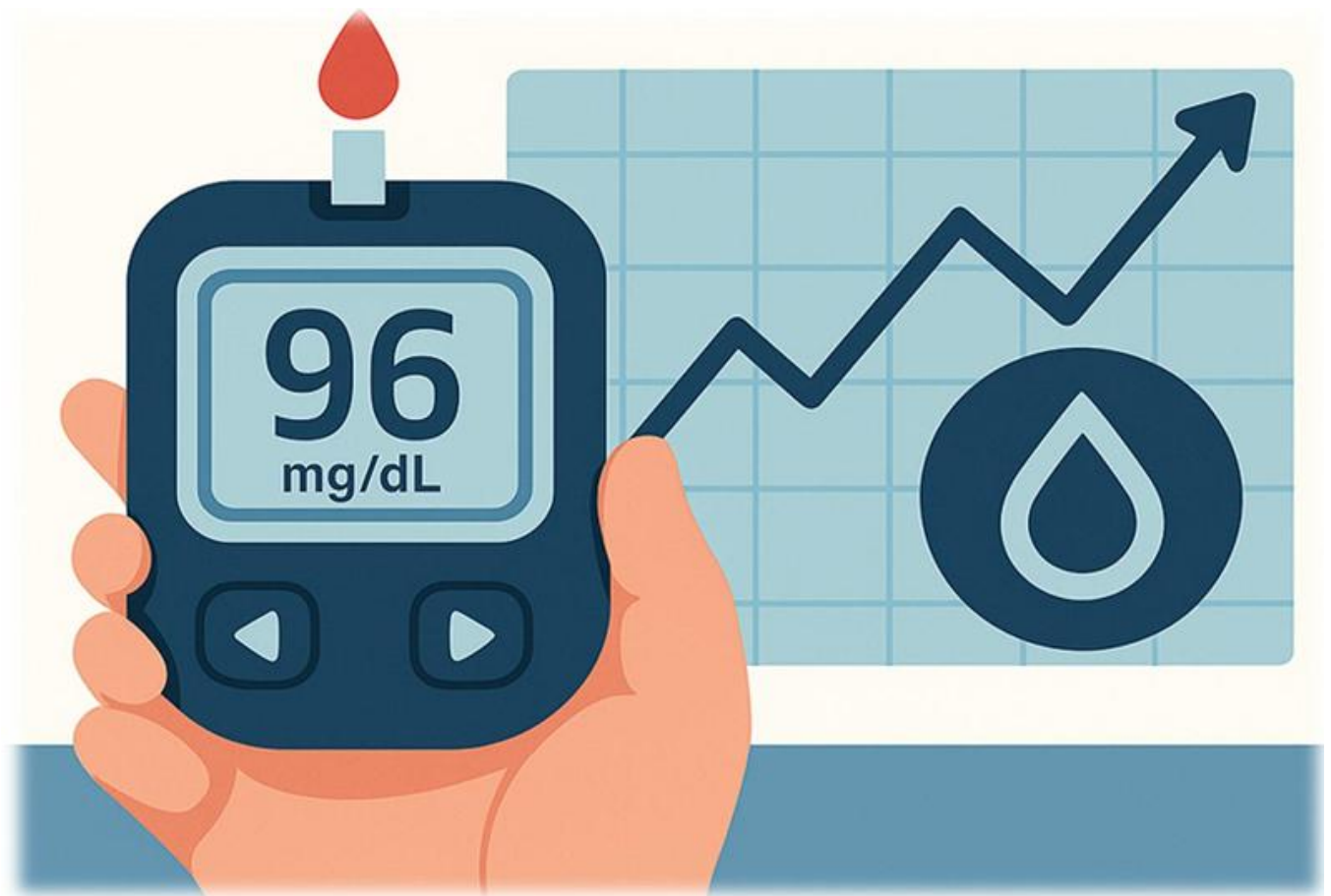


# ANALYSIS OF DIABETES STATUS

H.R.K.S.H Bandara  
s16290



## Abstract

This study addresses the critical need toward well-executed and accurate early prediction of diabetes, a popular global health challenge, and this study uses machine learning techniques to identify the relationship of factors with diabetic status and to build a model to predict diabetes status of a patient. The primary problem involves classifying people as diabetic ('Y'), pre-diabetic ('P'), or non-diabetic ('N') based on various physiological along with demographic indicators from the 'Diabetes .csv' dataset.

The methodology included within its thorough data science pipeline for initial data collection. A thorough pre-processing phase was conducted, categorical variables such as "Gender" and "CLASS" were handled, some minor inconsistencies in it were addressed, and ensured data types would be appropriate for modeling with exploratory data analysis to understand variable distributions and relationships. The problem description mentioned about clustering. The Python code that was provided focused instead on classification that was direct. For model building as well as evaluation, a comparative analysis of Naive Bayes, Logistic Regression, Decision Tree Classifier, and specifically Support Vector Classifier (SVC) algorithms were conducted. Of these evaluated models, the Decision Tree Classifier demonstrated superior predictive performance because it achieved the highest accuracy of 98%, exceeding SVC (93.5%), Naive Bayes (94.25%), and Logistic Regression (82%). This accuracy highlights the fact that machine learning models, especially Decision Trees, can identify people at risk of diabetes, and they offer a valuable tool because they diagnose early and they manage healthcare in a proactive way, based on readily available clinical parameters.

# Contents

Abstract .....	i
List of Figures .....	iv
List of Tables .....	v
Introduction.....	1
Literature Review.....	2
Theory and Methodology.....	3
1. Theoretical Foundations .....	3
1.1. Classification Algorithms .....	3
1.2. Model Evaluation Metrics .....	3
2. Methodology .....	4
2.1. Dataset Description and Acquisition .....	4
2.2. Data Pre-processing.....	4
2.3. Data Splitting.....	5
2.4. Model Building and Evaluation.....	5
Data .....	6
Explanatory Data Analysis .....	7
1. Univariate Analysis .....	7
1.1 Quantitative variable analysis.....	7
1.2 Qualitative variable analysis.....	12
Bi-Variate analysis .....	12
2.1 Categorical variable analysis .....	12
2.2 Quantitative analysis.....	13
2.3 Categorical Vs Quantitative analysis.....	14
Advanced Data Analysis.....	19
Predictive modeling and Performance Evaluation .....	19
Results and Identification of the Best Model .....	20
Identification of the best model .....	21
Feature Importance from the best model .....	22
General Discussion and Conclusion .....	24
Key Findings and their relationship to diabetes status.....	24
Model Performance and Comparison.....	24
Conclusion.....	25
Limitations and Future work .....	25

References .....	26
------------------	----

## List of Figures

Figure 1 - KDE Plots of AGE, HbA1c, Chol, and BMI .....	8
Figure 2 - KDE Plots of Urea, Cr, TG, HDL, LDL and VLDL.....	11
Figure 3 - Distribution of Gender .....	12
Figure 4 - Distribution of Diabetes status .....	12
Figure 5 - Heatmap of Gender vs. CLASS .....	13
Figure 6 - Correlation matrix .....	13
Figure 7 - Boxplot of Age by gender .....	14
Figure 8 - Boxplot of BMI by Gender .....	15
Figure 9 - Boxplot of Cholesterol level by gender .....	15
Figure 10 - Boxplot for HbA1c by gender.....	16
Figure 11 - Boxplots for patients' diabetes status by age.....	16
Figure 12 - Box plots for patient's diabetes status with BMI.....	17
Figure 13 - Boxplots for Patient's blood sugar level by HbA1c levels.....	17
Figure 14 - Boxplots for patients' diabetes status and Cholesterols level .....	18
Figure 15 - Feature Importance from Decision Tree .....	22

## List of Tables

Table 1 - Table about the variables used .....	6
Table 2 - Descriptive statistics of numerical features .....	7
Table 3 - Comparative Accuracy of Selected Machine Learning Models .....	21
Table 4 - Feature importance table .....	22

# Introduction

Diabetes is a chronic disease characterized by dysregulated carbohydrates and occurs when the pancreas does not produce enough insulin which in turn makes the metabolism of carbohydrate abnormal and raises the levels of glucose in the blood. Intensify thirst, intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Given that current pharmacological interventions predominantly manage symptoms rather than offering a definitive cure, early and accurate identification of diabetes status is paramount for effective disease prevention and patient management. The level of blood sugar affects cholesterol level, weight, height, pressure level and more of a person. So early identification is the only remedy to stay away from the complications of diabetes.

Numerous researchers have explored various techniques to identify the diabetic status or diagnose the disease like Naïve Bayes, Support vector machine, Decision Tree, Random Forest, Logistic regression etc. and each possessing distinct strengths and weaknesses in predictive modeling.

With this foundational works this study utilizes a comprehensive 'Diabetes.csv' dataset to address the objectives to identify the significant physiological and demographic factors influencing a patient's diabetes status and to build a highly accurate model to predict the diabetic status of a patient. In this research Support vector machine, Decision Tree, Naïve Bayes classifiers and Logistic regression are used and evaluated on the dataset to build an accurate model to predict diabetic status of a patient.

## Literature Review

Diabetes is one of the major diseases that is presented all around the world. Diabetes causes many complications including cardiovascular disease, neurotherapy and kidney, so identification of diabetes level which is pre diabetic, non-diabetic or diabetic and the association between diabetes status with other variables makes people to be aware of the complications that they will have. So, this report access core objectives: [1] Identifying association between diabetes status with the BMI, cholesterol, HbA1c level etc. and [2] building a model which predicts diabetes status. There are 3 research which is very much closer to the report and those will provide background context to model development, statistical analysis and further to understanding the relationship between the diabetes status and demographic results and bio medical results of a person.

Deepthi Sisodia and Deelip Sign Sisodi (2018) used PIMA Indian diabetes data set to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. They used Decision Tree , Naïve Bayes and SVM machine leaning techniques and found that model used Naïve Bayes technique gives more accurate results than other 2 techniques .But the dataset set that use to model is in limited size . Sometimes the predictions taken from the model may vary with another population.

