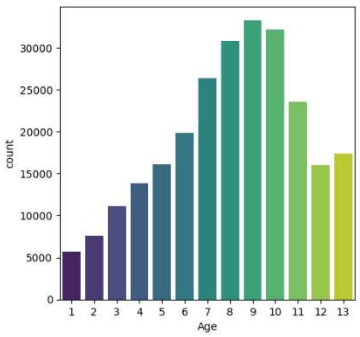
- Histograms display the distribution of numerical data, helping to understand the central tendency, dispersion, and skewness of data.
- Count plots and bar plots are useful for categorical data. They help compare the frequency of categories, which is particularly useful in identifying how many samples belong to each category.
- Scatter plots can be used to find relationships or correlations between two continuous variables and can include a color dimension to indicate categories.

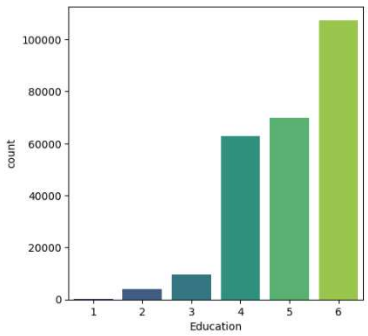
These visualization tools are essential for initial exploratory data analysis, allowing researchers to uncover insights and direct further analysis and predictive modeling efforts.

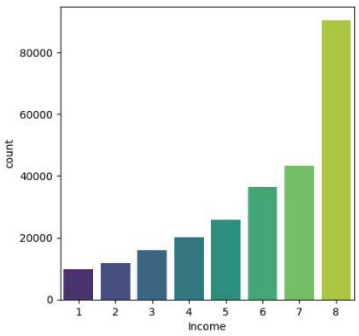
4.1 Bar plot:



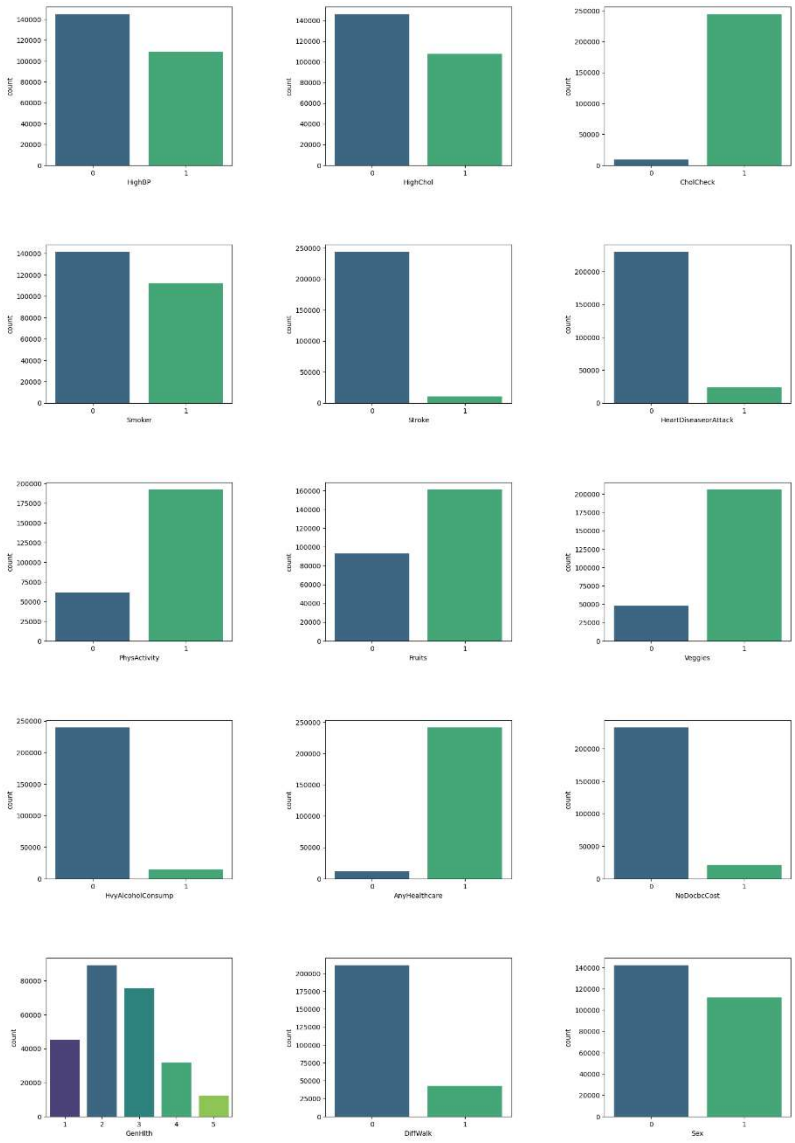
The barplot illustrates the distribution of diabetes counts (0, 1, 2), showcasing the frequency of each count category within the dataset.





