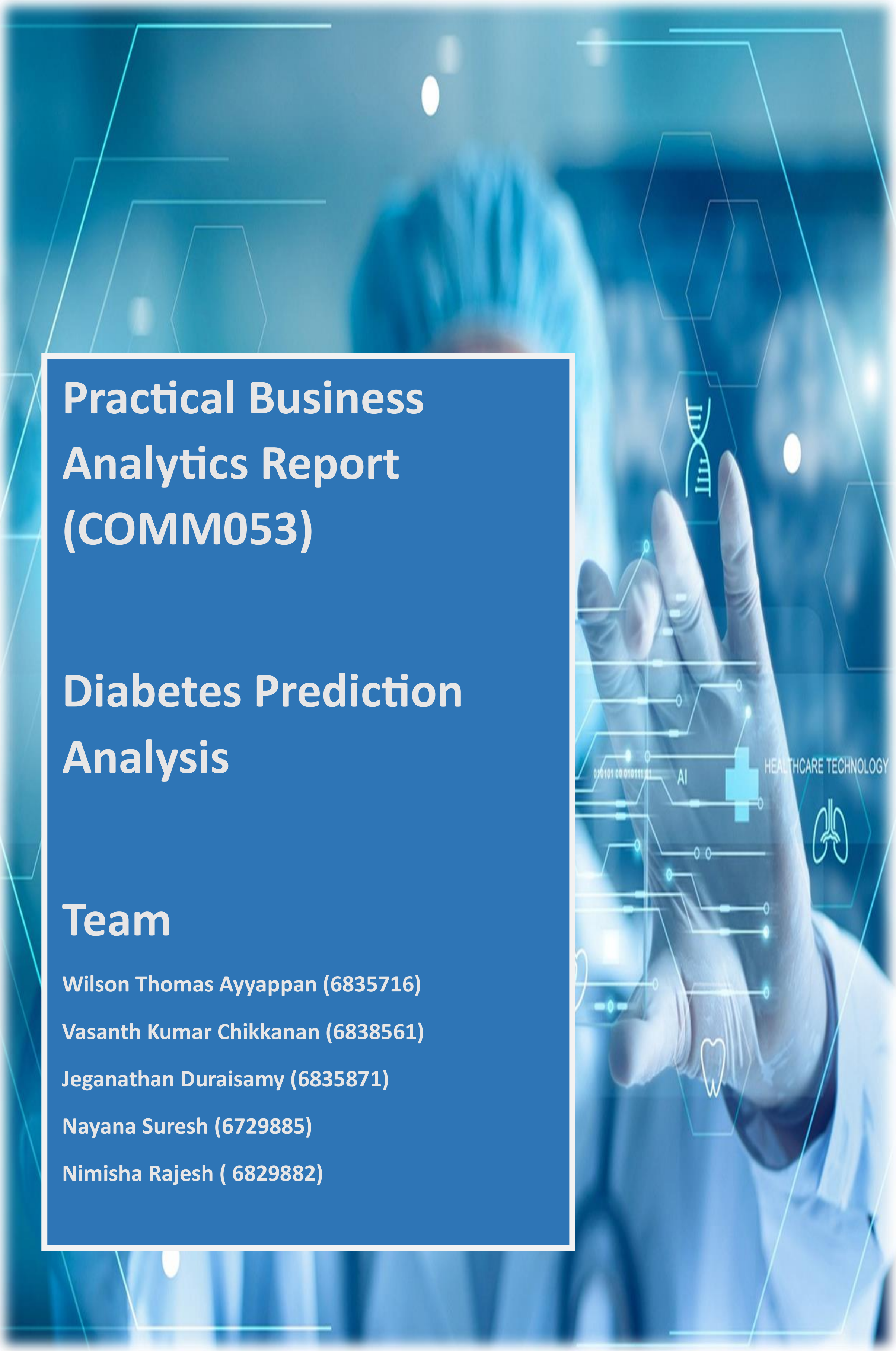


Practical Business Analytics Report (COMM053)

Diabetes Prediction Analysis

Team

Wilson Thomas Ayyappan (6835716)

Vasanth Kumar Chikkanan (6838561)

Jeganathan Duraisamy (6835871)

Nayana Suresh (6729885)

Nimisha Rajesh (6829882)

TABLE OF CONTENTS

1. Problem Definition:

1.1 Problem Understanding

- The problem definition for the diabetes prediction dataset revolves around developing a predictive model using machine learning techniques to assess the likelihood of an individual developing diabetes.
- The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative).
- The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level that are mentioned above.
- This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information.
- This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans.

1.2 Business Outcome:

- Diabetes predictive analysis can benefit business in a number of ways. It can improve patient outcomes, lower treatment costs, and provide early intervention to improve healthcare management. It might also contribute to the improvement of preventative medicine and the creation of individualized healthcare solutions.
- Insurance firms can provide more competitive options by providing more individualized plans and a better understanding of individual risks.
- Diabetes prediction analysis encourages cooperation between scientists, medical practitioners, and IT specialists, resulting in a more integrated and creative approach to diagnosing, treating, and preventing diabetes.

2.Dataset Description

The goal of the Diabetes Prediction dataset is to make it easier to create and validate machine learning models that can forecast a person's risk of developing diabetes or existence of the disease. In healthcare contexts, this type of dataset is very helpful for early detection, individualized treatment plans, and preventative medicine.

1) **Primary Goal**: The main objective of this dataset is to use certain variables from the dataset to determine a person's likelihood of developing diabetes.
2) **Target Variable(diabetes):** The binary result of whether a person has diabetes is often represented as 1 for those who do not have diabetes and 0 for those who have the diabetes.
3) **Features**: A variety of variables that are known to be connected to the risk of diabetes are usually included in the dataset. These could consist of:
    a. Demographic data: Age, Gender.
    b. Medical history: Presence of conditions like hypertension or heart disease.
    c. Lifestyle factors: Smoking history, physical activity levels.
    d. Physiological measurements: Body Mass Index (BMI), blood pressure.
    e. Clinical tests: Blood glucose levels, HbA1c levels, insulin levels.
4) **Source:**
    a. The dataset was sourced from the Kaggle website.
    b. Link to Kaggle Website: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data
5) **Feature Explanation**:
    a. Age: An important risk factor for diabetes is age. As one ages, the risk rises because of altered body composition, decreased physical activity, and other age-related variables. Due to the increased risk associated with aging, age is a key predictor.
    b. Gender: Diabetes can affect people differently depending on their gender. Different reactions to insulin, lifestyle choices, and biological variations could all be contributing factors. Different

biological and lifestyle variables can contribute to different diabetes prevalence and risk factors in different genders.

c. Hypertension (High Blood Pressure): High blood pressure is frequently linked to a higher chance of diabetes. It may be a sign of underlying cardiovascular problems, which are frequently connected to diabetes.
d. Heart Disease: Cardiovascular issues more generally linked to an elevated risk of diabetes may be indicated by a history of heart disease. People who have a history of cardiovascular issues or who have heart disease may be more susceptible to diabetes. Diabetes and heart disease share common risk factors like obesity and high blood pressure.
e. History of Smoking: Smoking can cause insulin resistance, which is a major contributing factor to the development of diabetes, as well as alter blood circulation. Nicotine is associated with cardiovascular illnesses, which are linked to diabetes, and it can cause insulin resistance.
f. BMI (Body Mass Index): BMI is a weight-and-height-based indicator of body fat. Due to obesity’s role as a primary risk factor, a higher BMI is frequently associated with a higher risk of diabetes.
g. HbA1c Level: Prolonged periods of high blood sugar are indicated by elevated HbA1c readings, which are the characteristic feature of diabetes. Diabetes is diagnosed and tracked using higher haemoglobin A1c values, which also signify worse blood glucose control.
h. Blood Glucose Level: Increased blood glucose levels may be a sign of diabetes. Blood Glucose Level can be determined in a number of methods, including glucose tolerance tests and fasting glucose levels. Having high blood glucose is the main sign of diabetes.
i. Diabetes (Target Variable): The Diabetes variable is a critical component, as it represents the target variable that the predictive model is trying to forecast. Based on the other predictors, the model attempts to forecast this variable, which is often binary (yes/no or 0/1).

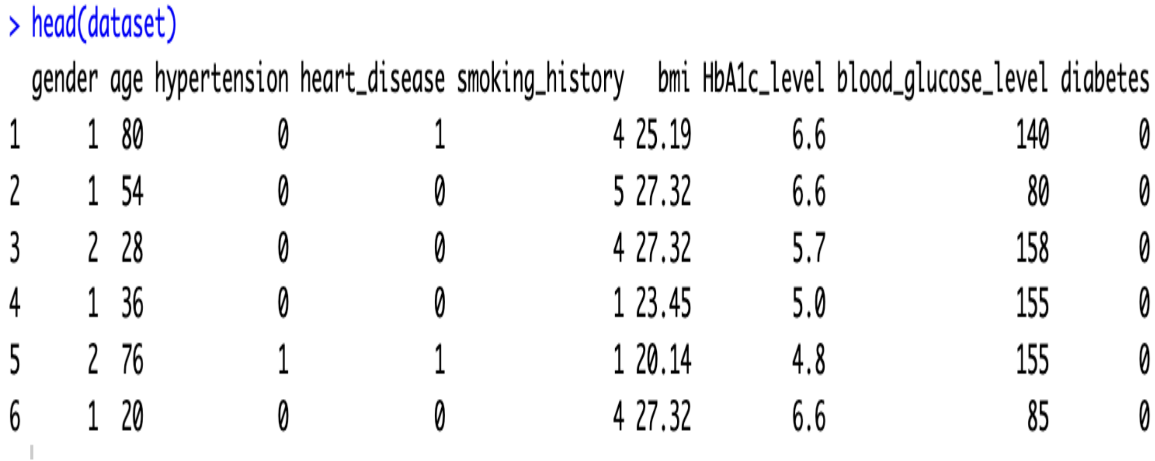
3.Data Understanding:

3.1 Import data

To access and read the data from external file outside the R environment. This step is where we get the data to build models on.

