University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Data Science Project

7PAM2002-0509-2023

Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

DIABETIC PREDICTION USING MULTIPLE MACHINE LEARNING ALGORITHMS

Student Name and SRN:

Jaswanth Jagadabhi and 21087866

Supervisor: William Copper

Date Submitted: 29-08-2024

Word Count: 7320

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I have not used chatGPT, or any other generative AI tool, to write the report or code (other than where declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Jaswanth Jagadabhi

Student Name signature: jaswanth jagadabhi

Student SRN number: 21087866

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

DIABETIC PREDICTION USING MULTIPLE MACHINE LEARNING ALGORITHMS

Acknowledgments

I want to express my gratitude to everyone who made this research possible. Grateful to the data scientists and the healthcare practitioners for contributing and assisting with the research. There are too many individuals that played important roles in my research process, but I owe much to my academic advisor for their encouragement and critical comments that improved the value of this paper. I also thank peers and colleagues for useful discussions and encouragement during the studies' process. Last but not least, the authors of the used dataset are also thanked, without which this research could not have been conducted. This research would not have been possible if it wasn't for their assistance.

Abstract

This study aims to investigate the following factors involved in early diagnosis of cases of diabetes with a database of 100,000 entries and some other constituent variables; including but not limited to Age, Sex, Hypertension status, Smoking profile, Heart diseases, HbA1c level, Body Mass Index, Blood glucose level. Linear Regression, Support Vector Regression, Logistic Regression and Random Forest Classifier are the various algorithms used in the research in order to evaluate the competency of the models in predicting diabetes.

Some elementary examination reveals no null values, which improves reliability for future modeling with the dataset. Therefore, Linear Regression and Support Vector Regression show moderate accuracy at best and high SE error while possessing low influences. However, in the case of Logistic Regression the accuracy is as high as 95 percent. 84% but is not very good at categorizing cases of diabetes illustrated by the high rates of false negatives in its analysis. Random Forest classifier can be validated as better than other models with 97% accuracy. 81% but also 45% of the cases which reveal that the application, especially in diagnostic stage, has difficulties in recognizing diabetics, which is indicated by TNs in the confusion matrix. These machine learning method shows that the approach of those participants based on their BMI and blood glucose will split them into three different risk zones for diabetes.

Table of Contents

Chapter 1: Introduction

1.1 Introduction

Diabetes is a non-curable and persistent condition which impacts a large number of people and is defined by the body's incapacity to maintain proper blood glucose ranges. The incidence of diabetes is steadily increasing and becoming a global threat to health-care systems, hence early diagnosis and control of the condition remains paramount to decrease calamitous consequences. New approaches in the application of ML show some promising opportunities for increasing the probability of an accurate diagnosis and prediction of diabetes as well as its management. Such reasoning makes it possible to identify individuals that may require early attention from the healthcare providers and thus, necessary action can be taken in good time.

This paper deals with the use of different machine learning classifiers to diagnose diabetes based on significant indicators that include blood glucose level, BMI, and HbA1c. The addition of many ML algorithms like Logistic Regression, Random Forest and K-Means clustering helps the system to analyze more and the prediction becomes more accurate.