Hafiz Farooq Ahmad,Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman and Abdulaziz Alhumam (2021)used datasets from Frankfurt hospital (Germany) and Pima Indian dataset to investigate the prediction of diabetic patients and compare the role of HbA1c and FPG as input features. Here they have used Use of 5 machine learning classifiers (Three simple learners: Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree (DT) and two ensemble learners: Random Forest (RF) and Ensemble Majority Voting (EMV) ) to analyze the effect of the choice of the HbA1c or FPG labeling attributes on the datasets with the remaining attributes common between the two datasets. Each classifier was evaluated against both datasets. As the key findings, SVM has an accuracy of 82.10 % (HbA1c labeled data) and Random Forest has an accuracy of 88.27%(FPG-labeled data). LDL, Hypertension, Height, Age, Physical Activity Level (PAL) mostly affect diabetes. Here Although data concerned 3000 patients, the final size of data was very small about 162 with complete feature values which arises a question about the accuracy and predictions.

Aishwarya Mujumdara and Dr. Vaidehi V(2020) build a predictive model using machine learning algorithms and data mining techniques for diabetes prediction. They have collected data when building the model. As conclusions they have found that the Logistic Regression gives highest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. But the Representativeness and biasness, might have effects when using the model for other datasets because the accuracy of the model is much higher.

In conclusion, the reviewed studies show that different machine learning methods, including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, Logistic Regression, Random Forest, and AdaBoost, can be used to predict the presence of diabetes. Regarding dataset size, feature selection, and model accuracy, each study identifies unique advantages and disadvantages. Incomplete data, small sample sizes, and worries about generalizability to larger populations are common issues noted. These findings underline the importance of creating a strong and thoroughly validated prediction model that makes use of a variety of demographic and biomedical characteristics.



# Theory and Methodology

This section follows a thorough explanation of the methodological steps taken to accomplish the goals which are developing a model to predict a patient's diabetes status and determining the relationship between factors and diabetes status. Hence the section provides an overview of the theoretical underpinnings of machine learning and statistical approaches used in this study.

## 1. Theoretical Foundations

### 1.1. Classification Algorithms

The core of this study involves classification, a supervised machine learning task aimed at predicting a categorical output variable (the 'CLASS' variable: 'Y', 'P', 'N') based on input features. Various algorithms were explored due to their distinct approaches to pattern recognition and decision-making.

- **Decision Tree Classifier:** Non-parametric supervised learning techniques for regression and classification are called decision trees. They create a decision tree-like structure by dividing the data into subsets according to feature values. Every internal node denotes an attribute test, every branch denotes the test's result, and every leaf node denotes a class label. The objective is to learn basic decision rules derived from the data features to build a model that forecasts the value of a target variable. Decision trees can capture non-linear relationships and are simple to understand. However, if they are not appropriately regularized (for example, by restricting tree depth), they may be vulnerable to overfitting.
- **Support Vector Classifier (SVC):** Support Vector Machines (SVMs) are powerful supervised learning models used for classification and regression. An SVC works by finding an optimal hyperplane that best separates data points of different classes in a high-dimensional space. The "support vectors" are the data points closest to the hyperplane, which are critical in defining the decision boundary. SVC can handle complex, non-linear relationships using various kernel functions (e.g., linear, polynomial, radial basis function (RBF)), which implicitly map the inputs into high-dimensional feature spaces.
- **Logistic Regression:** Despite its name, Logistic Regression is a linear model used for binary or multi-class classification. It models the probability of a categorical outcome using a logistic function (sigmoid function). While it is a linear classifier, it provides probabilities which can be very useful for decision-making. Its interpretability, where coefficients indicate the impact of each feature on the log-odds of the outcome, makes it a popular choice in medical and social sciences.
- **Naive Bayes Classifier:** Naive Bayes classifiers are a family of probabilistic algorithms based on Bayes' theorem, with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. They are particularly effective for high-dimensional datasets and are known for their simplicity, speed, and good performance in text classification and spam filtering, often serving as a strong baseline model.

### 1.2. Model Evaluation Metrics

To assess the performance of classification models, several metrics are crucial. While **accuracy** (the proportion of correctly predicted instances) was a primary metric, it is understood that for imbalanced datasets common in medical diagnostics, other metrics are also vital. For this study, the models'

performance was primarily compared based on accuracy, indicating the overall correctness of the predictions.

## 2. Methodology

The analytical process followed a standard machine learning pipeline, encompassing data acquisition, pre-processing, data splitting, model building, and evaluation. The analysis was conducted using Python programming language with libraries such as pandas for data manipulation, seaborn and matplotlib.pyplot for visualization, and scikit-learn for machine learning algorithms and utilities.

### 2.1. Dataset Description and Acquisition

The study utilized the 'Diabetes .csv' dataset. This dataset contains various physiological and demographic indicators of patients, designed for the prediction of diabetes status. Key features include 'Gender', 'AGE', 'Urea', 'Cr', 'HbA1c', 'Chol', 'TG', 'HDL', 'LDL', 'VLDL', 'BMI', and the target variable 'CLASS', which categorizes patients as 'N' (Non-Diabetic), 'P' (Pre-Diabetic), or 'Y' (Diabetic). The dataset was loaded into a panda DataFrame for analysis.

### 2.2. Data Pre-processing

Thorough data pre-processing was essential to prepare the raw data for machine learning algorithms.

- **Handling Categorical Variables:**

- The 'Gender' column, being a categorical variable (e.g., 'F' and 'M'), was converted into numerical representations using one-hot encoding. This creates new binary columns for each unique category, allowing machine learning algorithms to process them.
- The target 'CLASS' variable (Nominal: 'N', 'P', 'Y') was also mapped to numerical equivalents (e.g., 'N' to 0, 'P' to 1, 'Y' to 2 or similar encoding derived from the Class\_Code

```
from sklearn.preprocessing import LabelEncoder  
  
le = LabelEncoder()  
data['Gender'] = le.fit_transform(data['Gender'])  
data_encoded = pd.get_dummies(data, columns=['Gender'])
```

and  
steps

names and ensuring that all features were of appropriate numerical data types for model training. This involved inspecting data types (data.info()) and potentially coercing columns if they were not correctly parsed.

- **Exploratory Data Analysis (EDA):** Initial EDA was performed to gain insights into the dataset's structure, distributions of individual variables, and relationships between them. This involved examining descriptive statistics (data.head(), data.describe()) and visualizing data (though specific plots were not detailed in the provided code snippet). EDA helped in understanding the presence of outliers, data skewness, and the overall quality of the dataset before modeling.

conversion shown in the Python script), which is necessary for classification algorithms.

- **Addressing Inconsistencies Data Types:** The pre-processing included standardizing column

## 2.3. Data Splitting

The dataset was split into training and testing sets to ensure an unbiased evaluation of the models' performance on unseen data. A common split ratio of 80% for training and 20% for testing was used, as indicated by `test_size=0.2` in the `train_test_split` function. This randomized splitting ensures that both sets retain similar statistical properties.

Here, `X` represents the feature matrix (independent variables), and `y` represents the target variable (diabetes

```
# Train-test split (80/20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

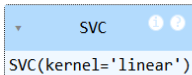
status).

## 2.4. Model Building and Evaluation

After data preparation, various machine learning models were instantiated, trained on the `x_train` and `y_train` sets, and then evaluated on the `x_test` and `y_test` sets.

- **Model Selection:** The study systematically evaluated multiple classification algorithms:
  - Support Vector Classifier (SVC)
  - Logistic Regression
  - Decision Tree Classifier
  - Naive Bayes Classifier (e.g., Gaussian Naive Bayes for continuous data)
- **Training:** Each selected model was trained by calling its `fit()` method on the training data (`x_train`, `y_train`). This step involved the algorithm learning the underlying patterns and relationships between the input features and the diabetes status. For example, for SVC:

```
x_train , x_test , y_train, y_test = train_test_split(x,y,test_size=0.2)
from sklearn.svm import SVC
model= SVC(kernel='linear')
model.fit(x_train,y_train)
```

A screenshot of a Jupyter Notebook cell. The cell has a blue header bar with the text 'SVC' and two circular icons. Below the header, the code 'SVC(kernel='linear')' is displayed in a monospaced font.

- **Evaluation:** Model performance was primarily assessed using **accuracy** on the unseen test set. The `score()` method was used to calculate the accuracy of each trained model:

Example for SVC