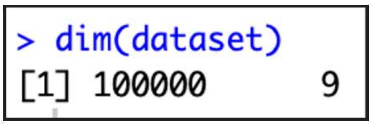
3.2 Data Information:

1) **Head:** head() function - which returns the first few rows of the dataset



```
> head(dataset)
  gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1      1  80            0             1               4 25.19         6.6                 140        0
2      1  54            0             0               5 27.32         6.6                  80        0
3      2  28            0             0               4 27.32         5.7                 158        0
4      1  36            0             0               1 23.45         5.0                 155        0
5      2  76            1             1               1 20.14         4.8                 155        0
6      1  20            0             0               4 27.32         6.6                  85        0
```

2) **Dimension of the Dataset:** This dataset comprises 100,000 rows and 9 columns, according to the output [1] 100000 9.

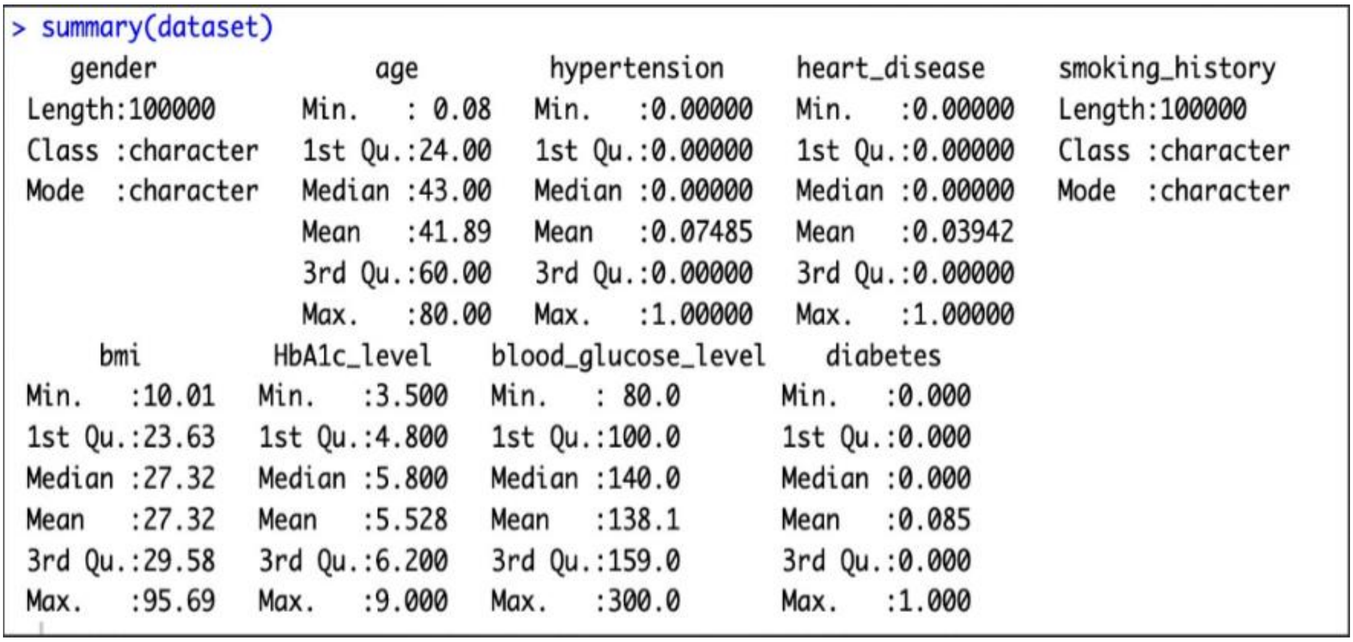


```
> dim(dataset)
[1] 100000      9
```

3) **Summary of the Dataset:**

- Displaying the output of summary().
- Statistical overview of every column in the dataset is given by this function. Blood glucose level, gender, age, blood pressure, heart disease, smoking history, BMI (body mass index), HbA1c level (a blood sugar control measure), and diabetes status are among the columns in the collection that seem to be related to medical or health data.

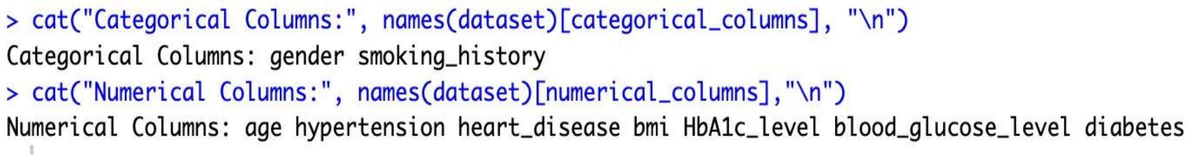
- For each variable, the summary includes: Minimum value (Min.),1st Quartile (25th percentile),Median (50th percentile),Mean (average),3rd Quartile (75th percentile),Maximum value (Max.)



```
> summary(dataset)
    gender                age         hypertension      heart_disease     smoking_history   
 Length:100000      Min.   : 0.08   Min.   :0.00000   Min.   :0.00000   Length:100000     
 Class :character   1st Qu.:24.00   1st Qu.:0.00000   1st Qu.:0.00000   Class :character  
 Mode  :character   Median :43.00   Median :0.00000   Median :0.00000   Mode  :character  
                    Mean   :41.89   Mean   :0.07485   Mean   :0.03942                     
                    3rd Qu.:60.00   3rd Qu.:0.00000   3rd Qu.:0.00000                     
                    Max.   :80.00   Max.   :1.00000   Max.   :1.00000                     
      bmi         HbA1c_level    blood_glucose_level    diabetes    
 Min.   :10.01   Min.   :3.500   Min.   : 80.0       Min.   :0.000  
 1st Qu.:23.63   1st Qu.:4.800   1st Qu.:100.0       1st Qu.:0.000  
 Median :27.32   Median :5.800   Median :140.0       Median :0.000  
 Mean   :27.32   Mean   :5.528   Mean   :138.1       Mean   :0.085  
 3rd Qu.:29.58   3rd Qu.:6.200   3rd Qu.:159.0       3rd Qu.:0.000  
 Max.   :95.69   Max.   :9.000   Max.   :300.0       Max.   :1.000  
```

4) Displaying Categorial and Numerical Columns:

- Categorical Columns: gender and smoking history are the categorical variables in the dataset.
- Numerical Columns: age, hypertension, heart disease, bmi, HbA1c_level, blood glucose level and diabetes are the numerical variables in the dataset.



```
> cat("Categorical Columns:", names(dataset)[categorical_columns], "\n")
Categorical Columns: gender smoking_history
> cat("Numerical Columns:", names(dataset)[numerical_columns],"\n")
Numerical Columns: age hypertension heart_disease bmi HbA1c_level blood_glucose_level diabetes
```

5) Distribution of all variables:

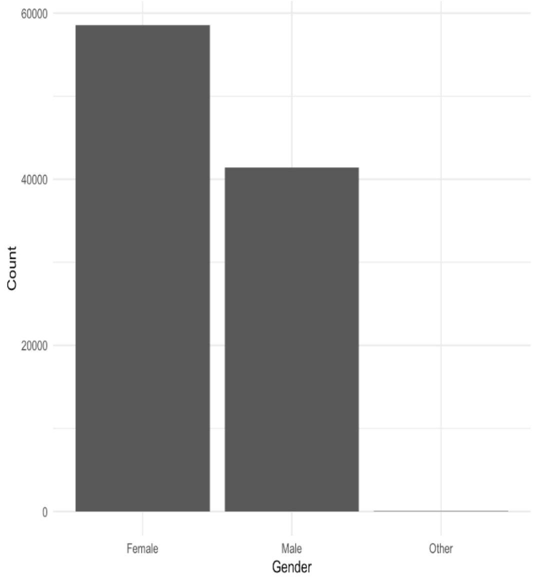
Distribution of all the variables in the dataset is analysed to understand its distribution. Two of those distribution insights have been explained below:

i. Gender:

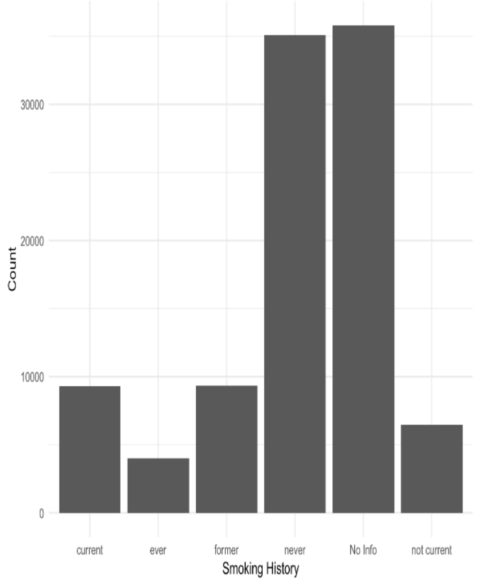
  - Nearly 60,000 people are in the "Female" group, which has the greatest figure.
  - Almost 40,000 people are in the "Male" group, which has the second-highest total.
  - In comparison to the other two categories, the "Other" category has the lowest count.

ii. Smoking History:

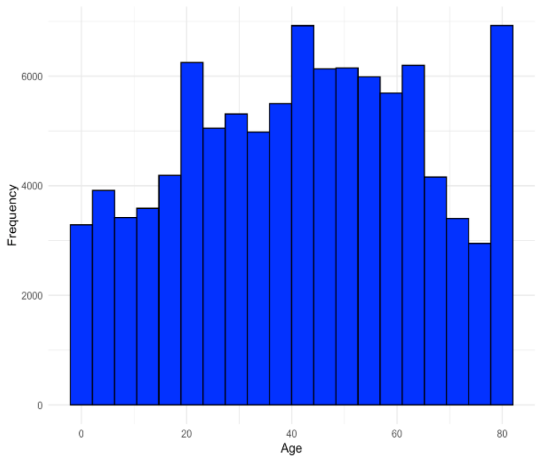
  - Current: It shows that people who don’t smoke at the moment, based on the size of the bar.
  - Ever: This category also has a relatively small bar, indicating a small number of individuals who have smoked at some point.
  - Former: Former smokers had a much higher bar, indicating a greater proportion of people who smoked in the past but stopped.
  - Never: People who have never smoked are represented by the highest bar in the sample by a substantial margin, making them the largest category.
  - No Info: A significant portion of the population does not have any information available about their smoking history.
  - Not current: The little bar in this category, which resembles the 'current' and 'ever' categories, indicates a lower number of people.