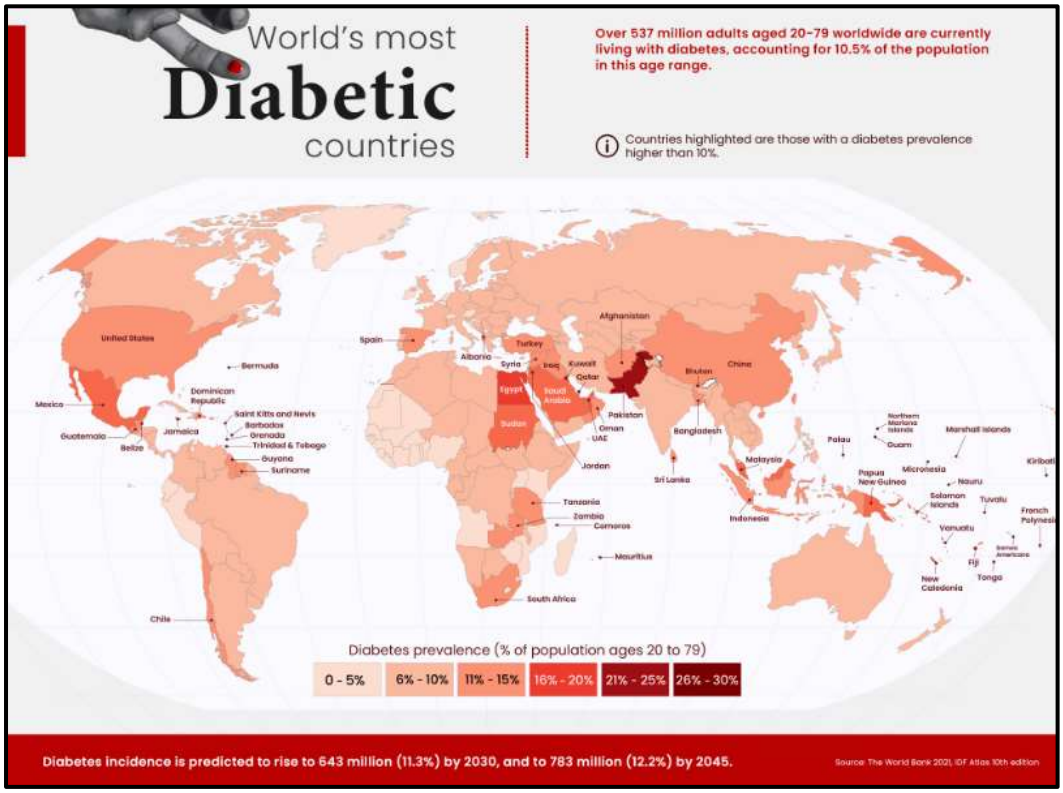


Figure 1.1.1: World's Most Diabetic Country
(Source: visualcapitalist.com, 2023)

1.2 Aim and Goals

Aim

The aim is to recognize crucial medical and also demographic predictors of diabetes and implement adequate approaches for classifying patients utilizing multiple machine learning approaches.

Goals

- To recognize significant medical and also demographic predictors of diabetes.
- To establish and develop regression and classification approaches for accurate diabetes classification.
- To implement regression models predicting the blood glucose levels according to the demographic alongside medical features.

- To evaluate the clustering approaches for identifying the distinct patient groups having the varying diabetes risk profiles, alongside the characteristics.

1.3 Research Questions

- Which demographic and medical features are the most crucial predictors of diabetes?
- How accurately may regression and classification models classify the patients as diabetic or non-diabetic according to their particular characteristics?
- How adequately can the regression approaches predict the blood glucose levels utilizing demographic data along with medical data?
- What distinct patient groups having varying diabetes risk profiles may be recognized utilizing clustering approaches, and also what are their defining features?

1.4 Background Research

Diabetes has been perceived as a developing worldwide medical problem, requiring early diagnosis alongside intervention to prevent severe complexities. Within this evaluation, the study is to develop accurate prescient methodologies for the diabetic prediction utilizing the dataset obtained (Butt *et al.* 2021). The following methods, involving "linear regression", "logistic regression", "Random Forest Classifier", and also "K-Means clustering", will be utilized. These strategies will recognize crucial indicators of diabetes and characterize patients as diabetic or non-diabetic. The broad dataset improves the prescient models' precision. The following Ethical contemplations are tended to as the dataset is anonymized and agrees with GDPR necessities (Jaiswal *et al*. 2021). This study will add to comprehending diabetes risk factors and supporting the advancement of enhanced analytic and preventive approaches.