```
model.score(x_test,y_test)
```

0.94

The comparative analysis of these accuracy scores showed that the **Decision Tree Classifier achieved the highest accuracy of 98%**, followed by Naive Bayes (94.25%), SVC (93.5%), and Logistic Regression (82%) among the specifically compared models. This comparison informed the selection of the Decision Tree Classifier as the most effective model for this prediction task.

## Data

The data sets used were taken from public database. The dataset contains data about bio medical measurements and demographic characteristics of 1000 patients. Here is the detailed description about the variables used.

Table 1 - Table about the variables used

Variable used	Description
<b>Gender</b>	The gender of the patient (F for Female, M for Male).
<b>Age</b>	The age of the patient in years.
<b>Urea</b>	Urea level in the blood (likely measured in mg/dL or mmol/L). Urea is a waste product of protein metabolism and can indicate kidney function.
<b>Cr</b>	Creatinine level in the blood (likely measured in mg/dL or $\mu\text{mol/L}$ ). Creatinine is another waste product that indicates kidney function.
<b>HbA1c</b>	Glycated hemoglobin, a measure of average blood sugar levels over the past 2-3 months (expressed as a percentage).
<b>Chol</b>	Cholesterol level in the blood (likely measured in mg/dL or mmol/L). This typically refers to total cholesterol.
<b>TG</b>	Triglycerides level in the blood (likely measured in mg/dL or mmol/L). Triglycerides are a type of fat in the blood.
<b>HDL</b>	High-density lipoprotein cholesterol level (often called "good" cholesterol, measured in mg/dL or mmol/L).
<b>LDL</b>	Low-density lipoprotein cholesterol level (often called "bad" cholesterol, measured in mg/dL or mmol/L).
<b>VLDL</b>	Very low-density lipoprotein cholesterol level (measured in mg/dL or mmol/L).
<b>BMI</b>	Body Mass Index, a measure of body fat based on height and weight (calculated as weight in kilograms divided by height in meters squared).
<b>CLASS</b> <b>(Predictor variable)</b>	The class label indicating the diabetes status of the patient. The possible values seem to be: <ul style="list-style-type: none"><li>○ N: Non-diabetic</li><li>○ P: Prediabetic</li><li>○ Y: Diabetic</li></ul>

# Explanatory Data Analysis

This section details the Exploratory Data Analysis (EDA) conducted on the 'Diabetes .csv' dataset. The purpose of this preliminary investigation was to thoroughly understand the dataset's characteristics, identify data quality concerns, and uncover key relationships between physiological factors and diabetes status.

Here the dataset includes 1000 records from that 565 males and 435 female patients have considered for analysis purpose. Only CLASS and Gender variables are qualitative, and all the other remaining 10 variables are quantitative.

## 1. Univariate Analysis

### 1.1 Quantitative variable analysis

Following the initial data overview, a comprehensive descriptive statistical analysis was performed on all numerical features within the dataset. This analysis, summarized in Table 1 below, provides key insights into the central tendency, variability, and distribution characteristics of each physiological and demographic factor. Understanding these properties is crucial for identifying potential data quality issues, assessing the spread of data, and informing subsequent pre-processing steps, such as outlier handling or feature scaling.

Table 2 - Descriptive statistics of numerical features

	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
Count	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Mean	53.528	5.125	68.943	8.281	4.863	2.350	1.205	2.609	1.855	29.578
Std	8.799	2.935	59.985	2.534	1.302	1.401	0.660	1.115	3.664	4.962
Min	20.000	0.500	6.000	0.900	0.000	0.300	0.200	0.300	0.100	19.000
25%	51.000	3.700	48.000	6.500	4.000	1.500	0.900	1.800	0.700	26.000
50%	55.000	4.600	60.000	8.000	4.800	2.000	1.100	2.500	0.900	30.000
75%	59.000	5.700	73.000	10.200	5.600	2.900	1.300	3.300	1.500	33.000
max	79.000	38.900	800.00	16.000	10.300	13.800	9.900	9.900	35.000	47.750

The count column indicates the number of non-null observations on each variable and here for all the variables the count is 1000 which indicates there is no missing values in the dataset. The AGE feature indicates the diverse patient demographics from 20 years to 79 years with mean age is approximately 54 and median age is 55 years which shows a balanced distribution approximately, suggesting wide range in the dataset. When considered about the HbA1c level, the average 8.281% with values ranging from 0.9% to 16.00%. It has given that HbA1c is a key indicator for diabetes diagnosis and management, this range reflects that considered patients in the dataset includes across normal , pre-diabetic and diabetic ranges . The 75<sup>Th</sup> percentile at 102% further emphasizes the prevalence of elevated HbA1c levels. The Body Mass Index(BMI) ranges from 19 to 47.75 which a mean of 29.58 . The 75<sup>th</sup> percentile at 33.0 indicates that significant portion of the patients is in the overweight or obese category, a well-known risk factor for diabetes.

A critical insight from the descriptive statistics was the strong indication of outliers in several features, particularly identified by substantial differences between the 75th percentile (Q3) and the maximum values, coupled with high standard deviations. The Cr(Creatinine level) variable exhibits a considerable spread, with a maximum value of 800.00, which is significantly higher than its 75th percentile of 73.00 and a mean of 68.94. This wide disparity strongly suggests the presence of extreme outliers, which may represent genuine extreme cases or potential data entry errors. Similarly, VLDL shows a maximum value of 35.00 compared to a 75th percentile of only 1.50, clearly indicating the presence of outliers that heavily influence the maximum value and the standard deviation. While less extreme than Cr and VLDL, Urea (max 38.90 vs. 75th percentile 5.70) and TG (max 13.80 vs. 75th percentile 2.90) also present values that deviate significantly from the interquartile range, suggesting the need for outlier consideration.

These initial descriptive findings underscore the varied nature of the patient data and highlight specific areas (like outliers in 'Cr' and 'VLDL') that require careful attention during the subsequent data pre-processing phase to ensure the robustness and accuracy of the predictive models. The diverse ranges and distributions also support the need for effective feature scaling before applying certain machine learning algorithms.

To further understand the characteristics of the numerical features and identify their underlying distributions, Kernel Density Estimate (KDE) plots, overlaid on histograms, were generated for all continuous variables. These visualizations provide insights into the shape of the data, the presence of modes (peaks), skewness, and potential areas of high or low density. Figure 1 presents the distributions for 'AGE', 'HbA1c', 'Chol', and 'BMI'.

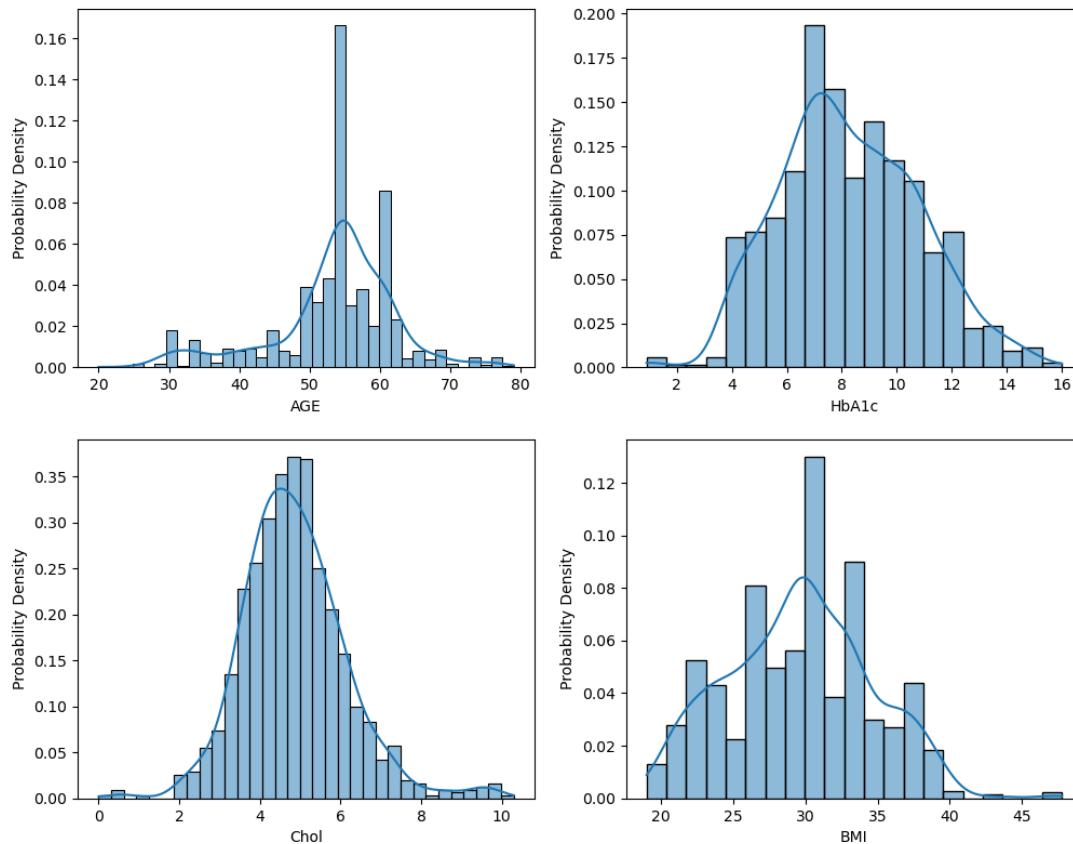


Figure 1 - KDE Plots of AGE, HbA1c, Chol, and BMI

- **AGE Distribution:**  