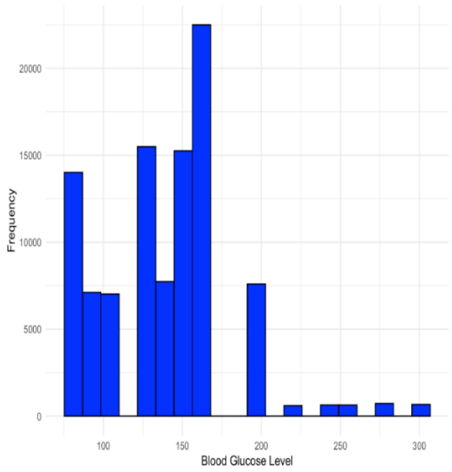
Gender



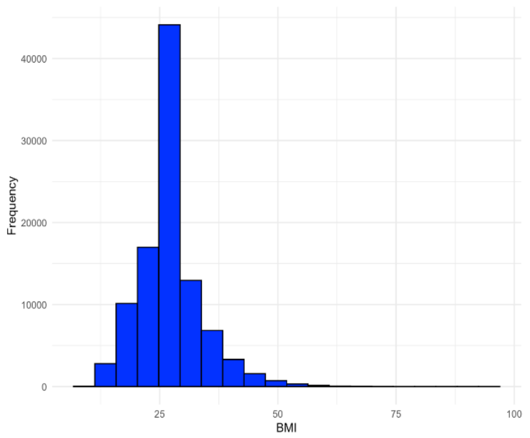
Smoking History



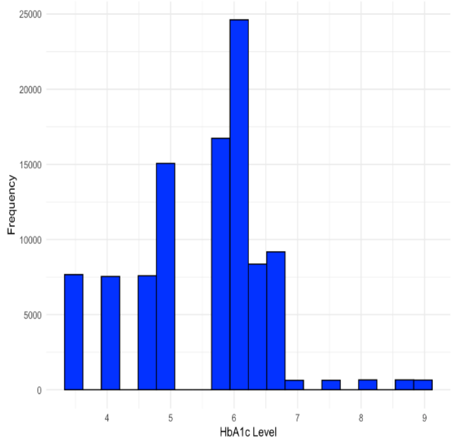
Age



Blood Glucose Level



BMI

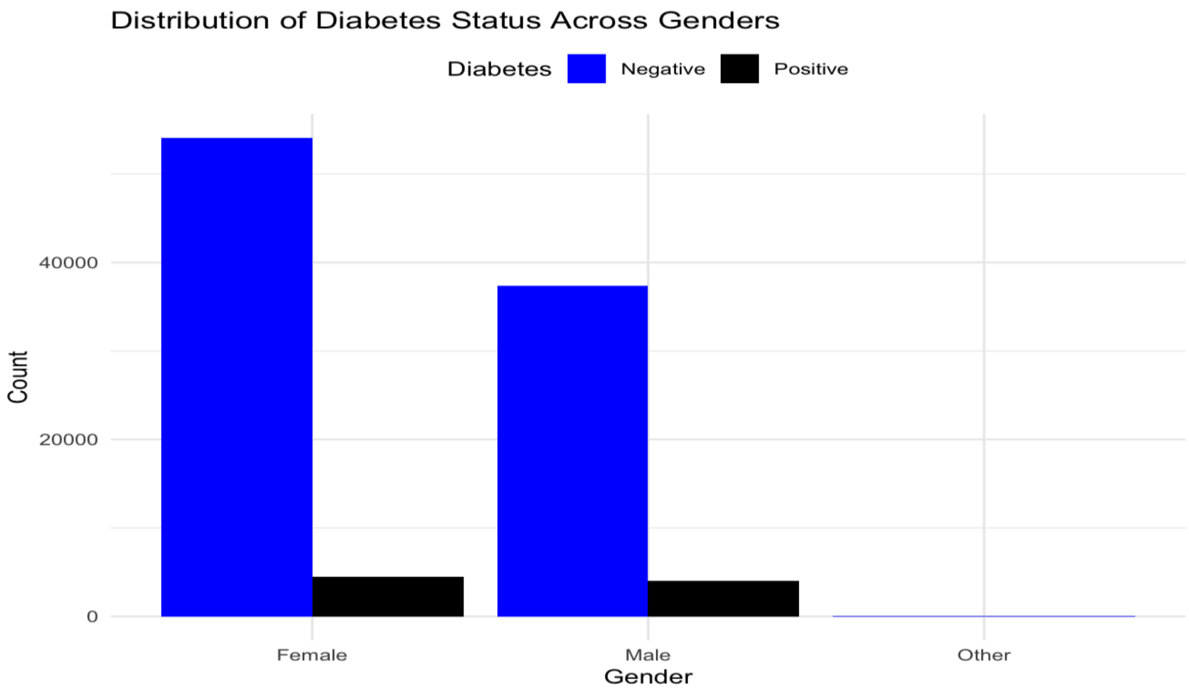


HbA1c

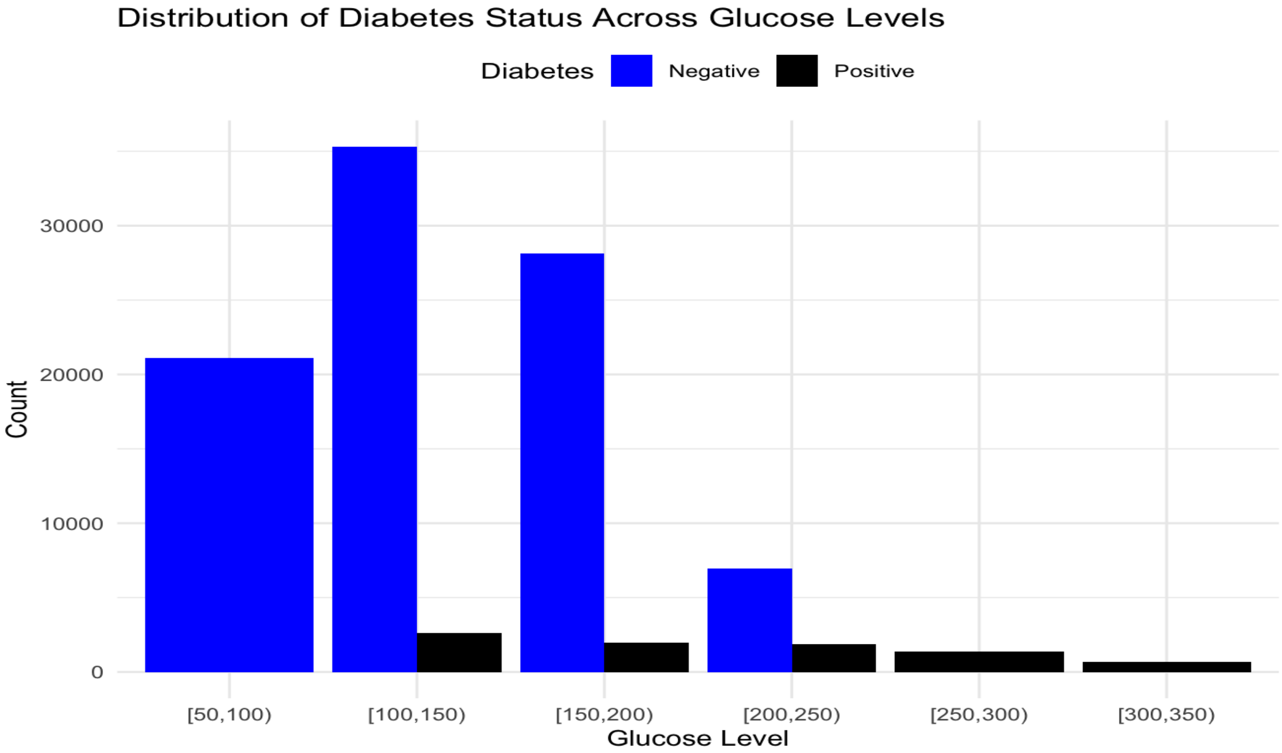
3.3 Exploratory Data Analysis:

1. **Relationship of target variable with other variables:**

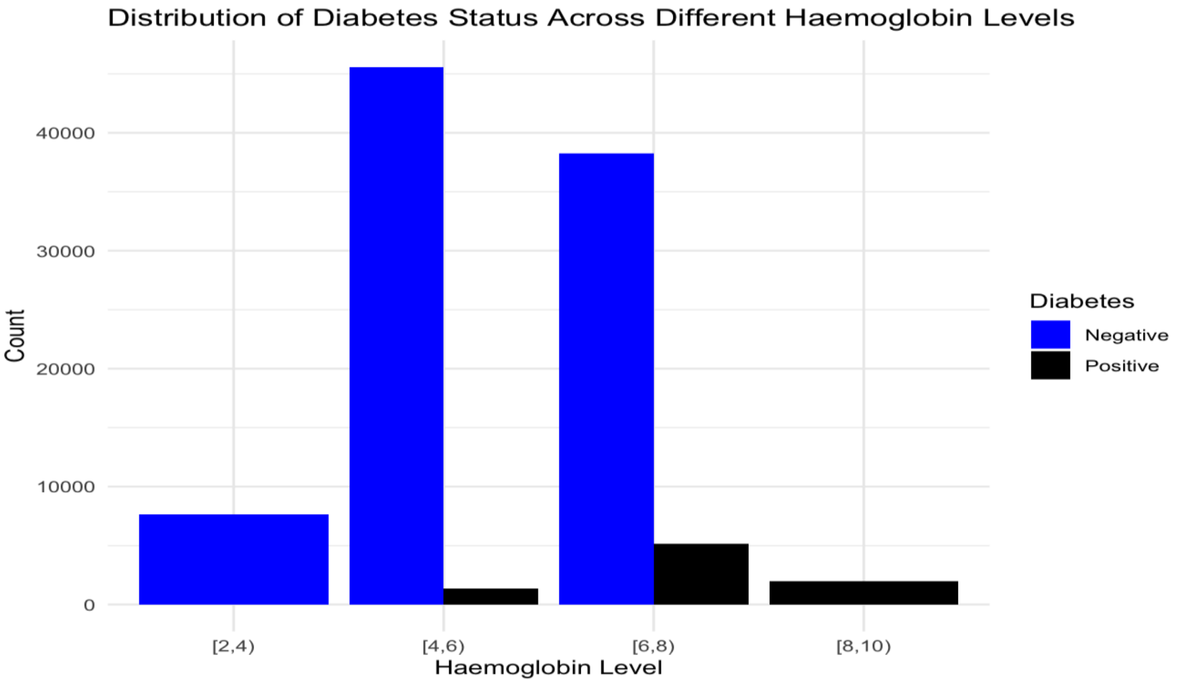
- This graph depicts that count of female who are not affected by diabetes is greater than the count of men who are not affected. Hence, we can say that women are less prone to get diabetes.



- From the below gram we can infer that people who have glucose level 100 – 300 have high chances of getting diabetes when compared to people in the range 300-350.Hence we can say more the glucose level, higher chance of getting diabetes.

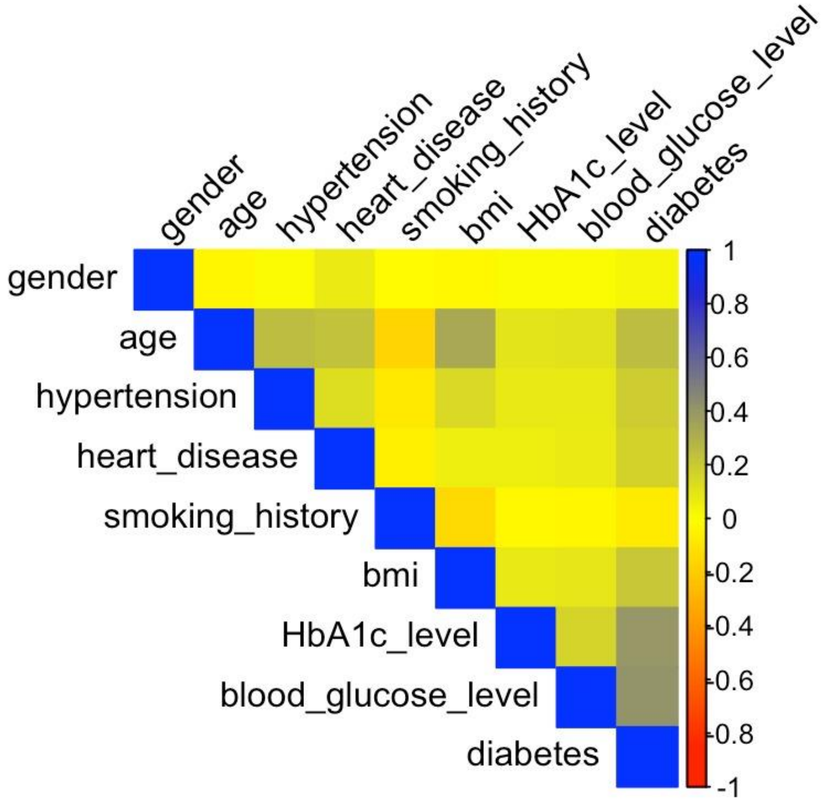


- From the below graph we can infer that people who have the haemoglobin level of (6,8) have high chances of getting diabetes when compared to people with the range of haemoglobin level (4,6). Hence, we can say more the haemoglobin level, higher chance of getting diabetes.



2. **Correlation Matrix:**

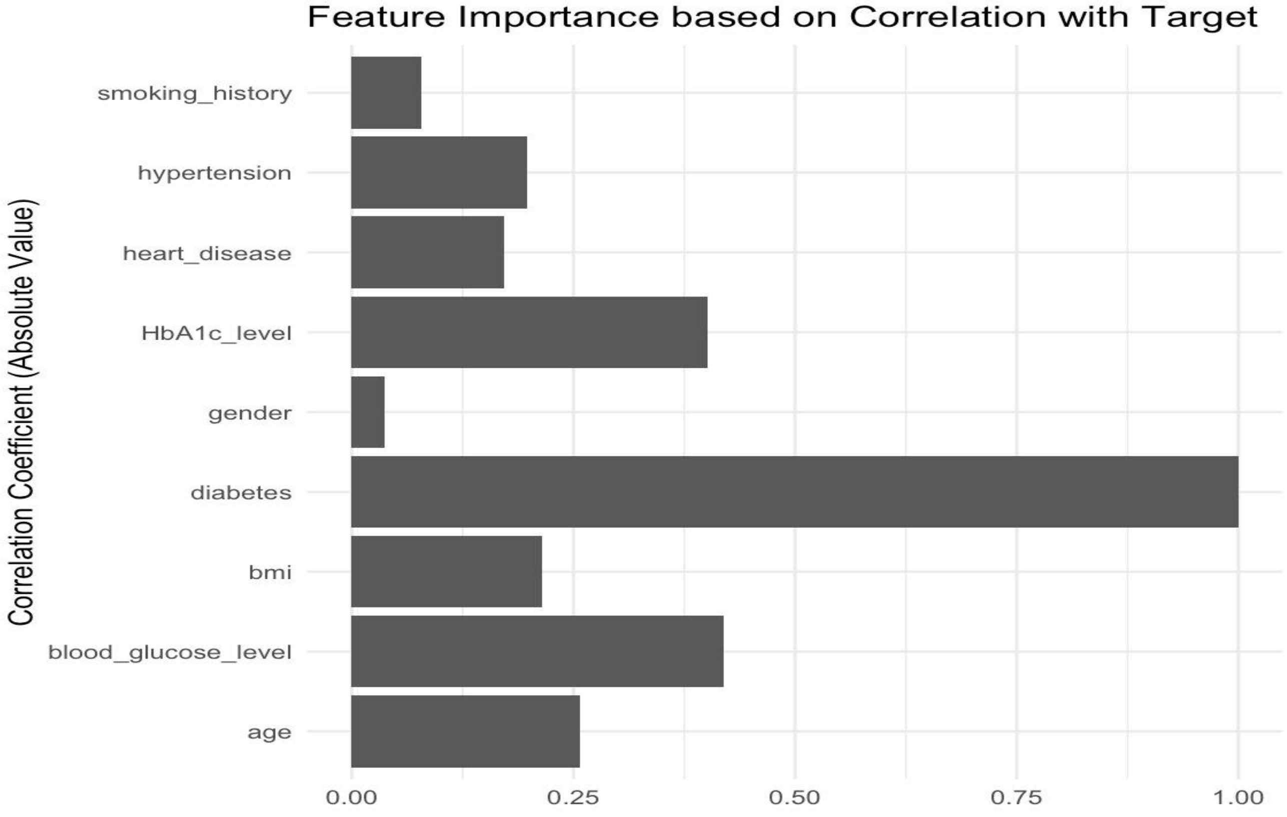
- A correlation matrix is a table showing how closely related two variables are, working best with variables that have a linear connection. It uses rows and columns to represent different variables, and a scatter plot can visually show how well these variables relate to each other.
- Value 1 in this correlation matrix indicates a positive relationship, where one variable increases, the other one also increases.
- Value -1 in this correlation matrix indicates negative relationship, where one variable increases, the other one decreases.
- Value 0 in this correlation matrix indicates no relationship between the two variables.
- In this case, we have less correlation between variables which makes variables independent of each other.



3. **Feature Importance of Correlation:**

- The chart evaluates the strength of a linear relationship in both directions between each feature and the target variable.
- The feature with the longest bar, "Age," has the highest absolute correlation with the target variable, indicating that it is the most significant predictor of all the features provided.
- “Blood glucose level" and "bmi" are the next most significant features, listed in descending order of importance.
- Based on their correlation with the target variable, features such as 'gender' and 'HbA1c_level' are of moderate significance.
- The variables in this figure with the lowest absolute correlation with the target variable are "smoking history," "hypertension," "heart disease,"

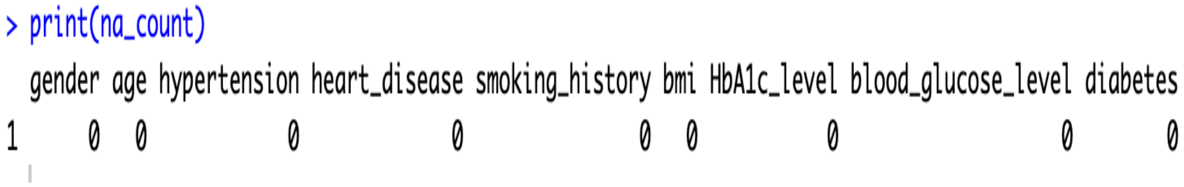
and "diabetes," suggesting that these are less significant predictors than the other variables.



4. Data Preprocessing:

4.1 Missing Values

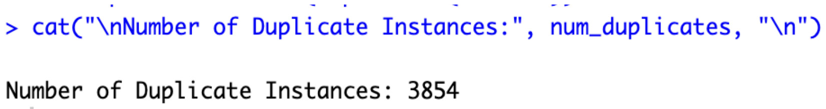
In our analysis of the diabetes prediction dataset, we found that it contains no missing values, indicating a high level of data completeness, making it well-suited for further analysing.