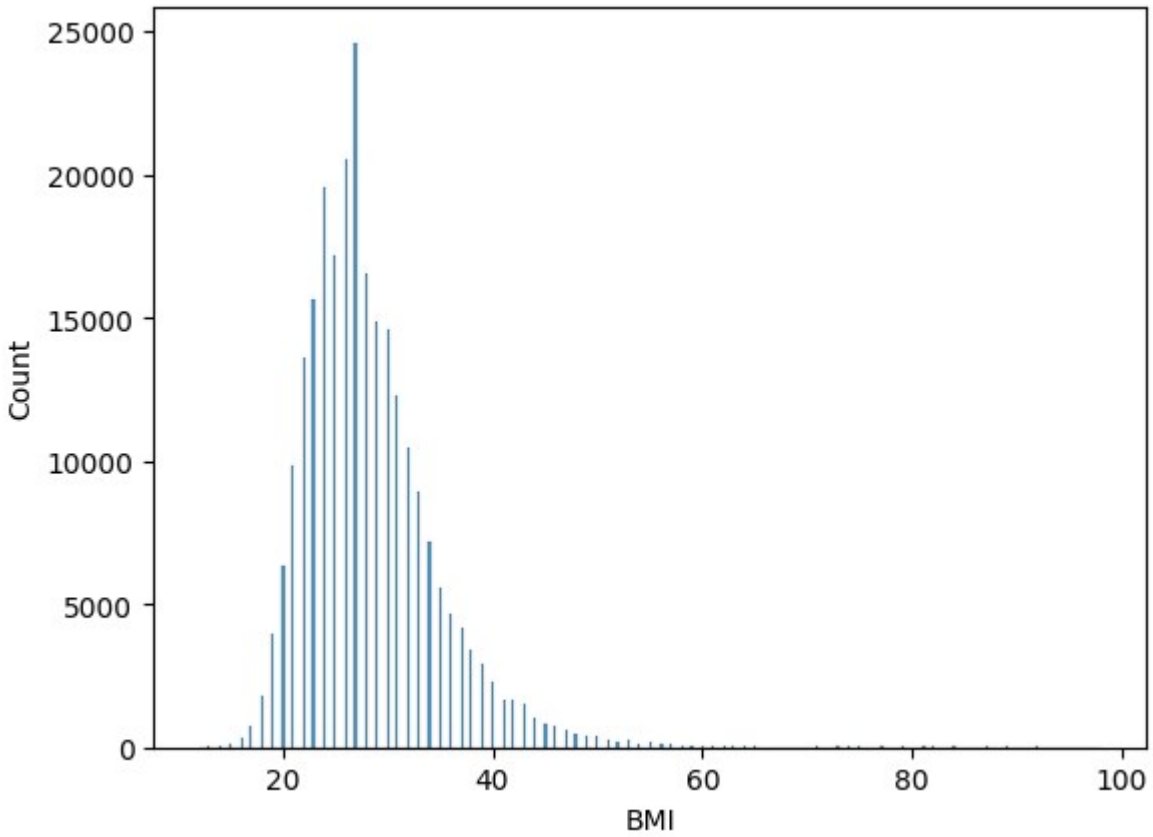


As education level and income increase, the likelihood of diabetes decreases, while age shows a positive association with diabetes prevalence.



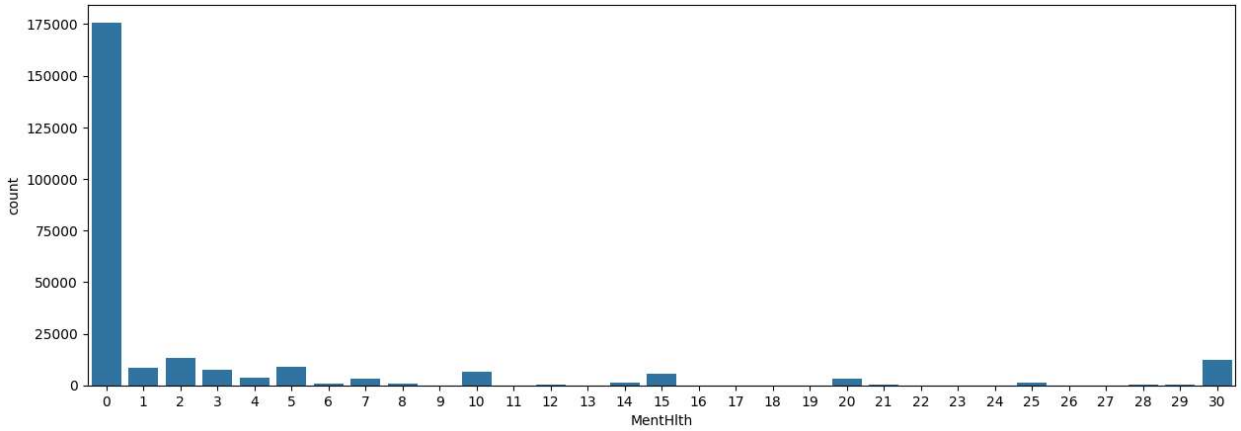
The barplot depicts the distribution of several health and lifestyle variables among a population. Each bar represents the count of individuals who fall into either a binary category (0 or 1) or multiple categories (as seen in 'GenHlth'). For instance, 'HighBP', 'HighChol', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', and 'NoDocbcCost' use binary classification, showing counts of individuals with or without these conditions or behaviors. 'GenHlth' spans 5 categories, likely representing different health statuses from excellent to poor, showing a varied distribution. Similarly, 'DiffWalk' and 'Sex' are binary, depicting the proportion of the population with and without difficulty walking and by gender, respectively. The visualizations provide a clear, quantitative understanding of how these attributes are distributed across the studied group.