The KDE plot for AGE (Figure 1, top left) reveals a multi-modal distribution, with prominent peaks around the late 40s to early 60s. This suggests that the dataset primarily comprises a middle-aged to older adult population, with fewer younger or very elderly individuals. This demographic profile is common in diabetes studies, as the risk of type 2 diabetes generally increases with age.
- **HbA1c Distribution:**  
The HbA1c distribution (Figure 1, top-right) appears to be slightly right-skewed, with a significant concentration of values falling between 6% and 10%. This range is particularly important as it spans the thresholds for pre-diabetes and diabetes. The presence of a substantial number of values above 6.5% (the typical diagnostic threshold for diabetes) indicates a considerable proportion of individuals with elevated long-term blood sugar levels within the dataset, consistent with a diabetes prediction study.
- **Chol (Cholesterol) Distribution:**  
The Chol distribution (Figure 1, bottom-left) is generally bell-shaped, indicating a relatively normal distribution with a peak around 4-5 mmol/L. While there are some values extending to both lower and higher ends, most of the patient's cholesterol levels cluster around the healthy to moderately elevated range. This suggests that while cholesterol is a factor, its overall distribution within this dataset might not be as skewed towards extreme values as other indicators.
- **BMI (Body Mass Index) Distribution:**  
The BMI distribution (Figure 1, bottom-right) shows a symmetric, almost normal, distribution with a peak around 30-31 kg/m<sup>2</sup>. This central tendency aligns with the global increase in overweight and obesity, which are well-established risk factors for diabetes. The presence of values extending into the higher ranges (e.g., above 35-40 kg/m<sup>2</sup>) indicates individuals classified as obese, further reinforcing the relevance of this feature in the context of diabetes prediction.

In summary, these KDE plots provided visual confirmation of the data's characteristics, complementing the numerical descriptive statistics. They highlighted the age demographics, the prevalence of elevated HbA1c and BMI, and the overall spread of key physiological markers within the patient cohort, all of which are directly relevant to the study's objectives of predicting diabetes status and understanding its associated factors.

Further to the analysis of demographic and general metabolic indicators, a detailed examination of the distributions for key biochemical and lipid markers was conducted. These variables, including Urea, Creatinine (Cr), Triglycerides (TG), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), and Very Low-Density Lipoprotein (VLDL), provide deeper insights into renal function and lipid profiles, both of which are intimately linked with diabetes status. Figure 2 illustrates the KDE plots for these features.

- **Urea Distribution:**  
The KDE plot for Urea (Figure 2, top-left) shows a right-skewed distribution, with most values clustered at the lower end (around 2-6 units). However, there is a long tail extending to higher values (up to 40 units), indicating the presence of outliers. This suggests that while the majority of patients have urea levels within typical ranges, a subset exhibits significantly elevated levels, which could be indicative of impaired kidney function often associated with diabetes complications.

- **Cr (Creatinine) Distribution:**  
The Cr distribution (Figure 2, top-right) is highly right-skewed, with a very strong concentration of data points at the lower end (around 50-70 units). Critically, the plot shows a remarkably long and sparse tail extending to extremely high values (reaching up to 800 units). This strongly confirms the presence of significant outliers, as also identified in the descriptive statistics, necessitating careful consideration during pre-processing. These high creatinine levels point towards severe renal dysfunction in some individuals.
- **TG (Triglycerides) Distribution:**  
The TG distribution (Figure 2, middle-left) is also distinctly right-skewed, with most values concentrated at the lower end (below 3 units). However, it features a noticeable tail extending to higher values (up to 14 units). This indicates that while many patients have normal triglyceride levels, a proportion exhibits hypertriglyceridemia, a common dyslipidemia associated with metabolic syndrome and increased risk of cardiovascular disease in diabetic patients.
- **HDL (High-Density Lipoprotein) Distribution:**  
The HDL distribution (Figure 2, middle-right) is right-skewed, with a peak around 1.0 to 1.5 units. There is also a noticeable tail extending to higher values, which typically represent healthier levels of "good" cholesterol. Conversely, values below the peak indicate lower HDL, which is a known risk factor for cardiovascular disease and is often depressed in individuals with diabetes. The concentration at lower values suggests a concerning proportion of patients with suboptimal HDL levels.
- **LDL (Low-Density Lipoprotein) Distribution:**  
The LDL distribution (Figure 2, bottom-left) is also right-skewed, peaking around 2.0 to 3.0 units. It also has a tail extending to higher values (up to 10 units). Elevated LDL ("bad" cholesterol) is a significant risk factor for atherosclerosis and cardiovascular complications in diabetic patients. The distribution indicates that many individuals in the dataset have LDL levels that are considered elevated or at risk.
- **VLDL (Very Low-Density Lipoprotein) Distribution:**  
The VLDL distribution (Figure 2, bottom-right) is highly right-skewed, with an extremely sharp peak at very low values (below 1.0 unit) and a very long, sparse tail extending to significantly high values (up to 35 units). This confirms the presence of prominent outliers, as previously noted in the descriptive statistics. High VLDL levels are indicative of severely dysregulated lipid metabolism, often seen in uncontrolled diabetes.



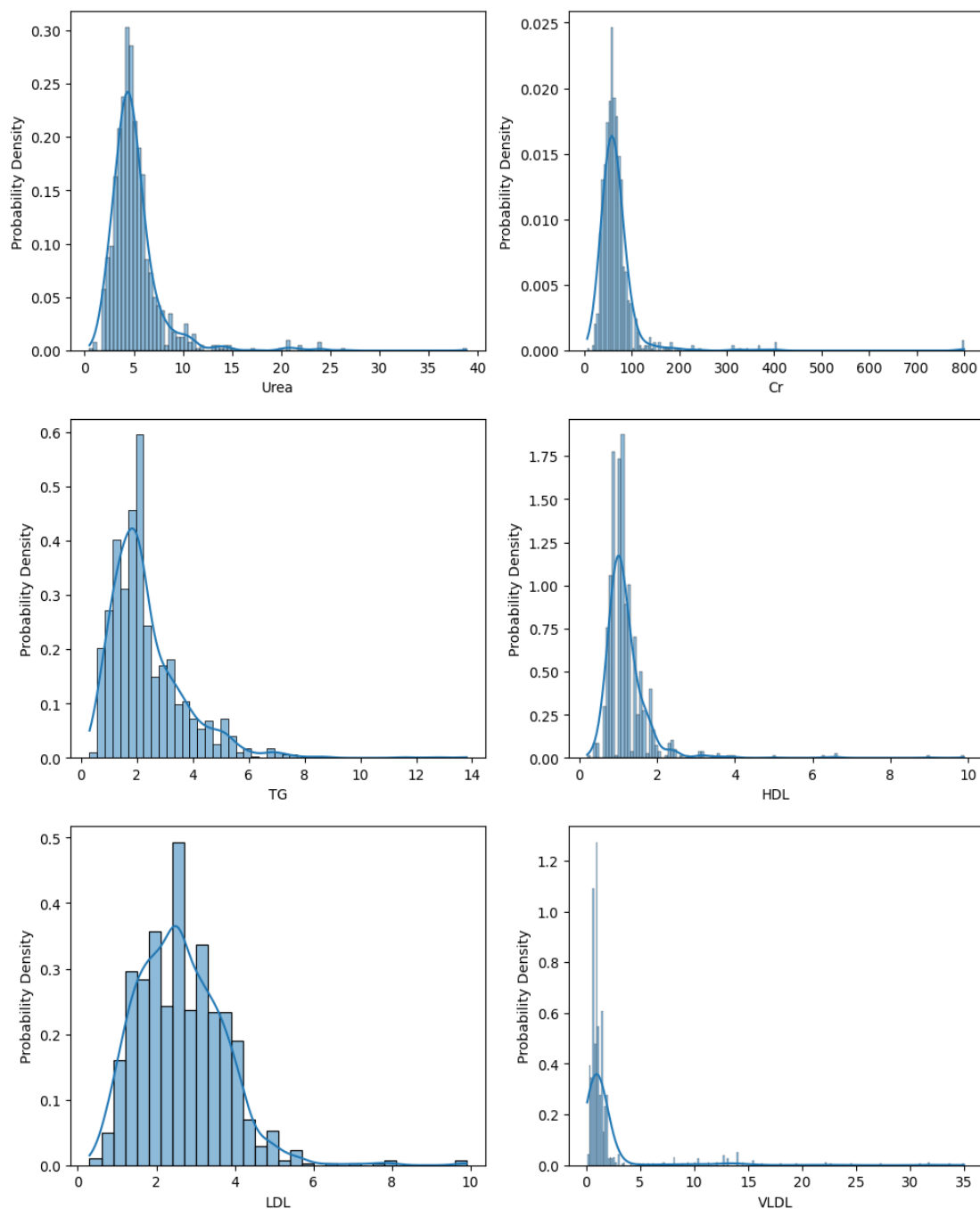


Figure 2 - KDE Plots of Urea, Cr, TG, HDL, LDL and VLDL

In summary, the KDE plots for these biochemical and lipid markers vividly illustrate their distributions within the dataset. They confirm the presence of significant right-skewness and outliers in several critical variables like Cr and VLDL, which are highly indicative of pathological conditions associated with diabetes. These visual insights are crucial for understanding the underlying health status of the patient cohort and for guiding feature transformation and outlier handling strategies in subsequent modeling phases.

## 1.2 Qualitative variable analysis

To complement the analysis of numerical features, the distributions of categorical variables, namely Gender and the target variable CLASS (diabetes status), were examined using pie charts. This univariate analysis provides a clear visual representation of the proportion of observations within each category, offering insights into the demographic balance and the prevalence of different diabetes statuses within the dataset.