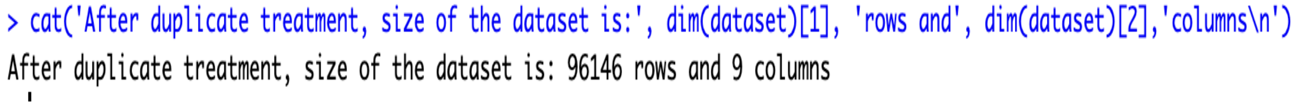
```
> print(na_count)
  gender age hypertension heart_disease smoking_history bmi HbA1c_level blood_glucose_level diabetes
1      0   0            0             0               0   0           0                   0        0
```

4.2 Duplicate Values

- Checking for and removing duplicate values is essential ensure the accuracy and reliability of data analysis and modelling.
- we identified and addressed duplicate entries. Upon detecting these duplicates, we carefully removed them to ensure the uniqueness and accuracy of our data. The dataset, now free of duplicates, is well-prepared for analysis.
- We found and deleted 3,854 repeated entries in our dataset. Now we have 96,146 rows and 9 columns left.



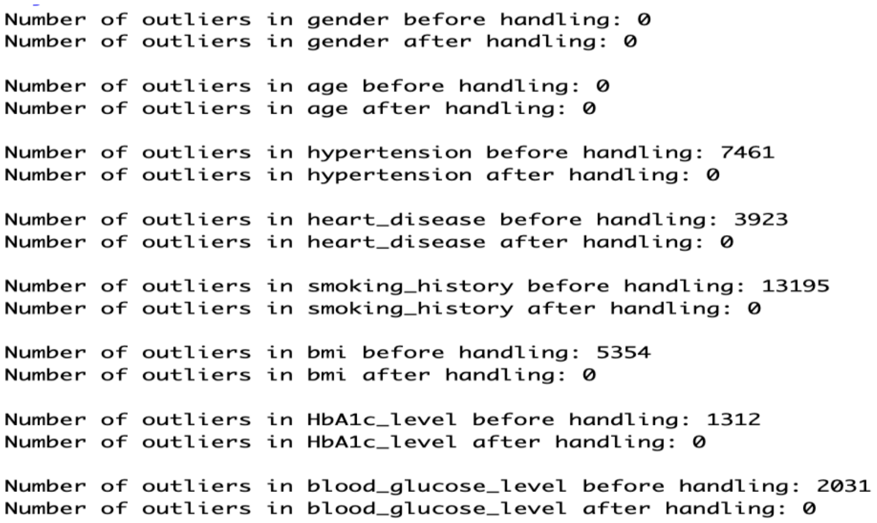
```
> cat("\nNumber of Duplicate Instances:", num_duplicates, "\n")

Number of Duplicate Instances: 3854
```



```
> cat('After duplicate treatment, size of the dataset is:', dim(dataset)[1], 'rows and', dim(dataset)[2],'columns\n')
After duplicate treatment, size of the dataset is: 96146 rows and 9 columns
```

4.3 Outliers

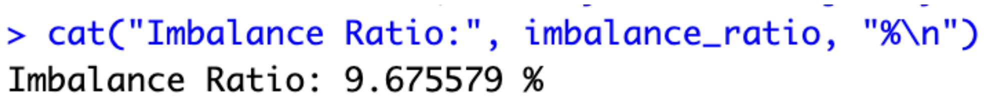
- Outlier checks are necessary to spot and correct unusual data points that can distort our analysis results.
- This step helps maintain the overall quality and accuracy of our dataset for reliable conclusions.
- After analysing we found the outliers in hypertension (7461), heart_disease (3923), smoking_history (13195), bmi (5354) and blood_glucose_level (2031), we handled the outliers and below is the result after handling the data.



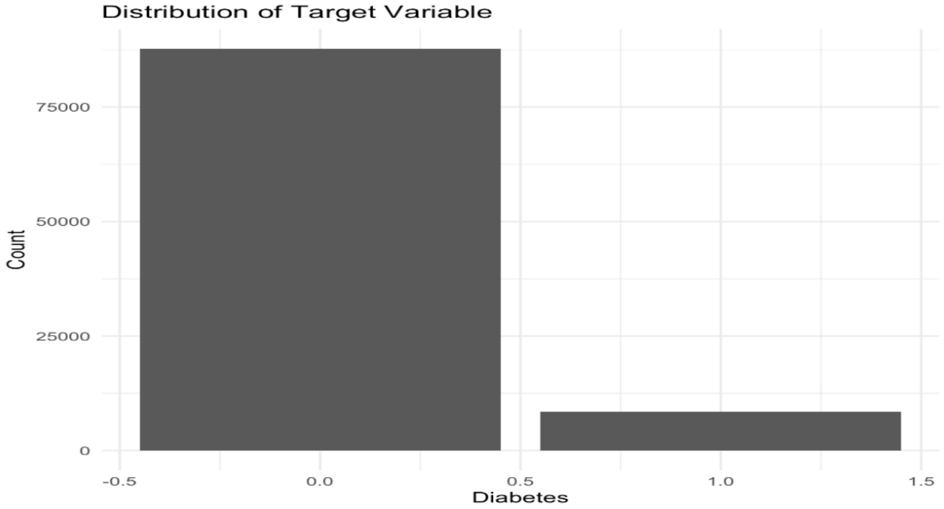
```
Number of outliers in gender before handling: 0
Number of outliers in gender after handling: 0

Number of outliers in age before handling: 0
Number of outliers in age after handling: 0

Number of outliers in hypertension before handling: 7461
Number of outliers in hypertension after handling: 0

Number of outliers in heart_disease before handling: 3923
Number of outliers in heart_disease after handling: 0

Number of outliers in smoking_history before handling: 13195
Number of outliers in smoking_history after handling: 0

Number of outliers in bmi before handling: 5354
Number of outliers in bmi after handling: 0

Number of outliers in HbA1c_level before handling: 1312
Number of outliers in HbA1c_level after handling: 0

Number of outliers in blood_glucose_level before handling: 2031
Number of outliers in blood_glucose_level after handling: 0
```

4.4 Imbalance Check:

- In our dataset reviews, we discovered that data is imbalanced in the distribution of data. Recognizing that such imbalances can skew analysis results, we took steps to balance the dataset.
- This adjustment is crucial for ensuring the accuracy and fairness of our predictive models. Now, with a balanced dataset, we can expect more reliable and representative outcomes from our analysis.

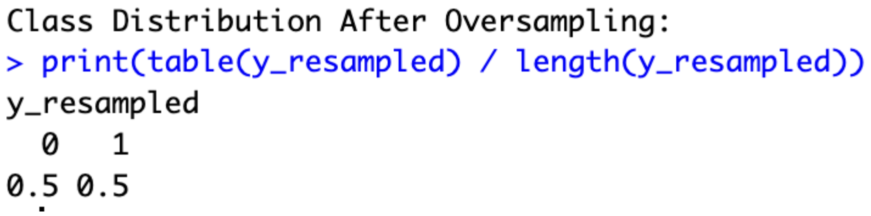


```
> cat("Imbalance Ratio:", imbalance_ratio, "%\n")
Imbalance Ratio: 9.675579 %
```



Oversampling: Imbalance Treatment

- Class imbalance in this given dataset is effectively addressed by the balance treatment approach that uses oversampling, such as SMOTE. It contributes to the development of a more balanced representation of patients with and without diabetes, resulting in a predictive model that is more accurate and trustworthy.
- To balance the dataset, oversampling entails inflating the number of cases from the minority class intentionally. Oversampling would specifically raise the number of diabetic cases in the sample when it comes to diabetes prediction.



```
Class Distribution After Oversampling:
> print(table(y_resampled) / length(y_resampled))
y_resampled
  0   1 
0.5 0.5 
```

5. Data Modelling:

5.1 Splitting the dataset:

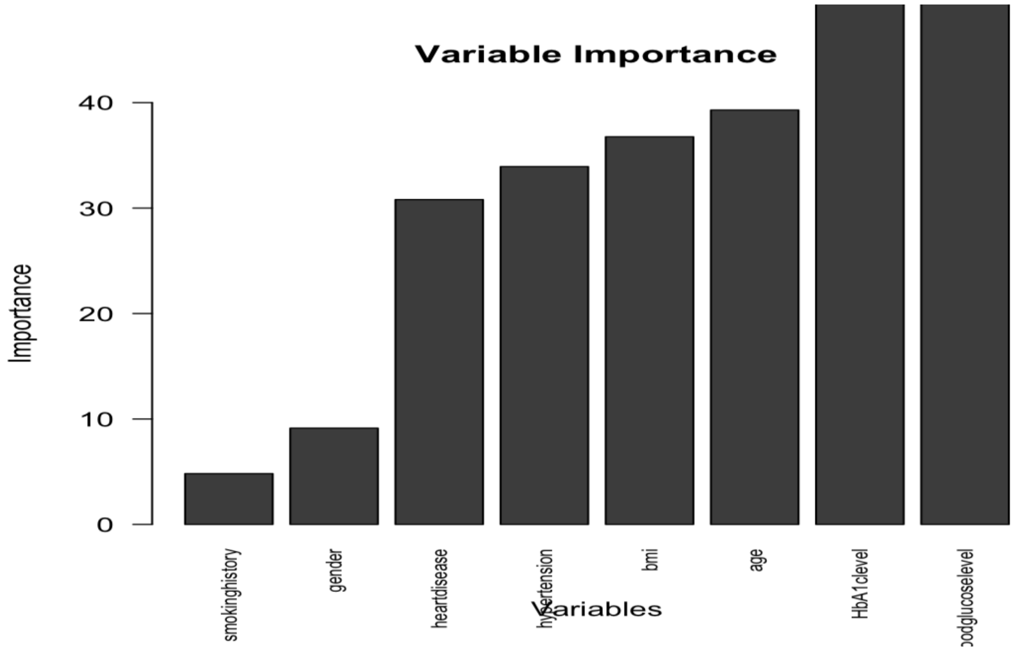
- This involves splitting the dataset into features (X) and the target variable (y) which will further be divided into Training and Testing sets using built-in function
- These train and test data are used further for model building

5.2 Impact Graph:

Machine learning models are frequently interpreted using Impact Graphs, which assist in determining the variables which are the most indicative of the model's results.