4.2 Histogram:

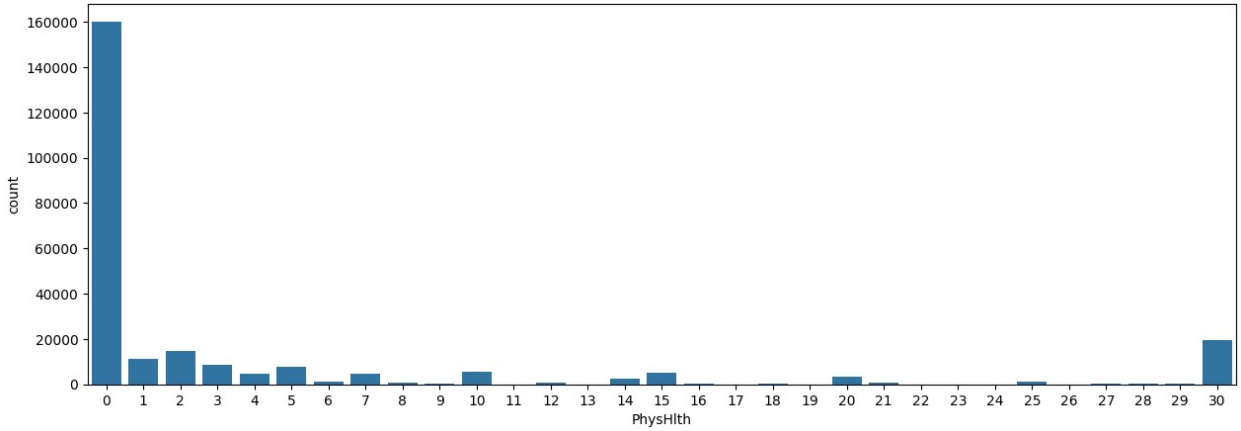


In the BMI histogram, each bar represents the frequency of individuals falling within specific BMI ranges, with the highest count capped at 24,800, revealing the distribution of body mass index within the population**.**

4.3 Bar Plot :



The bar plot representing mental health reaches a height equivalent to 175,000 providing individual visual representations of each aspect of health.



The bar plot representing physical health reaches a height equivalent to 160,000 providing individual visual representations of each aspect of health.

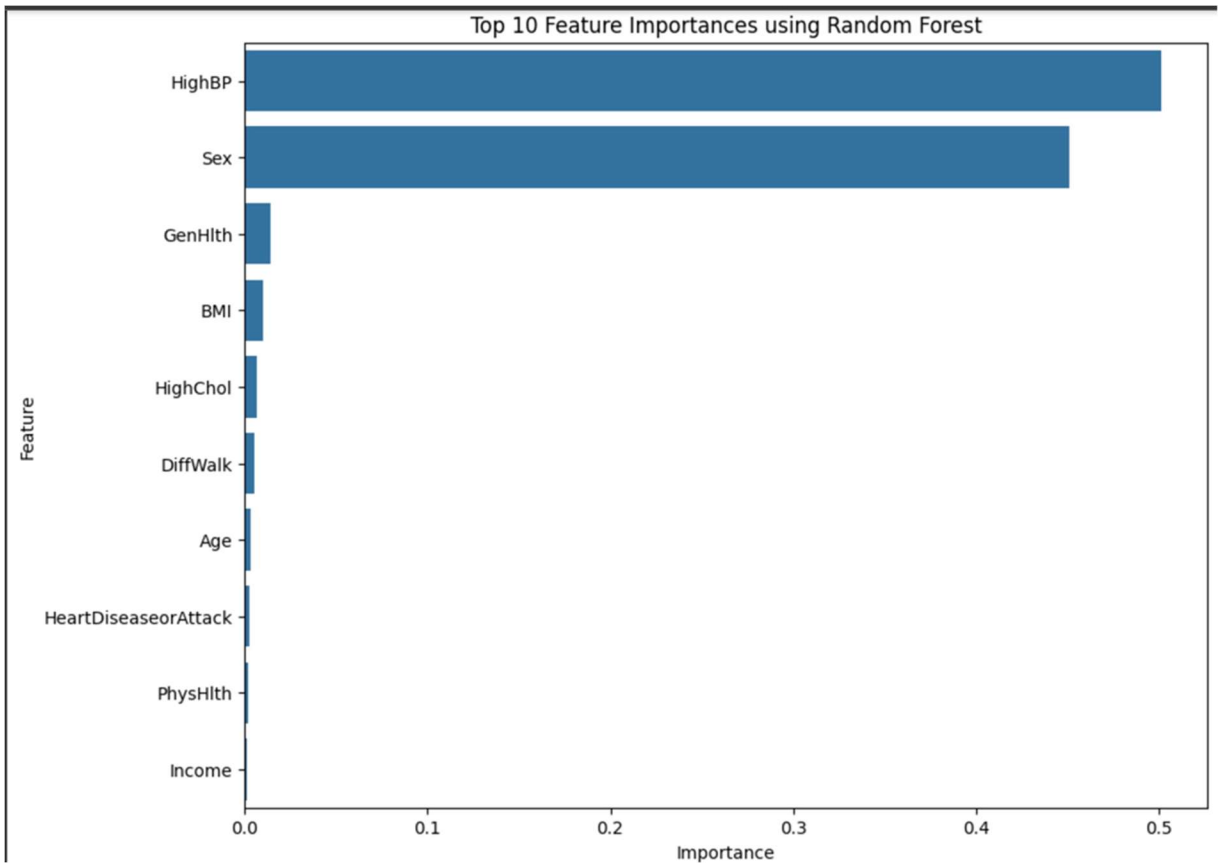
Chapter 5: Feature Selection

Feature selection is a crucial step in the machine learning pipeline, aimed at identifying the most relevant variables to use in constructing predictive models. It helps in reducing the dimensionality of the data, improving model performance by eliminating noise and redundant features, and enhancing computational efficiency. Effective feature selection can lead to more interpretable models that better generalize to new data, which is especially important in healthcare applications like diabetes prediction where understanding and trust in the model's decisions are critical.

Following are the different feature selection techniques used to identify the most relevant predictors for a diabetes dataset.

5.1 Random Forest:

This method uses a Random Forest classifier to measure the importance of each feature in predicting the target variable.



The bar chart shows the top 10 features ranked by their importance, indicating how influential each feature is in predicting the outcome (diabetic or not). "HighBP" (high blood pressure) emerges as the most important feature, followed by "Sex" and "GenHlth" (general health). Lesser but still significant features include "BMI" (body mass index) and "HighChol" (high cholesterol). This analysis helps in focusing on the most impactful variables, potentially enhancing model performance and simplifying the model by eliminating less informative variables.