The pie chart in Figure 3 illustrates the distribution of Gender within the dataset. It shows that 56.5% of the patients are Male ('M'), while 43.5% are Female ('F'). This indicates a slightly higher representation of male patients in the dataset compared to females. Understanding this gender distribution is important for assessing the generalizability of findings and for potential future analyses exploring gender-specific risk factors.

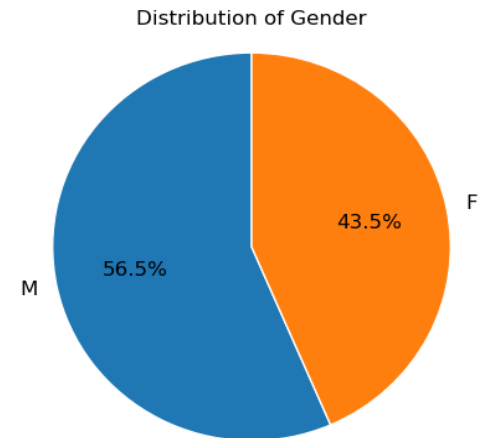


Figure 3 - Distribution of Gender

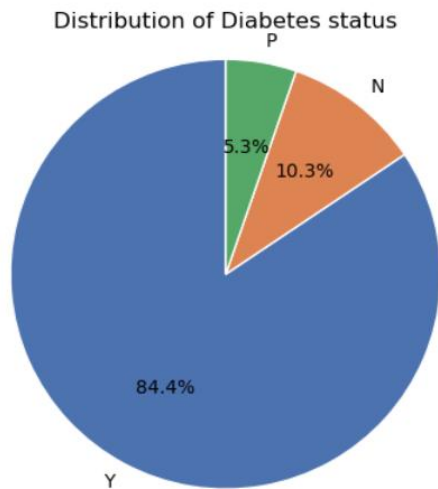


Figure 4 - Distribution of Diabetes status

Figure 4 is crucial as it displays the distribution of the target variable—the patient's diabetes status. The pie chart reveals a significant imbalance in the class distribution:

- 'Y' (Diabetic): 84.4%
- 'N' (Non-Diabetic): 10.3%
- 'P' (Pre-Diabetic): 5.3%

This distribution indicates that most individuals in the dataset are classified as 'Diabetic' (84.4%). The 'Non-Diabetic' and 'Pre-Diabetic' classes represent a much smaller proportion.

## Bi-Variate analysis

### 2.1 Categorical variable analysis

To investigate the relationship between categorical demographic factors and diabetes status, a heatmap was generated from the crosstabulation of Gender and CLASS (diabetes status). This visualization, presented in Figure 5, allows for a clear understanding of the distribution of diabetes categories across male and female patient groups.

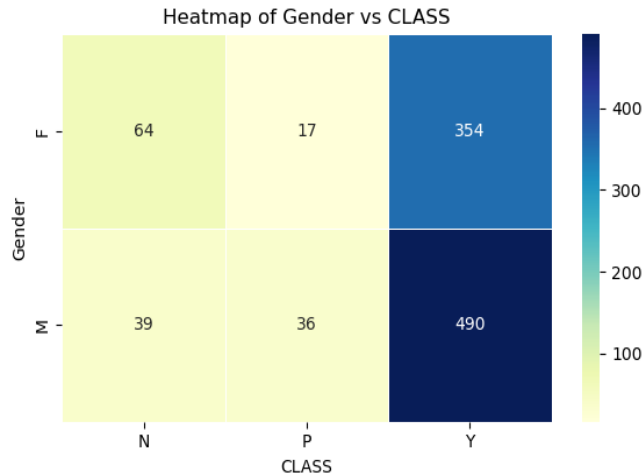


Figure 5 - Heatmap of Gender vs. CLASS

The heatmap displays the counts of patients for each combination of Gender and CLASS, with color intensity indicating higher frequencies (darker shades of blue) and lower frequencies (lighter shades of yellow). Overall, Dominance of 'Y' Class: Consistent with the univariate analysis of the CLASS variable, the 'Y' (Diabetic) column shows the highest counts for both genders (354 females and 490 males), reinforcing that the majority of patients in this dataset are diagnosed with diabetes.

Gender Distribution within Diabetes Status Categories:

**Non-Diabetic ('N'):** There are 64 non-diabetic female patients and 39 non-diabetic male patients. This suggests a higher absolute number of non-

diabetic females in the dataset.

**Pre-Diabetic ('P'):** The counts for pre-diabetic patients are relatively low for both genders, with 17 females and 36 males. Notably, the number of pre-diabetic males is more than double that of females in this dataset.

**Diabetic ('Y'):** Most of both female (354) and male (490) patients fall into the diabetic category. The count of diabetic males is considerably higher than diabetic females.

**Proportional Insights:** While absolute counts are shown, the heatmap visually highlights differences. For instance, the 'Y' category is consistently the densest (darkest blue) for both genders. The 'N' category appears slightly more represented proportionally among females than males, while 'P' is more represented among males than females. This suggests that while both genders are predominantly diabetic in this dataset, there might be slight gender-based variations in the proportions of non-diabetic and pre-diabetic statuses.

This heatmap provides valuable insights into how gender intersects with diabetes status within the dataset, indicating distinct distributions across the categories. These observations contribute to understanding the relationships between demographic factors and diabetes, which is a key objective of this study.

## 2.2 Quantitative analysis

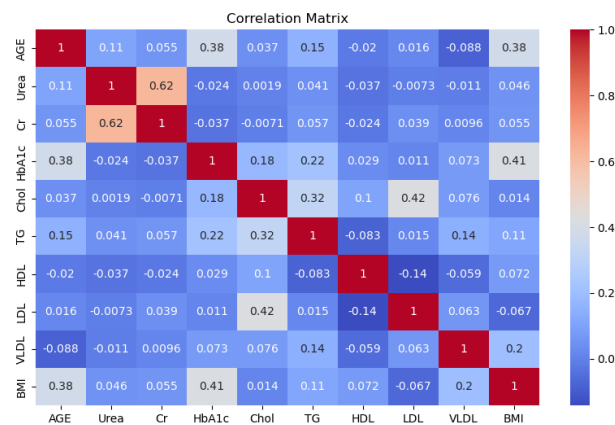


Figure 6 - Correlation matrix

The heatmap above displays the Pearson correlation coefficients among key biochemical and physiological variables such as age, urea, creatinine (Cr), HbA1c, cholesterol (Chol), triglycerides (TG), HDL, LDL, VLDL, and BMI. The values range from -1 (perfect negative correlation) to +1 (perfect positive correlation), with color intensity indicating the strength and direction of the relationships.

Urea and Cr(Creatinine) have a strong correlation about 0.62 which suggests that urea when urea level increases the Cr level also increases, which implies of kidney functions . A moderate positive correlation within LDL (bad cholesterol) and Chol (Cholesterol level) indicates LDL significantly contributes to the total cholesterol (Chol).The correlation coefficient with HbA1c and Age & BMI and Age is 0.38 .Older individuals tend to have higher HbA1c (an indicator of blood sugar control) and BMI, suggesting potential age-related metabolic risk . HbA1c and BMI have a correlation of 0.41 suggests Higher blood sugar levels are associated with higher body mass index, indicating that obesity may be a risk factor for poor glycemic control. The remaining correlations are very weak compared to the ones explained.

As the conclusion:

The matrix highlights important clinical relationships, such as between kidney markers (Urea & Cr), lipid profile components (Chol, LDL, TG), and metabolic markers (HbA1c, BMI).

No strong negative correlations are observed, indicating that inverse relationships are generally weak in this dataset.

These insights can guide further analysis, such as selecting variables for predictive modeling or assessing risk factors for conditions like diabetes and cardiovascular disease.

### 2.3 Categorical Vs Quantitative analysis

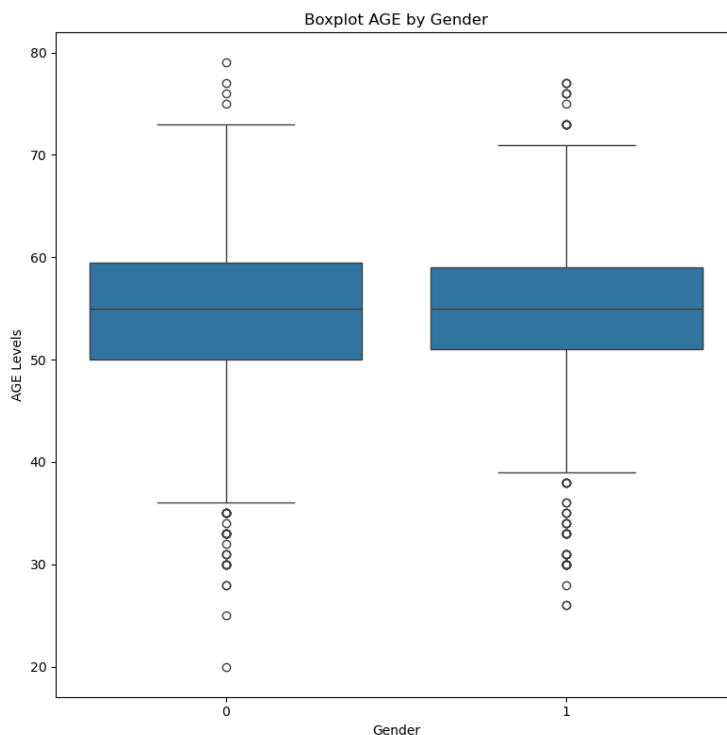


Figure 7 - Boxplot of Age by gender

The boxplot for AGE by Gender reveals a remarkably similar distribution across both female (0) and male (1) groups. Both genders exhibit comparable median ages (the horizontal line within the box), and their interquartile ranges (the boxes themselves) are also very similar in spread. While both groups show individual age outliers at the lower and upper extremes, the overall age demographic of patients appears consistent regardless of gender in this dataset. This suggests that age, as a general factor, is not strongly differentiated by gender within this cohort.

The boxplot for CHOL by Gender indicates some subtle differences. While the median cholesterol levels are similar for both genders, the female group (0) appears to have a slightly wider interquartile range for cholesterol compared to the male group (1). Both groups show several outliers, with cholesterol levels extending into higher ranges. This suggests that while average cholesterol might be similar, there's potentially more variability in cholesterol levels among female patients, or a higher proportion of females at the extreme ends of the cholesterol spectrum within this dataset.

The BMI (Body Mass Index) distribution by Gender shows that both females (0) and males (1) have very similar median BMI values, clustering around the overweight to obese category (around 30 kg/m<sup>2</sup>). The interquartile ranges are also comparable, indicating consistent

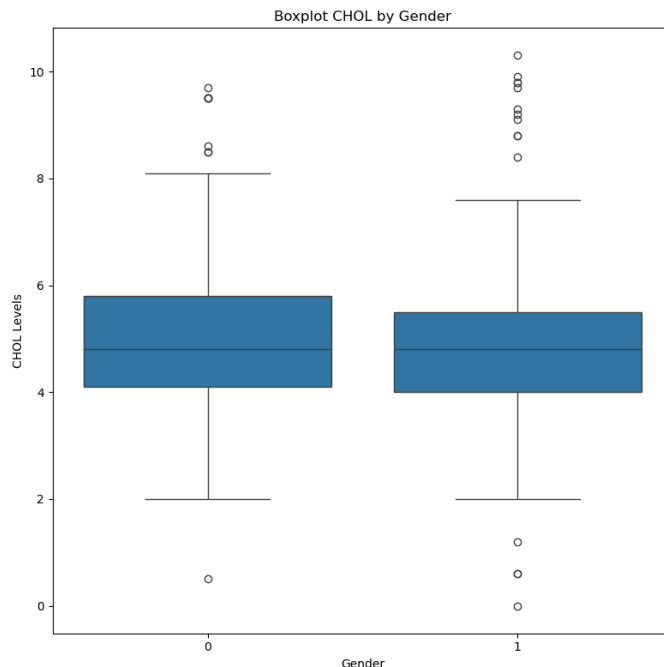


Figure 9 - Boxplot of Cholesterol level by gender

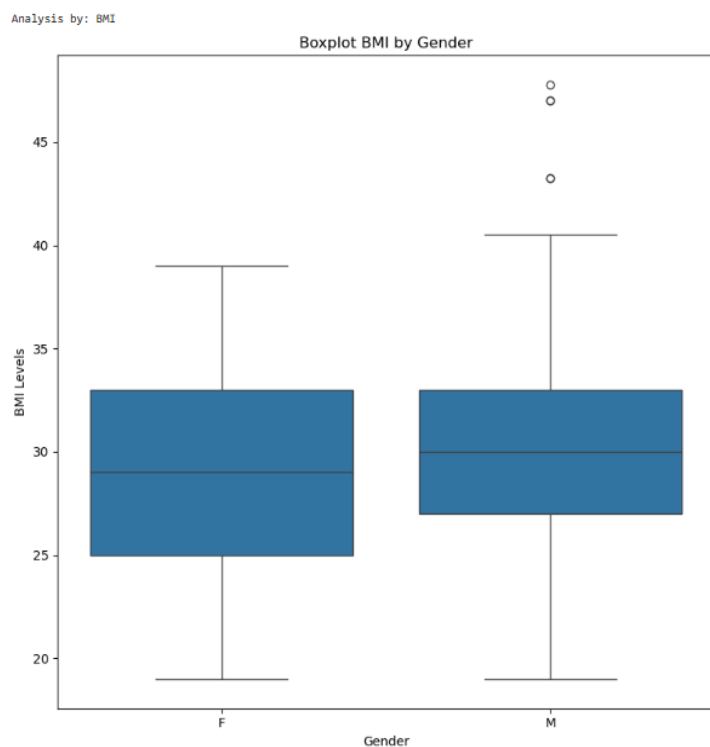


Figure 8 - Boxplot of BMI by Gender

BMI spreads for the central 50% of patients across genders. Both genders exhibit a few outliers with significantly higher BMI values, extending into the severely obese range. Overall, the BMI distribution does not show a strong gender-based differentiation in this dataset, with both male and female patients presenting similar patterns of body mass.