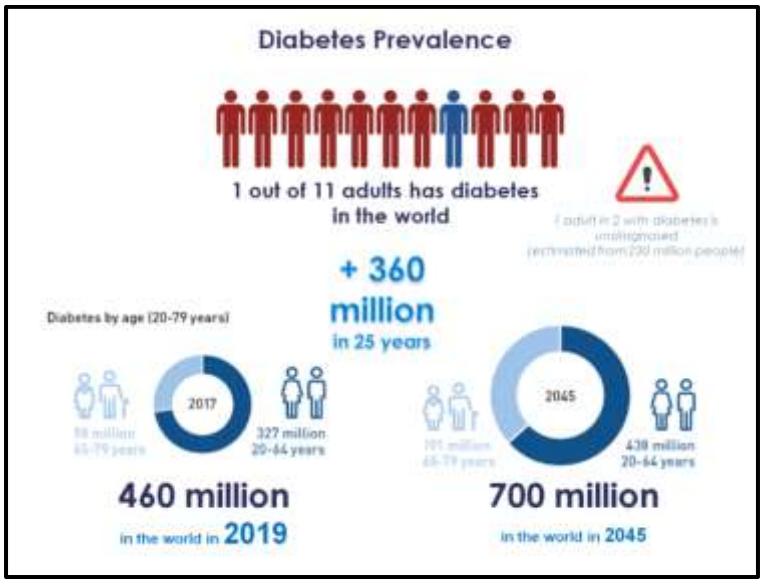


Figure 1.4.1: Diabetes Prevalence

(Source: www.pep2dia.com/prediabetes, 2021)

1.5 Research Rationale

The rationale for this study is grounded within the crucial requirement for the viable diabetes prediction alongside the avoidance approaches because of its rising worldwide predominance. Through using an exhaustive dataset from the Kaggle, which incorporates different clinical alongside demographic data, the study expects to enhance the comprehension of crucial diabetes indicators (Nahzat and Yağanoğlu, 2021). The following machine learning strategies will be utilized for predicting precise predictive approaches, which will assist early diagnosis alongside intervention. The utilization of anonymized, GDPR-agreeable data guarantees ethical norms are managed (Suresh *et al.* 2020).

1.6 Study Structure

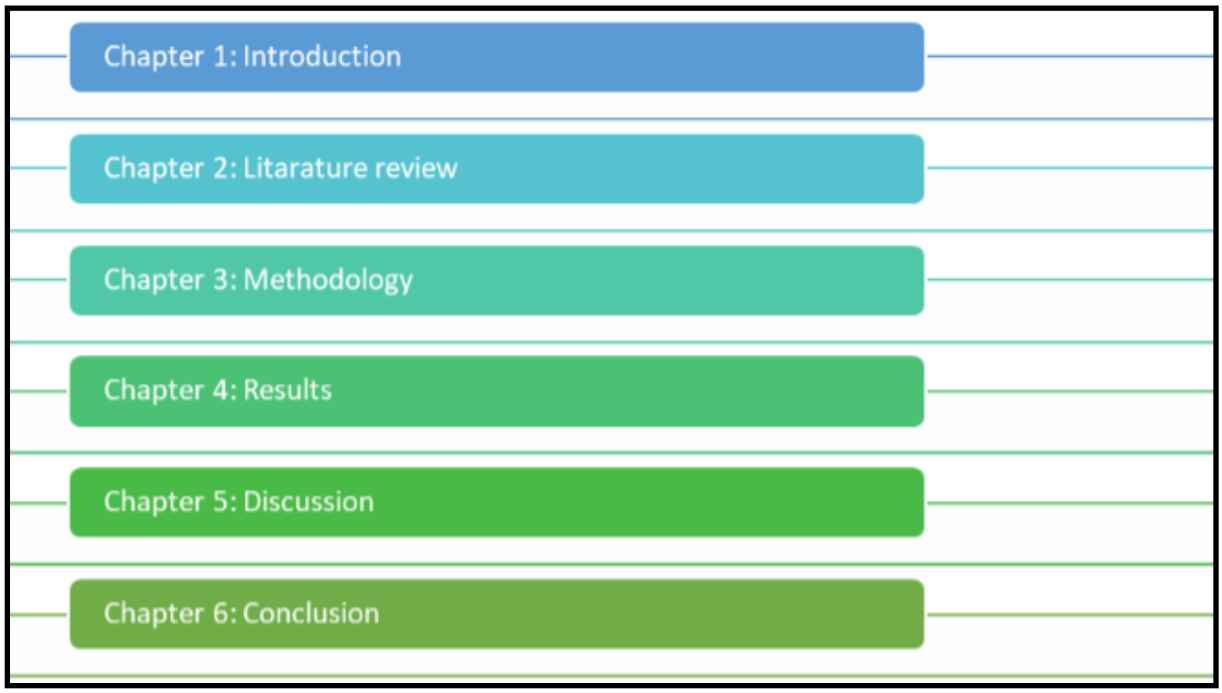


Figure: Research Structure

(Source: Self-developed)