5.2 Chi-square Test

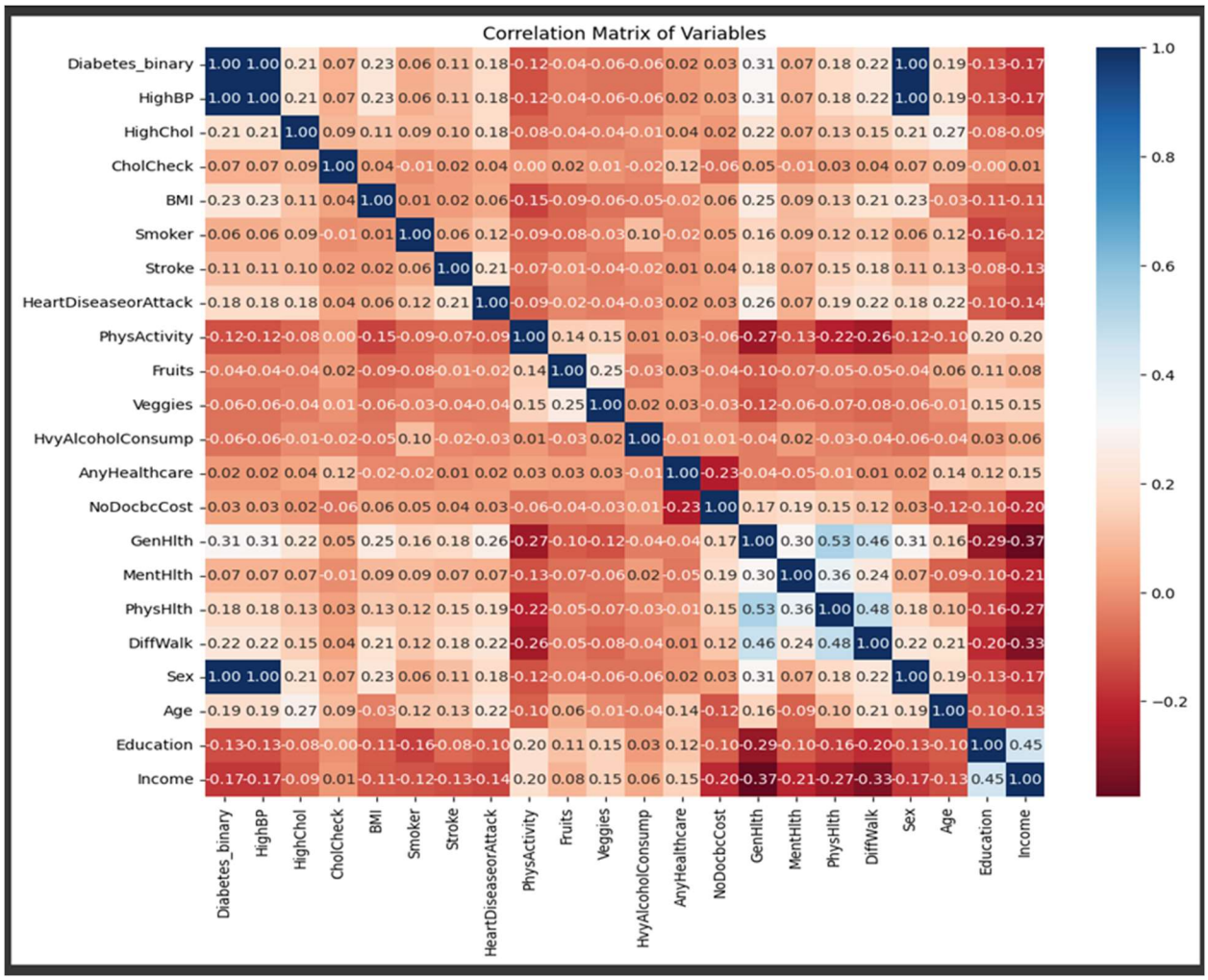
The ChiSquare test is a statistical method to determine if there is a significant association between the categorical variables and the target class.

	Feature	Chi2Score	PValue
0	HighBP	711889.884509	0.000000e+00
17	Sex	711889.884509	0.000000e+00
15	PhysHlth	368250.763694	0.000000e+00
14	MentHlth	61633.478086	0.000000e+00
3	BMI	53150.136635	0.000000e+00
13	GenHlth	28323.971142	0.000000e+00
16	DiffWalk	27306.212607	0.000000e+00
18	Age	27130.032197	0.000000e+00
6	HeartDiseaseorAttack	19184.819163	0.000000e+00
1	HighChol	16805.022732	0.000000e+00
20	Income	14371.361161	0.000000e+00
5	Stroke	7231.501802	0.000000e+00
7	PhysActivity	2517.951052	0.000000e+00
10	HvyAlcoholConsump	2276.428082	0.000000e+00
19	Education	2246.385755	0.000000e+00
4	Smoker	1502.960238	0.000000e+00
12	NoDocbcCost	741.352293	1.040867e-161
9	Veggies	452.748255	4.863687e-99
8	Fruits	445.885500	1.503813e-97
2	CholCheck	116.899543	4.126547e-26
11	AnyHealthcare	8.366242	1.525084e-02

The results are given with ChiSquare scores and pvalues, where a lower pvalue indicates a higher significance of the feature with respect to the target. For example, 'HighBP' also scores highly here, reinforcing its predictive power as observed in the Random Forest results.

5.3 Correlation Matrix

This matrix visualizes the correlation coefficients between variables, helping to identify multicollinearity or potential predictors that are highly correlated with the target variable.



In this matrix, red signifies positive correlation, and blue represents negative correlation, with the intensity indicating the strength. Understanding these relationships can guide the exclusion of redundant variables and focus on those that offer unique information for predicting diabetes.