- Out of all the variables displayed, "blood_glucose_level" has the greatest relevance score (about 40), making it the most significant predictor.
- "HbA1c_level" is another important variable that comes with the Second highest relevant score.
- "Body Mass Index" (BMI) and "Age" are also quite important, although not as much as "blood_glucose_level" and "HbA1c_level."
- "Heart_disease" and "hypertension" follow with scores of modest relevance.
- In comparison to the other categories, "Gender" has a lower score, meaning it has less predictive potential.
- "Smoking_history" has the lowest significance score out of all the factors displayed, indicating that it has the least predictive potential.
- We considered the threshold to be 35% where any variable having impact more than the threshold is derived into a new dataset

- Now that we have a new derived dataset based on threshold, we apply models on both original dataset which contains all the variables and the new derived dataset based on threshold
- Models built on both the datasets are then compared across metrices and the best out of it is chosen

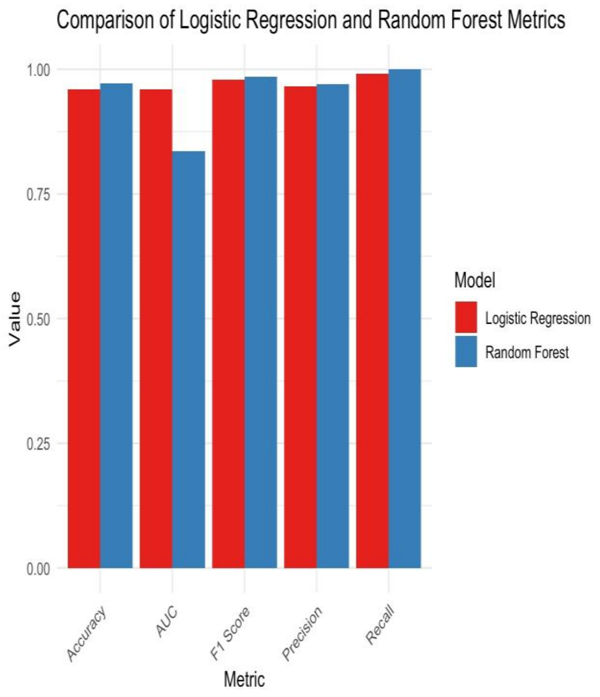


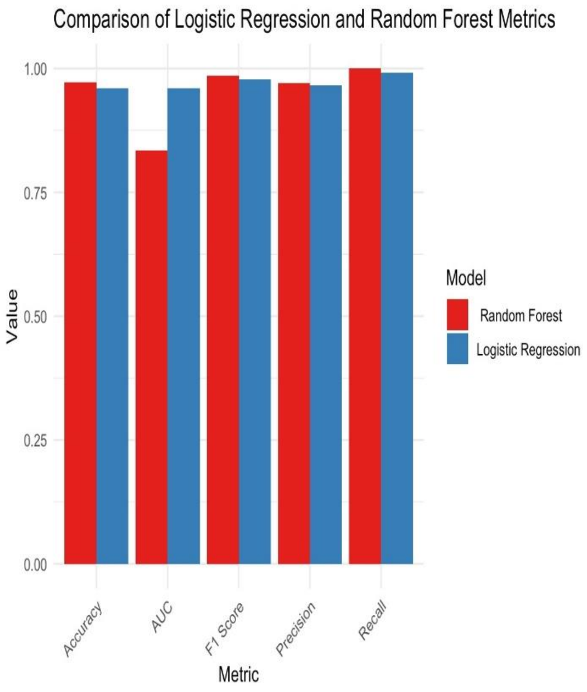
5.3. Logistic Regression (Common Classifier):

- For binary classification problems, logistic regression is a smart place to start because it's simple to understand and apply. It predicts the probability that a given data point belongs to a particular category.
- It facilitates comprehension of the relationship between one or more independent variables (such as age, weight, or blood sugar levels) and the dependent binary variable (the occurrence of diabetes).
- The results of Logistic Regression are easily interpreted, which facilitates the understanding of the significance of individual features (e.g., blood sugar levels, age). It can provide information on how certain characteristics, such age or body weight, relate to the chance of developing diabetes.

5.4. Random Forest : Wilson

- Random forest is a flexible machine learning technique that may be used for both classification and regression applications.
- In order to accurately predict diabetes, Random Forest manages the dataset's combination of categorical and numerical factors.
- The dataset contains a number of diabetes risk factors, including age, BMI, blood sugar levels, and more. The intricate connections and interactions between these variables can be captured by random forests.

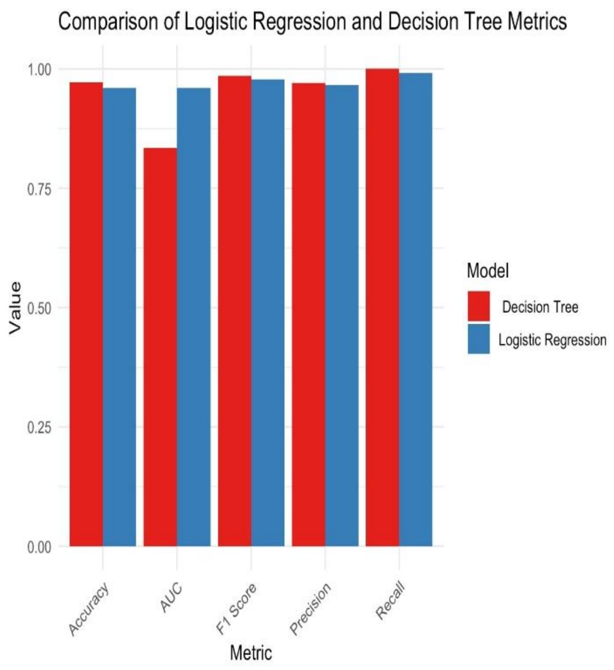


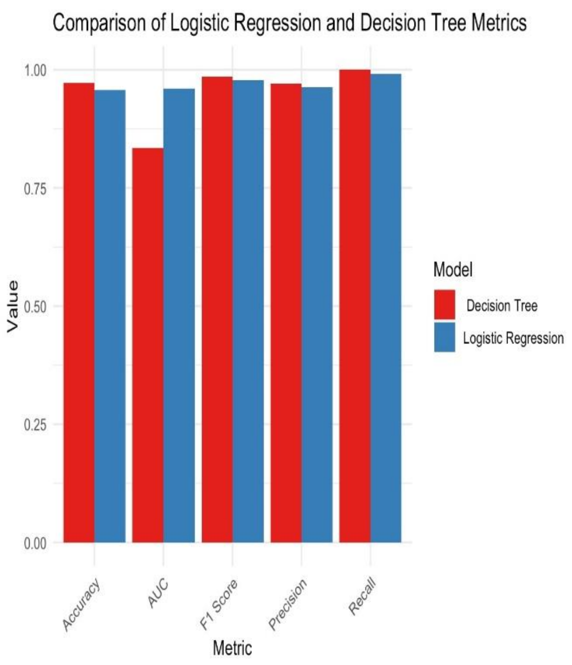


5.5. Decision Tree : Nimisha

- A decision tree is a kind of tree structure that looks like a flowchart, where the internal node represents a feature or characteristic, the branch represents a decision rule, and each leaf node represents the conclusion.
- There are numerical and categorical factors in the dataset, such as age, BMI, and blood glucose levels, as well as gender and smoking history. Decision trees are able to handle this type with ease.

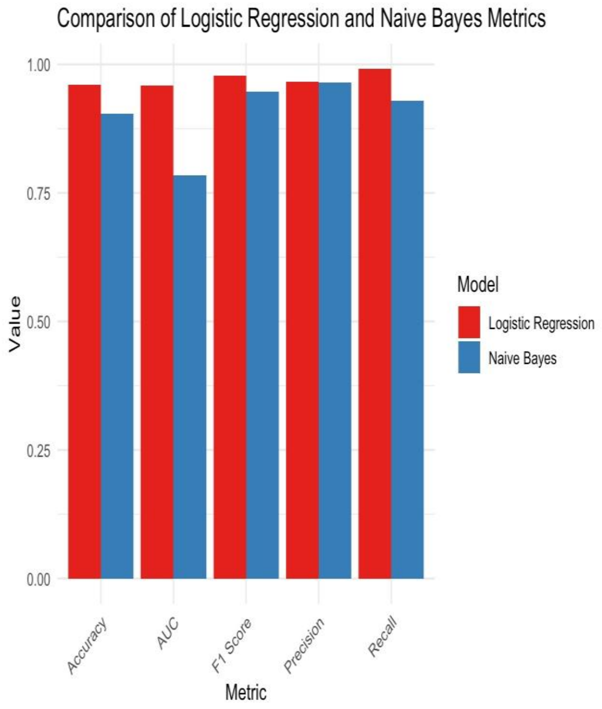
- Creating a decision tree shows which features are most important for predicting diabetes, helping us better understand and prevent the condition.