1.7 Summary

Within this study, accurate predictive approaches for diabetic prediction will be generated utilizing different machine learning approaches. The following clinical and demographic predictors of the diabetes will be recognized, alongside patients will be classified diabetic or the non-diabetic. The Regression approaches will predict the blood glucose levels, along with clustering will distinguish patient groups with shifting diabetes risk profiles. The Ethical norms are managed, guaranteeing important details into diabetes risk elements along with prevention.

Chapter 2: Literature Review

2.1 Introduction

For activities like diabetes detection, using the combination of the two approaches may yield better results generating stronger and reliable prediction models. Further research should be devoted to the development of more sophisticated mixes of the two approaches and to fine-tuning such approaches in different classification problems. One of its main advantages is also its technical simplicity, while the results are easily interpreted in terms of how predictor variables influence the likelihood of diabetes. However, as it was discussed, it is important to remember that even though logistic regression gives quite reasonable predictions, it might fail to handle non-linear relations and interaction of the predictors, which can be managed in a better manner with further techniques.

2.2 Analysis of Key Papers

2.2.1 *Joshi and Dhakal (2021)*- Effectiveness of Logistic Regression in Diabetic Prediction Models

Methods and Data: Fortunately, logistic regression has been shown to be an efficient technique for the binary classification problem, especially for diabetes. It brings out the working of a dependent variable that can have only one of two responses (diabetes presence or absence) along with multiple predictors and the probabilities of occurrence are captured by the logistic function. It would allow the definition of risk factors with an emphasis on the risk of developing diabetes, providing a comprehensible process of classification.

Final Thoughts and Findings: Logistic regression as to the method and data has been used in different works for predicting Diabetes with factors such as glucose level, BMI, age and family history. For instance, the study by Joshi and Dhakal (2021) exposed the use of both logistic regression and machine learning algorithms in predicting type 2 diabetes. The model was confirmed to be relatively stable in predicting such factors as glucose concentration and BMI that are identified in the literature to be relevant in the prediction of the risk. The authors were able to use logistic regression to show that they were accurate in predicting the risk of diabetes from these factors.

Review: The last conclusion is that logistic regression has a comparable achievement to other more sophisticated machine learning methods. Lynam et al (2020) opine that the ability of logistic regression to perform well in clinical settings is remarkable especially when it is used to differentiate between the two types of diabetes or when it is used to predict the risk factors of diabetes.

2.2.2 *Sheykhmousa et al., 2020*- Advancements in Random Forest Classifier for Accurate Predictions

Methods and Data: A large improvement in accuracy of the required binary classification, and the classifier's further evolution in the identification of diabetes, are shown in the Random Forest classifier. There are many methods of ensemble learning, in which multiple decision trees are combined to increase the model accuracy and minimise overfitting, which proved to be rather effective in predicting the risk of diabetes. Modern research like Jackins et al., 2021 involves analysis of the classifier performance in the clinic, disease prediction in general, and diabetes in particular.

Final Thoughts and Findings: Possible enhancements of the Random Forest classifier include improved compatibility with high-dimensional data, together with the efficient processes of dealing with missing values. It works during training by growing many trees and for any given input, it returns the mode of the classes to which the tree most closely points. Due to the reductions of variance it achieves, this technique also helps to avoid the overfitting problem that more simplistic models are particularly prone to. Data employed in these studies can be extensive patient data characterising variables as well as glucose levels, BMI, age and family history. These are the variables that are most important in the prediction of diabetes since they

contain all the main modifiable risk factors of diabetes. There has been an enhanced performance of Random Forest classifiers that impact on binary classification accuracy of diabetes. Thus the method of the ensemble can achieve higher accuracy and less sensitivity to overfitting as compared to the single decision trees or even more as compared to other traditional classification techniques (Sheykhmousa et al., 2020). Due to its general capacity to handle various forms of complicated data, it can be regarded as a useful tool in the domain of diagnosis in medicine.