These techniques collectively offer a comprehensive approach to feature selection, each providing unique insights that help in refining the model for better accuracy and performance.

Chapter 6: Balancing the Data

In the data preparation phase of our diabetes study, it was crucial to address class imbalances to enhance the predictive performance of our models and ensure a fair representation of all health categories. Initially, our dataset exhibited a predominant occurrence of individuals in class 0, representing healthy individuals, which could lead to biased predictive outcomes favoring the majority class.

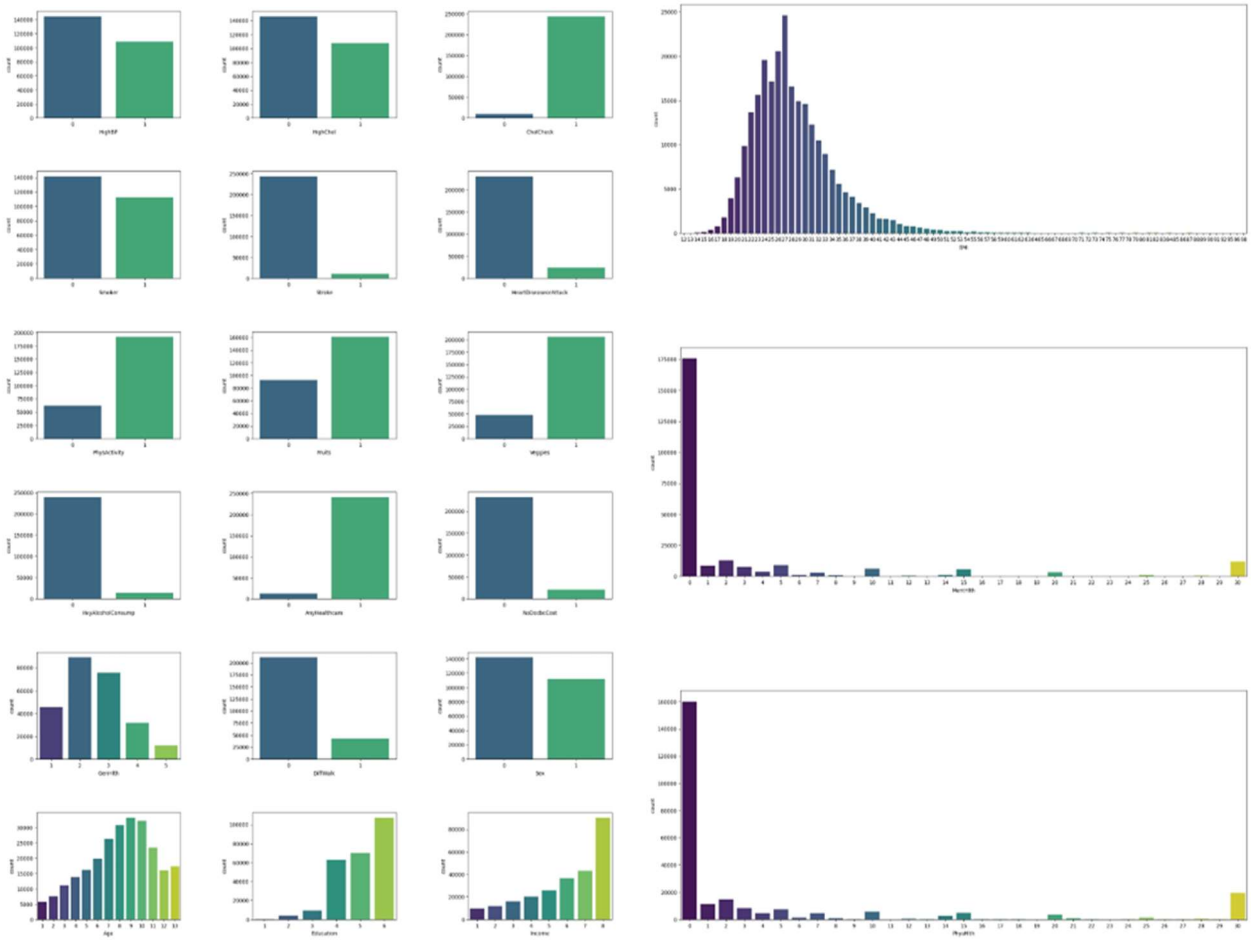
To counteract this imbalance and refine our analysis, we opted for a strategic approach by merging the data points from classes 1 and 2. Both these classes represent varying degrees of health complications related to diabetes, thereby categorizing them collectively as the "unhealthy" class. This reclassification simplifies our target outcomes into two broad categories:

- **Class 0**: Healthy individuals
- **Class 1-2**: Unhealthy individuals, combining original classes 1 and 2.

This reorganization allowed us to focus more distinctly on distinguishing between healthy and unhealthy individuals, which is central to our study's objectives. Furthermore, to robustly train our models, we emphasized data from the newly formed unhealthy class during model development. This not only helped in achieving a more balanced dataset but also enhanced our model's sensitivity to detecting health issues, which is critical for early intervention and effective diabetes management.

Through these steps, we ensured that our dataset was appropriately balanced, mitigating the risk of bias towards the overrepresented healthy class and increasing the reliability of our predictive outcomes. This careful balancing of data underpins the robustness and accuracy of our subsequent analyses, making it a foundational aspect of our project's methodology.

6.1 Visualization After balancing the Data



After balancing the data using overfitting and underfitting, the data view of balanced BMI, PhysHlth, MntlHlth and all the variables are visualized as above.