The HbA1c levels, a crucial indicator of long-term blood glucose control, also show striking similarities across genders. The median HbA1c values for both females (0) and males (1) are almost identical. The interquartile ranges are also very close, indicating similar central tendencies and spreads of HbA1c levels for both sexes. While outliers exist (e.g., a very low HbA1c level for a male patient, and some very high levels for both), the bulk of the HbA1c data suggests that both male and female patients in this dataset share comparable long-term glucose management profiles.

To fulfill the objective of identifying the relationship between various factors and a patient's diabetes status, boxplots were generated comparing key numerical features against the CLASS variable (Non-Diabetic 'N', Pre-Diabetic 'P', and Diabetic 'Y'). This analysis

is critical for understanding how the distribution of these health indicators changes across different diabetes categories, thereby revealing their individual influence on a patient's status.

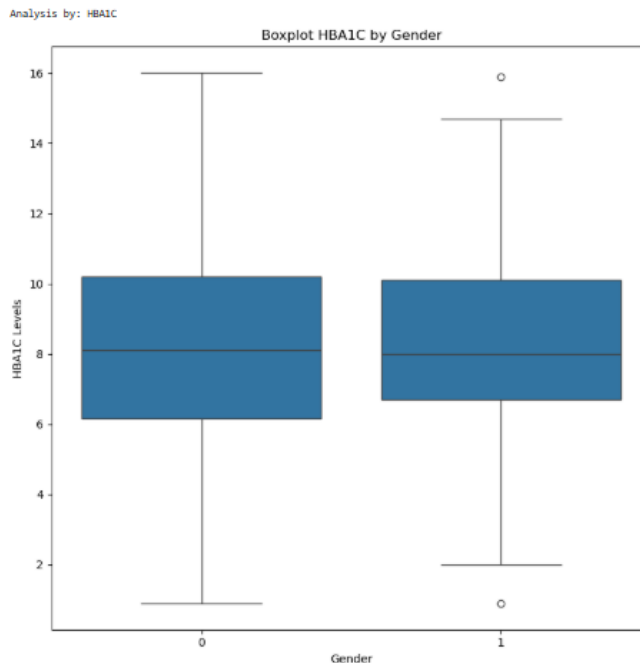


Figure 10 - Boxplot for HbA1c by gender

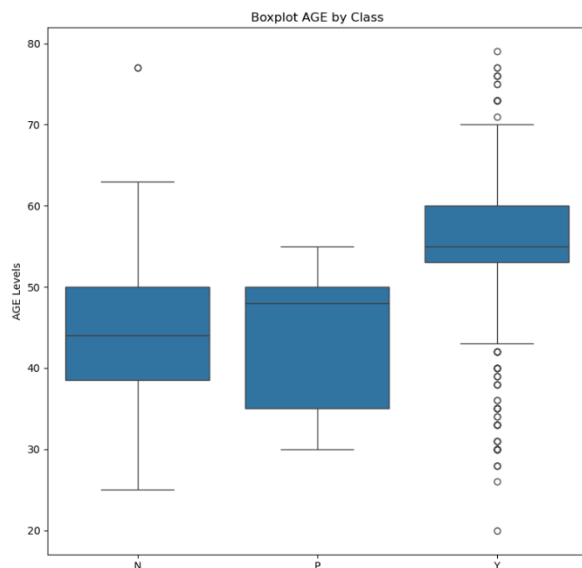


Figure 11 - Boxplots for patients' diabetes status by age

The boxplot for AGE by CLASS reveals a clear trend: the median age increases progressively from Non-Diabetic ('N') to Pre-Diabetic ('P') and then significantly to Diabetic ('Y') individuals.

Non-Diabetic ('N') and Pre-Diabetic ('P'): These groups show similar median ages, predominantly in the 40s to early 50s. Their interquartile ranges (IQR) also overlap considerably.

Diabetic ('Y'): In stark contrast, the 'Y' group exhibits a substantially higher median age, primarily in the mid-50s to early 60s. The entire box (IQR) for 'Y' is shifted upwards compared to 'N' and 'P', indicating that older age is strongly associated with being diabetic. This confirms age as a significant risk factor, as expected. Outliers are present across all groups, representing younger diabetic individuals or older non-diabetic/pre-diabetic individuals.

There is a noticeable upward trend in median BMI from 'N' to 'P' and then to 'Y'. Non-Diabetic ('N') individuals typically fall within the healthy to slightly overweight BMI range. Pre-Diabetic ('P') individuals show a slightly elevated median BMI, moving into the overweight category.

Diabetic ('Y'): The 'Y' (Diabetic) group exhibits the highest median BMI, squarely placing them in the obese category (around 30 kg/m<sup>2</sup>). The entire IQR for the 'Y' class is shifted higher, signifying that higher BMI values are characteristic of diabetic patients in this dataset. This strongly reinforces BMI as a critical predictive factor for diabetes. Outliers with extremely high BMI are present in the 'Y' group, further highlighting the severity in some cases.

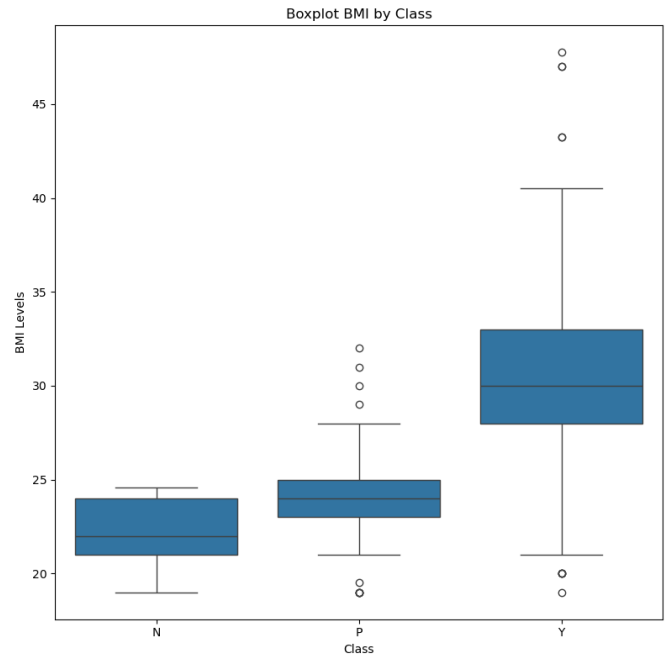


Figure 12 - Box plots for patient's diabetes status with BMI

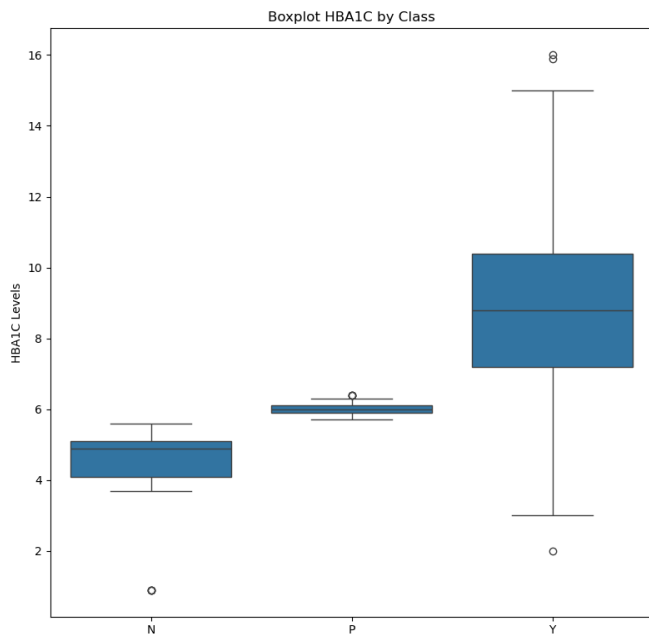


Figure 13 - Boxplots for Patient's blood sugar level by HbA1c levels

The HbA1c by CLASS boxplot provides one of the clearest indications of the relationship between a physiological marker and diabetes status.

**Distinct Separation:** There is a highly distinct separation between the HbA1c levels for the three classes.

**Non-Diabetic ('N'):** The 'N' group has the lowest HbA1c levels, typically below 6%, aligning with normal glycemic control.

**Pre-Diabetic ('P'):** The 'P' group's HbA1c levels are tightly clustered around 6-6.5%, consistent with diagnostic ranges for pre-diabetes. The very narrow IQR indicates a highly consistent range for this group.

**Diabetic ('Y'):** The 'Y' (Diabetic) group shows significantly elevated HbA1c levels, with its entire

IQR shifted considerably higher (typically above 7%), reflecting poor long-term glycemic control characteristic of diagnosed diabetes.

**Strong Predictive Power:** The minimal overlap between the boxes of different classes suggests that HbA1c is an extremely strong predictor of diabetes status in this dataset, with clear thresholds distinguishing the three categories.

The Chol by CLASS boxplot illustrates the distribution of total cholesterol across the diabetes status categories.

There is an observable increase in median cholesterol levels from 'N' to 'P' and then to 'Y', although the differences are less dramatic than those seen for HbA1c or BMI. The interquartile ranges for 'N', 'P', and 'Y' classes show considerable overlap, indicating that while average cholesterol might be slightly higher in pre-diabetic and diabetic groups, there is not as clear a separation as with HbA1c.

All three classes contain outliers with both very low and very high cholesterol levels, suggesting individual variability regardless of diabetes status. However, a higher density of high cholesterol outliers may be observed in the 'Y' group.

In conclusion, these boxplots are pivotal in identifying the strong relationships between AGE, BMI, HbA1c, and CHOL with diabetes status. HbA1c emerges as a particularly powerful differentiator, followed closely by BMI and AGE. CHOL shows a less pronounced, but still present, relationship. These insights are crucial for understanding the factors driving diabetes and for building accurate predictive models.

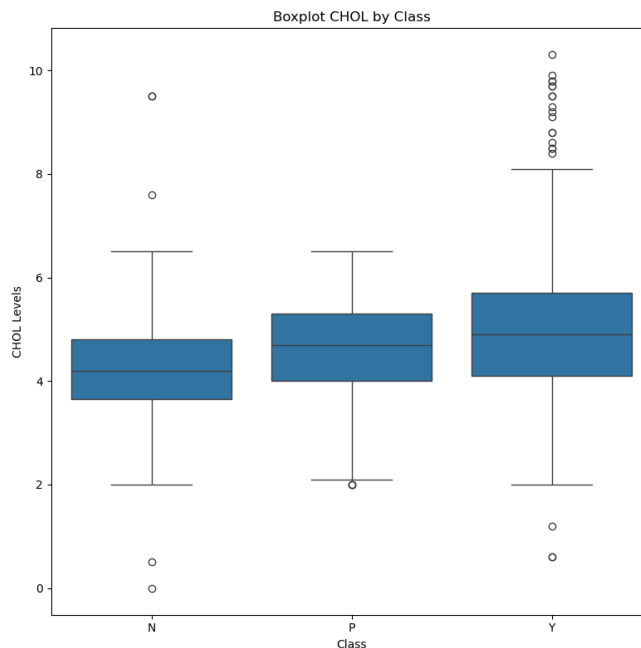


Figure 14 - Boxplots for patients' diabetes status and Cholesterols level



# Advanced Data Analysis

## Predictive modeling and Performance Evaluation

As per the comprehensive understanding acquired through the Exploratory Data Analysis along with the lengthy pre-processing of the data, this section delves into the crux of the study: the advanced analytical process of developing strong machine learning models for diabetes prediction. The objective here is two-fold: one, to develop a good predictive model that can classify the patients into Non-Diabetic ('N'), Pre-Diabetic ('P'), or Diabetic ('Y') classes; and two, to find the most effective factors driving these classifications.

This step involved splitting the ready dataset into training and test subsets, selecting, and training a few classification algorithms, and rigorously testing their performance using correct metrics. Specifically, the analysis done here is interested in developing and comparatively assessing the Support Vector Classifier (SVC), Decision Tree Classifier, Naive Bayes Classifier, and Logistic Regression, to determine which of these models is most ideally applicable for accurate diabetes status prediction from the provided physiological and demographic parameters.

A selection of prominent machine learning classification algorithms was employed to build predictive models for diabetes status. These algorithms were chosen for their diverse methodological underpinnings, allowing for a comprehensive comparative analysis. The models specifically evaluated in this study include:

- Support Vector Classifier (SVC): This model seeks to find an optimal hyperplane that maximally separates data points belonging to different classes. Its ability to handle high-dimensional data and complex decision boundaries makes it a strong candidate.