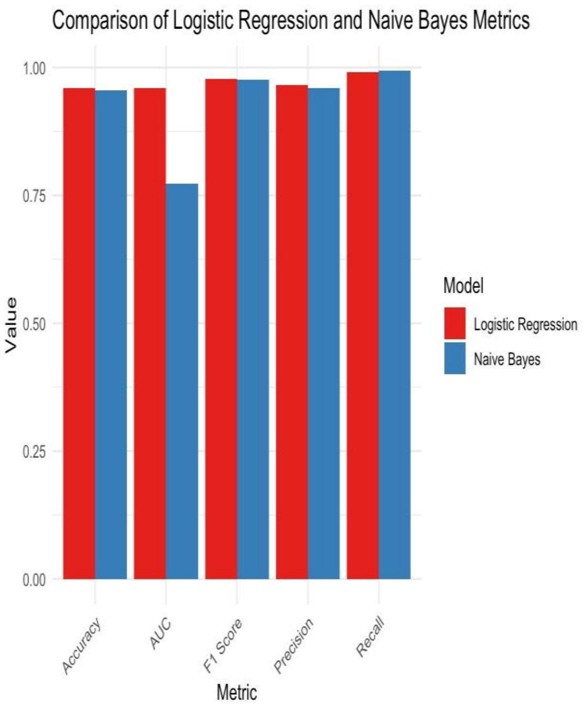




5.6. Naive Bayes : Nayana

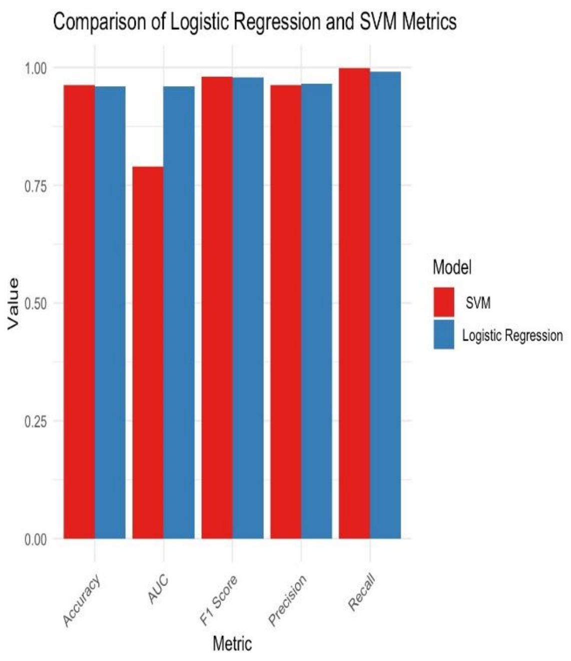
- When working with datasets that contain categorical input factors, such as gender or smoking history, Naive Bayes is a good choice. These characteristics can be processed effectively to predict the risk of diabetes.
- When extraneous features are included, Naive Bayes remains comparatively resilient. This is especially useful for medical datasets when certain variables may not directly affect the course of the disease.
- With its ability to handle multiclass classification issues well, Naive Bayes can provide insights into different levels or types of diabetes risks provided.

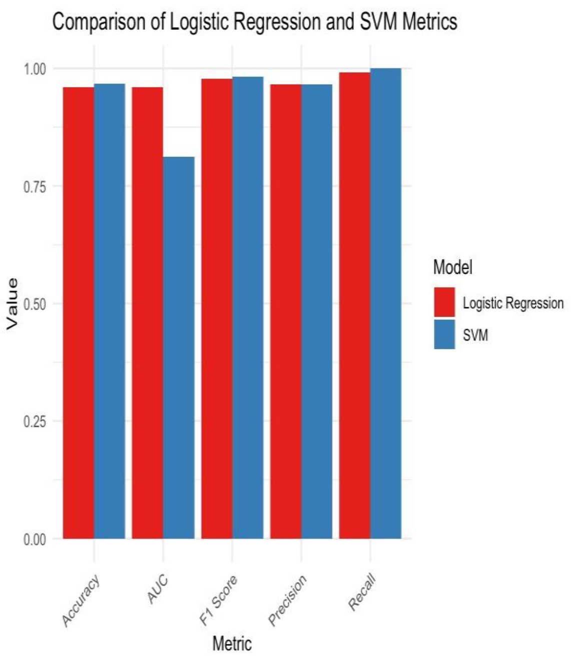




5.7. Support Vector Machine (SVM) : Jegan

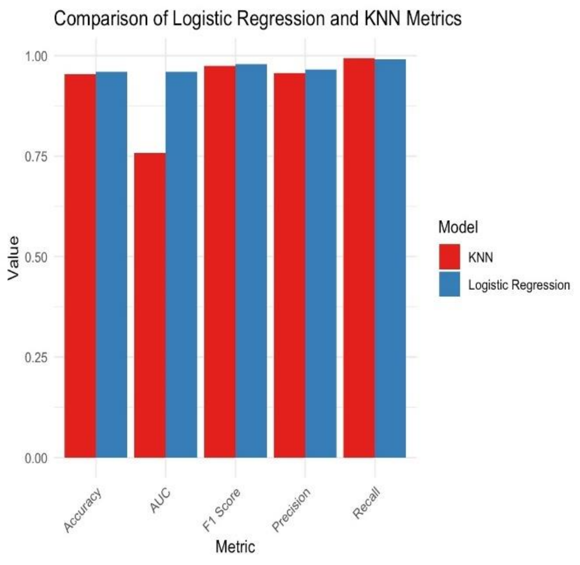
- Support Vector Machines (SVM) create a clear dividing line (called a hyperplane) to accurately and distinctly separate diabetic and non-diabetic patients for diagnosis.
- SVM with kernel functions like RBF, polynomial, can effectively represent non-linear relationships when there is a non-linear link between the features (e.g., blood glucose levels, BMI) and diabetes.
- SVM works really well when different groups are clearly separated. This is good for medical data, because healthy and sick people often show different patterns.

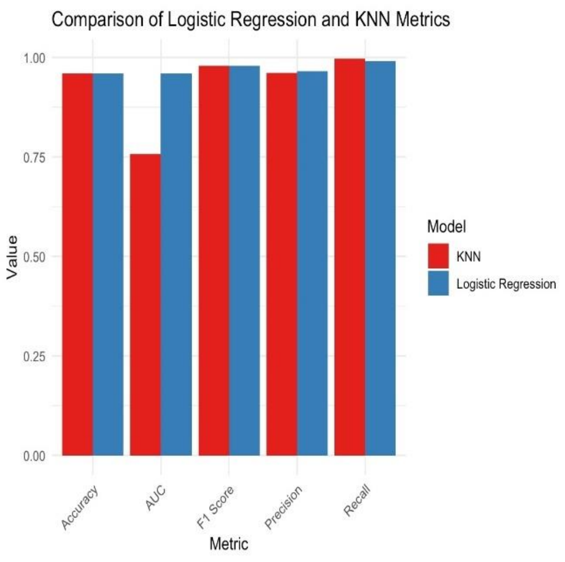




5.8. K- Nearest Neighbors (KNN) : Vasanth

- When applied to a diabetes dataset, the k-Nearest Neighbors (k-NN) algorithm provides a dependable and user-friendly method of estimating the likelihood of developing diabetes by comparing an individual's health information with that of the most similar people.
- Based on a variety of characteristics, KNN is used to categorize people as likely or unlikely to have diabetes (such glucose level, age, heart disease, etc.). The forecast made by the model would be based on the majority class (diabetes or not) of the 'K' closest people found in the dataset**.**
- If the diabetes prediction dataset has past cases similar to new data points, KNN may be a suitable option for finding patterns and categorizing new data points as either indicative of diabetes or not.





6.Conclusion:

6.1 Comparison of statistical metrices

1) Dataset 1:

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.9718000	0.9704676	0.9995993	0.9848180	0.8361
Decision Tree	0.9718667	0.9701704	1.0000000	0.9848594	0.8345
Naive Bayes	0.9041333	0.9651348	0.9287796	0.9466083	0.7838
Support Vector Machine (SVM)	0.9627333	09626142	09980328	0.9800036	0.7904
Logistic Regression	0.9600333	0.9661872	0.9910018	0.9784372	0.9594
K-Nearest Neighbors (KNN)	0.9531333	0.9571368	0.9932605	0.9748641	0.7572

2) Dataset 2:

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.9718667	0.9701704	1.0000000	0.9848594	0.8345
Decision Tree	0.9718667	0.9701704	1.0000000	0.9848594	0.8345
Naive Bayes	0.9561000	0.9598452	0.9935883	0.9764253	0.7731
Support Vector Machine (SVM)	0.9681667	0.9663792	1.0000000	0.9829022	0.8127
Logistic Regression	0.9578333	0.9630088	0.9920219	0.9773001	0.9594
K-Nearest Neighbors (KNN)	0.9601333	0.9615358	0.9962842	0.9786016	0.7572

6.2 Inference across deployed models

1) Random Forest:

- **Strengths**: Random Forest is an ensemble model that combines multiple decision trees, providing better generalization and reducing the risk of overfitting. It can handle high-dimensional data and is robust to outliers and noise.
- **Considerations:** Random Forest can be computationally expensive and may require more training time compared to individual decision trees.
- **Performance**: Performance is very consistent across both datasets, with a slight improvement in Recall in dataset 2. The AUC score is slightly lower in dataset 2.

2) **Decision Tree:**

- **Strengths:** Decision Trees are easy to understand and interpret. They can handle both numerical and categorical data, and can capture non-linear relationships and interactions between features.
- **Considerations**: Decision Trees can be prone to overfitting if not properly regularized. They may also be sensitive to small variations in the training data.
- **Performance** : The Decision Tree has identical metrics in both datasets, which is unusual and suggests overfitting might be occurring, as it's rare for a model to have exactly the same performance metrics on two separate datasets unless the data is very similar.