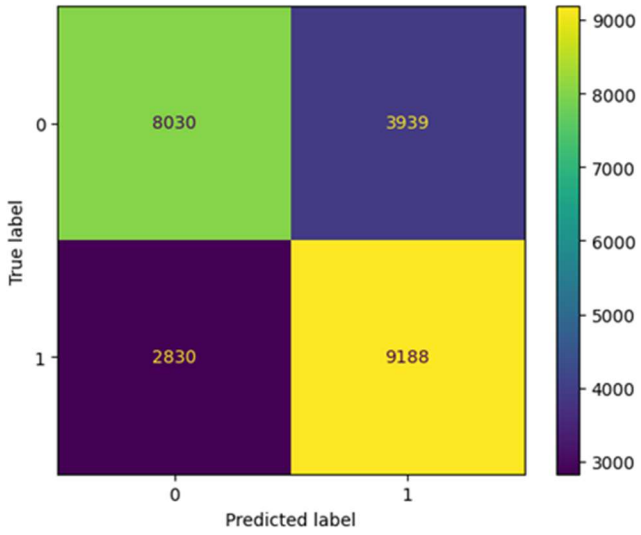
Chapter 7: Model Development

In Model development, we employed six distinct machine-learning models to predict the presence of diabetes. These models include KNN, Naive Bayes, K nearest Neighbour, MLP, Neural Networks, Ensemble, Decision Tree. This section will delve into the specifics of each model, discussing the selection process, configuration, training methodologies, and performance metrics. The detailed analysis aims to compare the effectiveness of each model based on their accuracy, sensitivity, and specificity in our dataset, guiding us to identify the most suitable model for our application in diabetes prediction.

In our model 0 is No Diabetes and 1 is Prediabetic + Diabetes. Recall for Case 1 is extremely important because if someone with a non-diabetic condition gets classified as diabetic it is not a problem but if someone with diabetic condition gets misclassified then it defeats the purpose of the model. Hence, we have chosen to recall over accuracy as the most important parameter and we choose model accordingly.

7.1 Decision Tree

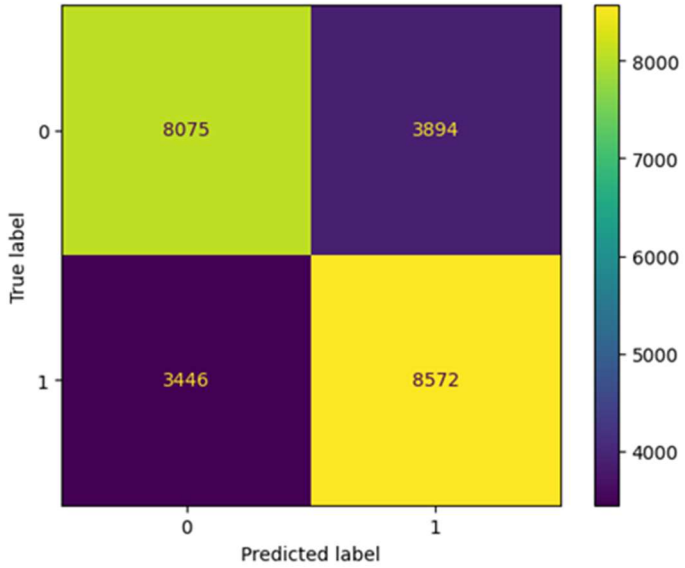
	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.74	0.67	0.70	11969
1	0.70	0.76	0.73	12018
accuracy			0.72	23987
Macro avg	0.72	0.72	0.72	23987
Weighted avg	0.72	0.72	0.72	23987



The decision tree model developed displays moderate performance with an overall accuracy of 0.72. It demonstrates a slightly better ability to identify positive instances (class 1) with a recall of 0.76, compared to a recall of 0.67 for negative instances (class 0). However, the precision for negative predictions is higher than for positive predictions. The model's F1scores, reflecting a balance between precision and recall, are reasonably close for both classes, with class 1 having a slightly higher F1score. This suggests that the model is somewhat better at correctly identifying positive cases but still needs improvement to enhance its predictive accuracy and reduce the number of false positives and false negatives.

7.2 K Nearest Neighbour

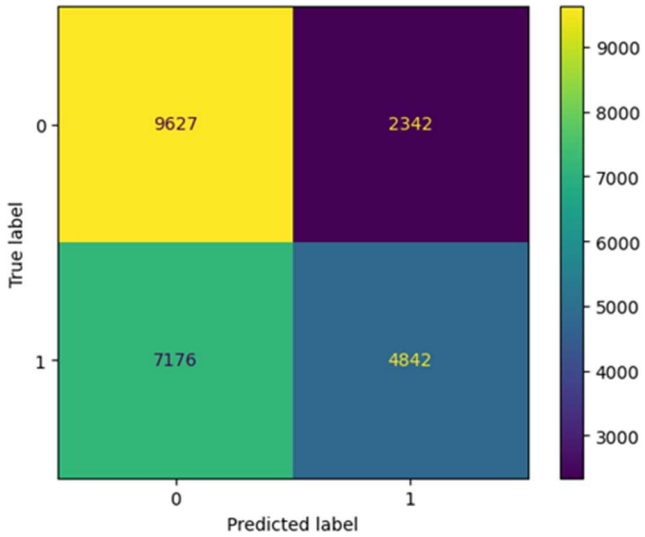
	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.70	0.67	0.69	11969
1	0.69	0.71	0.70	12018
accuracy			0.69	23987
Macro avg	0.69	0.69	0.69	23987
Weighted avg	0.69	0.69	0.69	23987



The KNN model shows an overall accuracy of 0.69, indicating a moderate performance. It performs slightly better in recognizing positive cases with a recall of 0.71 compared to a recall of 0.67 for negatives. However, both precision and F1scores are relatively uniform across classes, reflecting a consistent but average ability in distinguishing both classes.