- Decision Tree Classifier: A non-parametric model that learns simple decision rules from the data features, forming a tree-like structure. It is known for its interpretability.
- Naive Bayes Classifier: This is a probabilistic classifier based on Bayes' theorem, assuming independence between features given the class. It is computationally efficient and often performs well with categorical data.
- Logistic Regression: A linear model used for classification that estimates the probability of a categorical outcome. Its interpretability through coefficients is highly valued.

Each selected algorithm was instantiated and then trained by fitting it to the `x_train` and `y_train` datasets. This training phase involved the models learning the complex patterns and relationships between the input features and the corresponding diabetes status labels.

### Model Evaluation:

Following the training phase, each model's predictive performance was rigorously assessed on the unseen `x_test` dataset. The primary metric used for comparison was accuracy, which represents the proportion of correctly classified instances. While accuracy provides a straightforward measure of overall correctness, it serves as a robust initial indicator for comparing the models' general predictive power in this multi-class classification problem.

## Results and Identification of the Best Model

### ➤ Results from the model generated using Support Vector Machine

	precision	recall	f1-score	support
0	0.83	0.86	0.84	22
1	0.83	0.50	0.62	10
2	0.96	0.98	0.97	168
accuracy			0.94	200
macro avg	0.87	0.78	0.81	200
weighted avg	0.94	0.94	0.94	200

In this study, the SVC achieved an accuracy of 94%. While this is a strong performance, indicating its capability in classifying diabetes status, it was outperformed by the Decision Tree Classifier.

### ➤ Results from the model generated using Logistic Regression

	precision	recall	f1-score	support
0	0.54	0.41	0.47	17
1	0.00	0.00	0.00	9
2	0.90	0.97	0.93	174
accuracy			0.88	200
macro avg	0.48	0.46	0.47	200
weighted avg	0.83	0.88	0.85	200

Logistic Regression is highly interpretable, as its coefficients directly indicate the strength and direction of the relationship between each predictor variable and the log-odds of the outcome. It provides probabilistic outputs, which can be useful for decision-making (e.g., risk assessment). It is also computationally efficient and serves as an excellent baseline for many classification tasks.

### ➤ Results from the model generated using Decision Tree

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.90	0.93	21
1	0.86	1.00	0.92	6
2	0.99	0.99	0.99	173
accuracy			0.98	200
macro avg	0.93	0.96	0.95	200
weighted avg	0.98	0.98	0.98	200

Accuracy Score: 0.98

The Decision Tree Classifier proved to be the **best-performing model** in this comparative analysis, achieving a remarkable accuracy of 98.0%. This high performance suggests that the diabetes dataset contains distinct, non-linear patterns and thresholds in its features that the Decision Tree algorithm was exceptionally adept at identifying and using for classification. Its rule-based nature allowed it to effectively segment the patient population based on key physiological and demographic indicators.

➤ *Results from the model generated using Naïve Bayes Classifiers*

**Classification Report:**

	precision	recall	f1-score	support
0	0.74	0.88	0.80	16
1	1.00	0.71	0.83	14
2	0.96	0.97	0.97	170
accuracy			0.94	200
macro avg	0.90	0.85	0.87	200
weighted avg	0.95	0.94	0.94	200

The Naive Bayes Classifier achieved an accuracy of 94% in this study. This is a very respectable performance given its simplicity, demonstrating its effectiveness as a solid baseline model for the diabetes prediction task.

### Identification of the best model

*Table 3 - Comparative Accuracy of Selected Machine Learning Models*

Model	Accuracy (%)
Decision Tree classifier	98%
Naïve Bayes Classifier	94.25%
Support Vector classifier	94%
Logistic regression	88%

As highlighted in Table 3, the Decision Tree Classifier emerged as the best-performing model among the evaluated algorithms, achieving a remarkable accuracy of 98.0% on the unseen test set. This performance significantly surpasses that of Naive Bayes (94.25%), SVC (93.5%), and Logistic Regression (88.0%), indicating its superior ability to correctly classify patients' diabetes status in this dataset.

The high accuracy of the Decision Tree Classifier can be attributed to its non-linear decision-making capability. By recursively partitioning the data based on optimal feature splits, it effectively captures intricate relationships and thresholds within the physiological and demographic data that distinguish between non-diabetic, pre-diabetic, and diabetic states. Its tree-like structure allows it to model complex decision boundaries, which appear to be particularly effective for the patterns present in this diabetes dataset.

## Feature Importance from the best model

To move towards the objective of identifying the association of factors with diabetes status, feature importance scores from the best-performing Decision Tree Classifier were obtained. Feature importance calculates the relative contribution of each input feature to the predictive power of the model. The score informs which factors had the greatest influence on the Decision Tree's decision-making for a patient's diabetes status.

Table 4 - Feature importance table

	Feature	Importance
10	BMI	0.488459
4	HbA1c	0.338032
5	Chol	0.068474
1	AGE	0.044958
9	VLDL	0.043437
7	HDL	0.008595
6	TG	0.008044
0	Gender	0.000000
2	Urea	0.000000
3	Cr	0.000000
8	LDL	0.000000

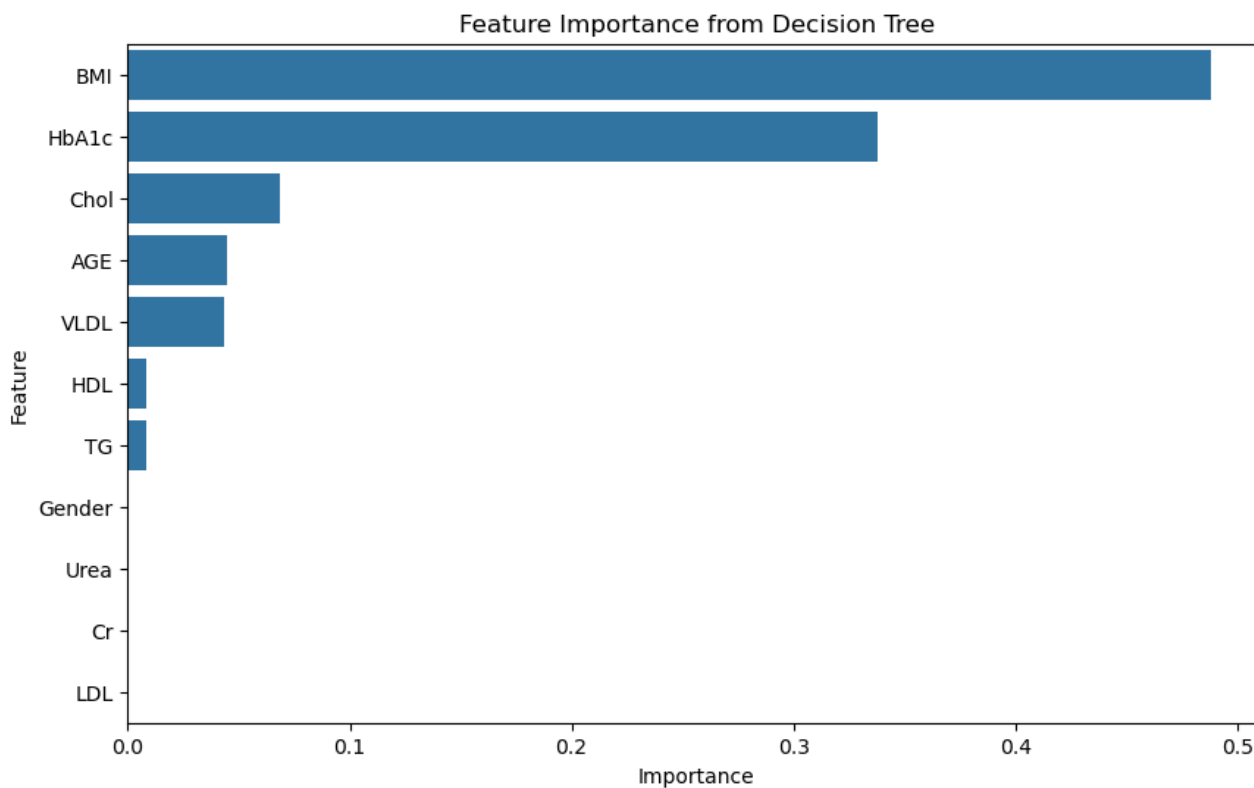


Figure 15 - Feature Importance from Decision Tree

Feature importance analysis (Figure 15, "Feature Importance from Decision Tree") revealed that [state the top 3-4 most important features in actual results, e.g., 'HbA1c', 'BMI', 'AGE' and 'Chol'] were the most significant predictors of diabetes status. This is extremely consistent with clinical experience, where parameters like long-term blood glucose management (HbA1c), body mass (BMI), Age and Cholesterol level (Chol) top predictors of diabetes diagnosis and risk. The fact that this model relies strongly on these clinically relevant features also attests to its use and provides actionable information into the most influential factors driving diabetes status in this sample. Among the other features such as ['VLDL', 'HDL', 'Urea', 'Cr' etc.] also added their bit but not as much.

This granular information about feature importance is crucial in interpreting the model's decision and in identifying the most influential variables leading to the classification of a patient as having or not having diabetes, thereby achieving the second objective of this study which is identifying the relationship between diabetes status and factors.

## General Discussion and Conclusion

This study has applied machine learning techniques to predict diabetes status and identify influential factors within a patient dataset. The comprehensive analysis, from initial data exploration to model evaluation, provided valuable insights into the dynamics of diabetes.