3) **Naive Bayes:**

- **Strengths:** Naive Bayes is a probabilistic model that performs well with small training sets and can handle both continuous and categorical features. It is computationally efficient and often provides fast predictions.
- **Considerations:** Naive Bayes assumes that features are conditionally independent given the class label, which may not hold in real-world scenarios. It may struggle with correlated features.
- **Performance :** There's an improvement in Recall and F1 Score in dataset 2, but the AUC has slightly decreased**.**

4) **Support Vector Machine (SVM):**

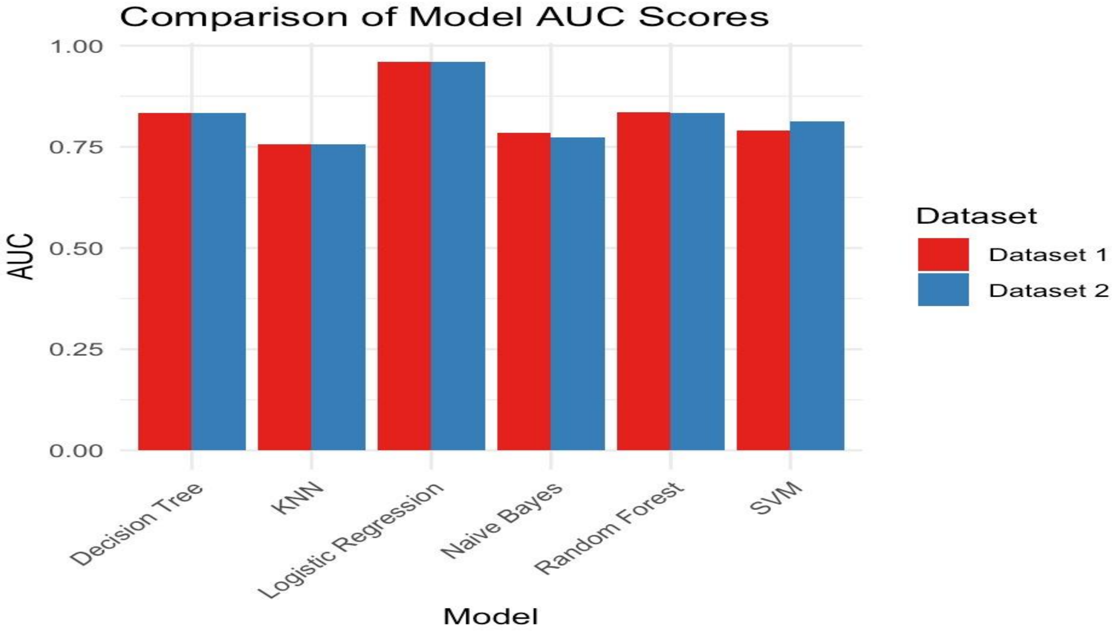
- **Strengths:** SVM is effective in cases where there is a clear margin of separation between classes. It can handle both linear and non-linear decision boundaries through the use of different kernels. SVM is robust to overfitting when the regularization parameter is properly chosen.
- **Considerations:** SVM can be sensitive to the choice of hyperparameters, such as the kernel and the regularization parameter C. It may also be computationally expensive for large datasets.

- **Performance :** SVM shows a notable improvement in Recall (perfect in dataset 2), with a significant increase in AUC, suggesting better generalization in the second dataset.

5) **Logistic Regression:**
   - **Strengths:** Logistic Regression is a simple and interpretable model that works well when there is a linear relationship between the features and the target variable. It can handle both binary and multi-class classification problems.
   - **Considerations:** Logistic Regression may struggle with complex relationships and non-linear interactions between features.
   - **Performance** : The performance is relatively stable across both datasets, with a negligible change in the metrics.

6) **K-Nearest Neighbors (KNN):**
   - **Strengths**: KNN is a non-parametric model that does not assume any specific functional form. It can handle complex decision boundaries and can be effective when there is local structure in the data.
   - **Considerations:** KNN can be sensitive to the choice of the number of neighbors (K) and the distance metric. It may also suffer from the curse of dimensionality when the number of features is large.
   - **Performance** : - There's a small improvement in Accuracy and F1 Score in dataset 2, but the AUC remains the same.

6.3. Comparison of AUC score:



Best Model Evaluation:

Upon examining the comparison of metrics across different models for both datasets, it becomes evident that the values fall within a similar range. This consistency is attributed to the nature of the dataset, where each variable exhibits independence from one another. Consequently, the models find it easier to learn from and make predictions on this type of dataset, resulting in metrics that are nearly identical across all models.

Additionally, we note that there are minimal variations in the metric values between dataset1 and dataset2. This can be attributed to the reduced impact of the dropped variables in dataset2, as these variables have limited influence. Consequently, the models display less disparity when comparing the two datasets.

To determine the best model, we should look for high performance in F1 Score and AUC, as these are more robust metrics for classification tasks.

1) AUC:
    a. Logistic Regression has the highest AUC in both datasets, which is a strong indicator of its overall performance.
    b. SVM also shows a significant increase in AUC in dataset

2) F1 Score:
    a. Random Forest and Decision Tree have the highest F1 Scores, but their perfect Recall suggests a potential overfitting issue.
3) Consistency:
    a. Logistic Regression and SVM show good consistency across datasets, with SVM having a perfect Recall in the second dataset.

Given these observations,

- The Support Vector Machine (SVM) stands out as the best model for dataset 2 due to its perfect Recall and substantial increase in AUC, indicating improved performance in distinguishing between the classes.
- Logistic Regression also appears as a strong candidate, considering its high and consistent AUC value.
- However, model selection should also take into account the context of the problem, the potential costs of false positives vs. false negatives, the interpretability of the model, and computational efficiency.
- If the cost of false negatives is high (which is often the case in medical diagnoses), a model with a higher Recall like SVM for dataset 2 might be preferred.

Therefore, our conclusion favours the usage of Logistic Regression as the superior model, despite the presence of strong competitors. This preference is primarily driven by several factors:

1. High Accuracy: The Logistic Regression model demonstrates a high level of accuracy in its predictions.

2. Consistent AUC Score: The model consistently achieves a favourable Area Under Curve (AUC) score, indicating good performance in distinguishing between classes.

3. Simplicity: Logistic Regression employs a simpler algorithm compared to other models, making it easier to understand and interpret.

4. Faster Training: The model can be trained relatively quickly, reducing the time required for model development and implementation.

5. Handling Imbalanced Data: Logistic Regression exhibits the capability to handle imbalanced datasets effectively, addressing potential challenges associated with imbalanced class distributions.

Considering these factors collectively, Logistic Regression emerges as the preferred model in this scenario as it also performs efficient in binary classification scenarios.

Individual Contribution:

1.Jeganathan Duraisamy:

Our team chose a dataset intended to predict diabetes following a thorough search. The first stage was to gain a full understanding of the dataset, beginning with the identification of the target variable that was necessary for our prediction study. Next, I evaluated each column to ascertain if the data were categorical or numerical—a critical distinction for the preprocessing stage and I found the structure and summarized the dataset.

I worked on the Support Vector Machine technique on my own, gained an understanding on how it operates, and applied it to both datasets, then I compared the efficiency in accordance with all the key metrics of the models constructed on both datasets, whereas SVM shows a notable improvement in Recall (perfect in dataset 2), with a significant increase in AUC, suggesting better generalization in the second dataset. Precision measures the number of true positives over all positive predictions. SVM shows slightly higher precision than Logistic Regression, suggesting it may be better at minimizing false positives.

2.Wilson Ayyappan:

After gaining insights from the data, I focused on analysing it. To understand the data and its characteristics, I plotted several graphs to understand its distribution. I used correlation graphs to see how other variables related to our target variable. Some variables were of greater impact to the target variable whereas some showed no impact at all. By looking at the dataset's past trends, I could make better conclusions through my analysis. After EDA, I focused on understanding the dependencies within it. I created a correlation graph to understand the distribution of all variables across the target variable.

Then I have learned about and used the Random Forest algorithm on both data sets. I have calculated all the key metrics and compared them with our common classifier Logistic Regression's metrics. I have observed that the performance is very consistent across both datasets, with a slight improvement in Recall in dataset 2 meanwhile, the AUC score is slightly lower in dataset 2.

A. Will

3.Nayana Suresh:

After getting the insights through the dependency graph, I've created a correlation matrix to understand the relation and correlation between the variables. Further I created a variable importance graph to identify their impact on the target variable. This helped me decide which variables are most suitable for model building and which ones have the least impact on the target variable. I have plotted a feature importance graph to help us set the threshold for data splitting. After splitting the dataset, Vasanth ,Nimisha and I worked on training the Logistic Regression model and calculated all the metrics as it is our common classifier.

Since we split the data into two sets, I specifically worked on implementing the Naïve Bayes algorithm on both datasets. Then I Worked on obtaining the all-relevant metrics and compared them to determine the best model.

Nayana.S

4.Vasanth Chikkanan:

After understanding the data through EDA, I have converted the categorical variables into numerical variables. Then I took a closer look at the dataset to identify any inconsistencies as a part of Data pre-processing. My examination revealed the absence of null values, negating the need for any treatment in that aspect. Additionally, I pinpointed duplicates across the variables, constituting roughly 0.3% of the dataset. Consequently, we opted to retain the original entries and eliminate the duplicates, leading to a reduction in the dataset's size.

Moreover, we derived a subset from the original dataset, focusing on specific variables determined by a threshold derived from the impact graph. Currently, I am delving into the K - Nearest Neighbours algorithm. After gaining a comprehensive understanding of its mechanics, I applied this algorithm to both datasets. Moving forward, I have found all the key metrics and compared those with the Logistic Regression metrices by plotting the bar graph for the analysis. Finally, I found a small improvement in Accuracy and F1 Score in dataset 2, but the AUC remains the same.

C.Va

5.Nimisha Rajesh:

In my part, I assessed the balance concerning our target variable and identified an imbalance with a ratio of 9.8. To address these inconsistencies, I chose oversampling over under sampling to ensure data preservation. After implementing oversampling techniques, we achieved a balanced dataset. Then I had a close look at finding the outliers and came up with 6 features containing outliers. I calculated Inter Quartile Range (IQR) in order to set the limit to derive outliers and I've managed to handle them. Following these steps, the data was ready for analysis and modelling.

Upon your suggestion, we created two datasets from the original one, considering specific variables based on a threshold from the impact graph. I individually studied the Decision Tree algorithm, its types, and its working principles, implementing it on both datasets. I then assessed the accuracy and other key metrics of both models and concluding that Logistic Regression outplayed Decision Tree in various aspects.

Nimisha