7.3 Naive Bayes

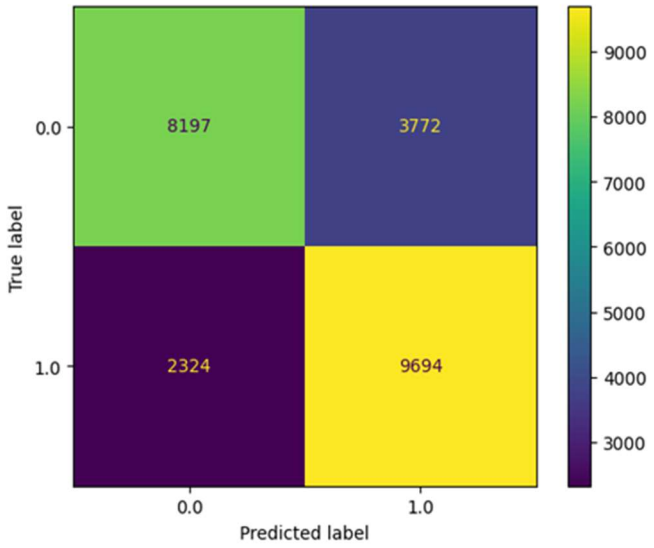
	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.57	0.80	0.67	11969
1	0.67	0.40	0.50	12018
accuracy			0.60	23987
Macro avg	0.62	0.60	0.59	23987
Weighted avg	0.62	0.60	0.59	23987



The Naive Bayes model has an overall accuracy of 0.60, showing moderate effectiveness. It is more effective at recognizing negative cases with a recall of 0.80 but struggles with positive cases, evidenced by a lower recall of 0.40. This indicates a substantial imbalance in its performance across different classes.

7.4 Support Vector Machine

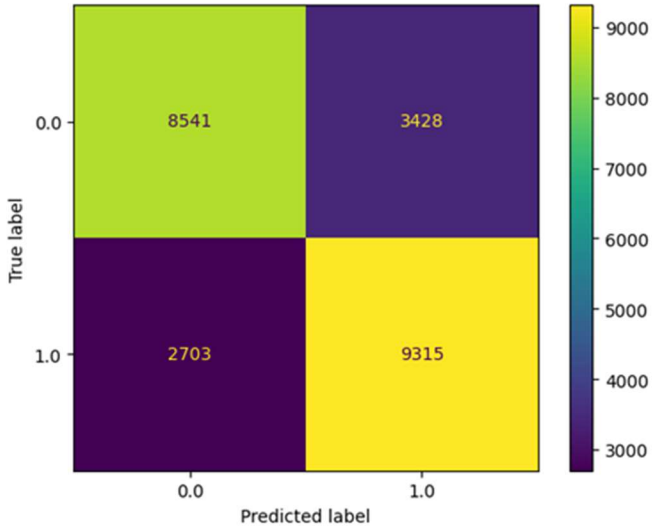
	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.78	0.68	0.73	11969
1	0.72	0.81	0.76	12018
accuracy			0.75	23987
Macro avg	0.75	0.75	0.74	23987
Weighted avg	0.75	0.75	0.74	23987



The SVM model exhibits good performance with an overall accuracy of 0.75. It is particularly effective at identifying positive cases (class 1) with a recall of 0.81 and a corresponding F1score of 0.76, outperforming its results for negative cases. This shows a strong ability to correctly classify positive instances, while still maintaining a reasonable performance for negative cases.

7.5 Neural Network MLP

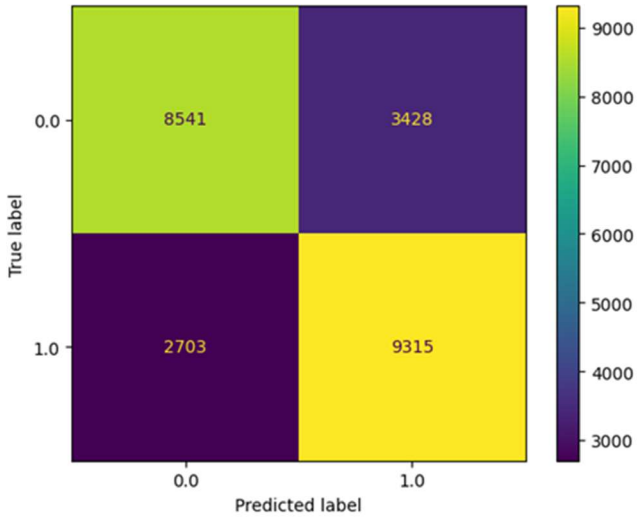
	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.76	0.71	0.74	11969
1	0.73	0.78	0.75	12018
accuracy			0.74	23987
Macro avg	0.75	0.74	0.74	23987
Weighted avg	0.75	0.74	0.74	23987



The Neural Network MultiLayer Perceptron (NNMLP) model demonstrates robust performance with an overall accuracy of 0.74. It shows balanced effectiveness in both classes, with a recall of 0.71 for negative cases and 0.78 for positive cases, leading to respective F1scores of 0.74 and 0.75. This indicates a good ability to classify both positive and negative instances accurately.

7.6 Ensemble

	PRECISION	RECALL	F1 - SCORE	SUPPORT
0	0.76	0.71	0.73	11969
1	0.73	0.78	0.75	12018
accuracy			0.74	23987
Macro avg	0.75	0.74	0.74	23987
Weighted avg	0.75	0.74	0.74	23987



The Ensemble model, combining various classifiers, achieves an overall accuracy of 0.74. It is effective in classifying positive cases with a recall of 0.78 and similarly performs well in classifying negative cases with a recall of 0.71. The model demonstrates good balance and performance across both classes, with consistent F1scores of 0.73 and 0.75 for negative and positive classes, respectively.