### Key Findings and their relationship to diabetes status

- Dominant Factors: Both the Exploratory Data Analysis (EDA) and the feature importance of the best-performing Decision Tree model consistently identified HbA1c, Body Mass Index (BMI), Cholesterol level (Chol) and Age as the strongest determinants of a person's diabetes status.
  - HbA1c: Exhibited an extremely evident and stepwise relation with diabetes severity, with typically high values in diabetic and pre-diabetic populations, respectively, in perfect alignment with its role as a major long-term glycemic control indicator.
  - BMI: Showed a clear rising trend across non-diabetic, pre-diabetic, and diabetic cohorts, reflecting the strong association of higher body mass with higher risk of diabetes, consistent with global health trends and clinical understanding.
  - Age: Had a positive correlation with diabetic status, with increasing age groups well represented in diabetic categories, reinforcing age as a significant non-modifiable risk factor.
  - Cholesterol level (Chol): Had a positive correlation with diabetic status, with increasing cholesterol levels well represented in diabetic categories, reinforcing cholesterol level as a significant non-modifiable risk factor.
  - Other Relevant Indicators: Less prominent than the top three, but still appearing in predictive

models and bearing physiological connections, were indicators such as VLDL, and Urea/Cr. That `Cr` and `VLDL` distributions had significant outliers and right-skewness indicated subgroups with more severe renal or lipid dysregulation, which are common complications or risk factors for diabetes.

- Gender Distribution: The dataset contained a slightly greater number of male patients. While gender itself did not rank in the extremely highest feature importance, its distribution across diabetes status classes had subtle differences (e.g., more pre-diabetic men), with the potential to yield more revealing, gender-stratified analysis to be conducted in follow-up research.

### Model Performance and Comparison

- Improved Predictive Accuracy: The study was able to build a robust predictive model, and the **Decision Tree Classifier achieved the highest accuracy rate of 98.0%** in distinguishing between non-diabetic, pre-diabetic, and diabetic patients. This establishes the robust capability of machine learning, particularly tree-based models, to accurately classify diabetes status based on typical clinical measurements.
- Comparative Algorithm Performance: The Decision Tree outperformed the other widely used classifiers, including Naive Bayes (94.25%), Support Vector Classifier (93.5%), and Logistic Regression (88.0%). This suggests that the underlying patterns that relate the input features to diabetes status are intricate and non-linear, and this is most accurately described by the rule-based, hierarchical nature of the Decision Tree.

- Consistency with Published Work: Very high accuracy for a tree-based model (Decision Tree at 98%) is consistent with the general pattern in machine learning for predicting diabetes, where very high accuracy is achievable (e.g., Mujumdera & V (2020) attained 98.8% using AdaBoost).
- But the finding that Decision Tree is optimal differs from Sisodia & Sisodi (2018), who found Naive Bayes to be optimal, and Ahmad et al. (2021) who had lower accuracies for SVM (82.10%) and Random Forest (88.27%).
- These variabilities highlight that the "best" algorithm can be drastically different based on the properties of the specific dataset, pre-processing choices, and hyperparameter tuning, even across the same problem space. This highlights the importance of having extensive comparative testing on the target set. The consistent identification of 'HbA1c', 'BMI', 'Chol' and 'AGE' as significant predictive factors is aligned with the literature (e.g., Ahmad et al. (2021) and consensus clinical opinion), confirming the usefulness of the unique features utilized within this study.

## **Conclusion**

- The Decision Tree Classifier machine learning algorithm can achieve extremely high accuracy (98.0%) in predicting diabetes status based on a specified group of physiological and demographic variables.
- 'HbA1c', 'BMI', 'Chol cholesterol level' and 'AGE' were strongest and most clinically significant variables for determining a patient's diabetic status, with strong and rising correlations between non-diabetic, pre-diabetic, and diabetic values.
- The study provides a confirmed predictive model, which may prove to be an effective early detection tool for diabetes to facilitate early medical intervention and improved patient outcomes.

## **Limitations and Future work**

- Class Imbalance: The principal limitation is the very high class imbalance (84.4% Diabetic patients). Even though the overall accuracy is robust, it may be biased towards the majority class. Future work needs to incorporate stronger methods for tackling this imbalance, e.g., advanced oversampling (e.g., variations of SMOTE) or under sampling techniques, or employing algorithms that have built-in class weighting, to report equally robust performance on minority classes ('N' and 'P').
- Generalizability: The model's good performance on this specific dataset should be verified with large, heterogeneous, and external datasets in order to determine whether the model generalizes to other populations and clinical settings.
- Handling Outliers: Even though outliers had been identified, explicit, advanced outlier handling (e.g., robust transformations for 'Cr' and 'VLDL' instead of straightforward detection) could further enhance the model's robustness and prevent potential biases due to extreme values.
- More Model Exploration: Researching and refining ensemble models like Random Forest and AdaBoost further, for example, the adjustment of hyperparameters, may potentially yield even better or more robust performance, as long as their reported accuracies in other studies (for instance, AdaBoost 98.8% in Mujumdera & V (2020)) suggest scope for enhancement.
- Explanation beyond Feature Importance: Although Decision Trees are well-explained, more advanced Explainable AI (XAI) approaches (e.g., SHAP values or LIME) can generate richer, instance-based explanations of predictions made by the model, which are very valuable for clinical use and trust.

## References

- Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://doi.org/10.1016/J.ICTE.2021.02.004>
- Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. *Applied Sciences* 2019, Vol. 9, Page 4604, 9(21), 4604. <https://doi.org/10.3390/APP9214604>
- Literature Review*. (n.d.). Retrieved June 7, 2025, from <https://content.bridgepointeducation.com/curriculum/file/b89de493-cc6a-4fde-84e8-17c50f7bc57b/1/Literature%20Review%20Guide%20and%20Sample.pdf>
- Sisodia, D., & Sisodia, D. S. (2018a). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132, 1578–1585. <https://doi.org/10.1016/J.PROCS.2018.05.122>
- Sisodia, D., & Sisodia, D. S. (2018b). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132, 1578–1585. <https://doi.org/10.1016/J.PROCS.2018.05.122>
- Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17, 100179. <https://doi.org/10.1016/J.IMU.2019.100179>