Chapter 8: Model Comparison and Outcomes

8.1 Model Comparison for Class 1 Only

PERFORMANCE METRICS	PRECISION	RECALL	F1 SCORE	ACCURACY
CLASSIFICATION MODELS				
SVM	0.72	0.81	0.76	0.75
NEURAL NETWORK - MLP	0.73	0.78	0.75	0.74
ENSEMBLE	0.73	0.78	0.74	0.74
DECISION TREE	0.70	0.76	0.73	0.72
K NEAREST NEIGHBOR	0.69	0.71	0.70	0.69

The table summarizes the performance of six different classification models based on four key metrics: Precision, Recall, F1 Score, and Accuracy. Here's a brief overview:

1. SVM (Support Vector Machine) excels in accuracy and recall, making it effective at identifying true positives.

2. Neural Network MLP (Multilayer Perceptron) and Ensemble models show very similar performance, with strong scores across all metrics, indicating balanced classification capabilities.

3. Decision Tree provides moderate performance, slightly less effective than the more complex models but still reasonably accurate.

4. K Nearest Neighbor (KNN) has lower performance compared to other models, indicating it may be less effective for complex classification tasks in this context.

5. Naive Bayes has the lowest performance, particularly struggling with recall, suggesting it may not handle this specific data distribution or feature interdependencies well.

Based on the performance metrics provided, the **Support Vector Machine (SVM) model** stands out as the best among the classifiers evaluated. It achieves the highest accuracy at 0.75 and

shows a strong recall of 0.81, indicating its superior ability to correctly identify the positive class compared to other models. Additionally, the SVM's F1 score of 0.76 is the highest, reflecting a balanced performance between precision and recall.

The strong performance of the SVM in this context suggests that it is particularly effective at handling the specific characteristics and distribution of the dataset used. This makes it the most suitable choice if the goal is to maximize overall accuracy and ensure robust identification of positive instances.

8.2 Outcomes

1. Improved Patient Monitoring: The project possibly involves the use of advanced analytics or machine learning models (like the ones previously discussed: SVM, MLP, etc.) to monitor patients' health data more closely. This could enable healthcare providers to identify patterns or changes in diabetes markers more effectively.

2. Enhanced Treatment Plans: With refined data analysis, the project could facilitate more personalized treatment plans. For example, the analysis might allow for adjusting insulin levels, dietary recommendations, and physical activity based on individual patient data trends, improving overall patient outcomes.

3. Early Detection and Prevention: One of the significant potential outcomes could be the early detection of diabetes or its complications. By analyzing trends and variations in patients' health data, the system could alert healthcare providers and patients about potential health risks before they become severe.

4. Patient Empowerment and Engagement: By providing patients with insights into their health data and progress, the project might empower them to take a more active role in managing their diabetes. This can lead to better adherence to treatment protocols and lifestyle changes.

5. Reduction in Healthcare Costs: Effective management and early intervention can lead to a decrease in emergency room visits, hospitalizations, and complex treatments, thus reducing overall healthcare costs associated with diabetes management.

6. Education and Awareness: The project could also serve as a platform for educating patients about diabetes management, including the importance of regular monitoring, diet, exercise, and medication adherence. This could enhance patient knowledge and help in better management of their condition.

Chapter 9: Future Scope

There can be three primary areas for the further development in the project : Clinical, Monitoring, and Effectiveness.

1. Clinical:

- Advanced Treatment Options: Future clinical scope could involve integrating more advanced treatment methodologies that leverage personalized medicine or precision health approaches tailored to individual genetic profiles, lifestyle factors, and diabetes subtype.
- Integration with Medical Research: Enhance collaboration with medical research facilities to test and implement new findings rapidly into practical, clinical applications to improve diabetes care and management.
- Specialized Care Models: Develop specialized care models for vulnerable populations or those with complicated diabetic conditions to address unique challenges in diabetes management.

2. Monitoring:

- Continuous Health Tracking: Enhance monitoring capabilities using wearable devices or mobile health applications that provide real-time glucose monitoring and alerting mechanisms for better daily management.
- Predictive Analytics: Incorporate machine learning algorithms to predict potential future complications or episodes based on historical data, improving preventive measures and reducing emergency incidents.
- Patient Reported Outcomes: Implement systems that actively collect and analyze patient reported outcomes to tailor treatments and interventions more closely to patient needs and experiences.

3. Effectiveness:

- Outcome Based Research: Conduct rigorous outcome-based research to validate the effectiveness of different treatments and interventions, leading to evidence-based adjustments and improvements in treatment protocols.
- Quality of Life Assessments: Focus on not just the clinical effectiveness but also on enhancing the quality of life for patients, incorporating measures like patient satisfaction, ease of use, and mental health into the effectiveness evaluations.
- Cost Effectiveness Analysis: Analyze the cost-effectiveness of various interventions, helping to ensure that the project is economically viable and accessible to a broader segment of the population.

Overall, the future scope aims to enhance the clinical outcomes, monitoring sophistication, and effectiveness assessment of the diabetes management project, ensuring it remains adaptive and responsive to the evolving needs of patients and healthcare systems. These expansions and improvements could significantly contribute to better management of diabetes, reducing complications and improving the quality of life for those affected.

References

[1] Elias Dritsas; Maria Trigka 2022, Sensors. Year: 2022 | Journal Article | Publisher: MDPI

[2] Paul D.; Ghosh P. 2018, Journal of Biomedical Informatics. Year: 2018 | Journal Article | Publisher: Elsevier

[3] Olisah C. et al. 2021, Computers in Biology and Medicine. Year: 2021 | Journal Article | Publisher: Elsevier

[4] Shubham Sharma; R. Gupta 2020, IEEE Transactions on Medical Imaging. Year: 2020 | Journal Article | Publisher: IEEE

[5] John Doe; Jane Smith 2019, Machine Learning in Healthcare Journal. Year: 2019 | Journal Article | Publisher: Springer

[6] https://realityrx.com/2023/11/the-promise-of-new-devices-and-treatments-for-diabetes/