Review: The development trend of Random Forest classification for predicting diabetes further proves that more and more advanced machine learning technology is being implemented in the healthcare field. The model's capacity to produce an accurate prognosis and its capability to integrate a number of features make it suitable for future applications in early diabetes diagnosis and prognosis.

2.2.3 *Chen and Liu (2020)*- Support Vector Regression and K-Means Clustering Applications

Methods and Data: SVR has great utility and has the capability of improving the precision in attaining binary classification such as the detection of diabetes, Machine Learning is also very efficient in performing high dimensionality of subclasses. SVR is an expansion of Support Vector Machines (SVM) explicitly intended for issues managing nonstop result, and it has been stretched out for paired order also. Chen and Liu (2020) have effectively shown the utilization of a SVR model joined with the Machine Learning and sludge shape tumultuous boundary transformation in light of the Machine Learning calculation that advances the prescient precision and vigor of the model.

Final Thoughts and Findings: Therefore, the manner in which SVR and Machine Learning has been applied in binary classification could be seen as an improvement especially when diagnosing medical conditions such as diabetes. What is more, approach the data and fine-tuning SVR models increase classification precision and the overall performance of the methods. The research also indicates that the use of higher levels of algorithmic and approach can solve issues of lack of probability in forecasts, and model stability.

2.2.4 *Muneer and Fati (2020)*- Comparative Analysis of Supervised and Unsupervised Learning Techniques

Methods and Data: Muneer and Fati (2020) made a comparative analysis of the approaches for the detection of cyberbullying on the microblogging platform, Twitter, which involved classification problems of the binary type. Cross-validating a few supervised learning methods as well as calculated relapse and support vector machines with approach algorithms were undertaken. Concerning the assessment systems, the models were tried with the datasets containing both marked and unlabeled occurrences. In a comparable report, Islam et al. (2020) worked on bosom disease forecast with different machine-learning approaches, for example, decision trees, random forests, and support vector machines. Their work on the point zeroed in on a ton of information pre-handling and component extraction to enhance characterization.

Final Thoughts and Findings: The two works talk about the benefits and burdens of involving supervised and unsupervised learning for paired arrangement. According to Muneer and Fati (2020), it has been postulated that supervised methods yield higher classification renown in contrast to unsupervised methods that acknowledge the fact that CRF has engineered precise labels. Islam et al. (2020) noticed that though decision trees and SVMs are highly accurate in prediction, using some basic yet important components such as feature extraction or data segmentation based in an unsupervised manner will enhance the model to a great extent. These outcomes underline that in the process of selecting the most appropriate approach, it is necessary to use peculiarities of the data and purposes of classification

2.3 Literature gap

An assessment of the literature on machine learning techniques and their application in predictive analytics in the medical domain including but not limited to diabetes detection identifies the following research gaps. Although Muneer and Fati (2020) and Islam et al. (2020) provide a detailed review of supervised, unsupervised, and other ML approaches, their primary concentration is on non-healthcare settings and several forms of disease. Thus, there is a shortage of studies which directly advertise these methods in the context of diabetes detection, especially with the datasets that would reflect the differences in the population and risk factors. The reported precision of diabetes binary classification could be improved by more specific studies based on both types of ML: while using target-oriented supervised and complex NMF-based unsupervised algorithms with precise feature selection and improved data augmentation methods.

Most related works employ benchmark datasets and classical supervised statistical learning techniques, while the medical data is seldom uniform and heterogeneous in practice. It was computed that there is a definite need for utilising better methods namely deep learning and hybrid models as well as for employing larger and a more diverse sample size. Therefore, the current research points precisely to the necessity of further investigations that should fill the existing gaps.

2.4 Summary

An examination of the most recent literature shows substantial progress and research limitations with regard to deploying machine learning approaches towards identifying diabetes. The findings from other domains of supervised and unsupervised learning showcase their utility, but a comparison has been scarce, particularly for diabetes detection. Much prior research employs simple algorithms and past databases; thus a lot of features remain untapped and unanalysed. It is widely felt that accuracy in binary classification ought to be improved, through the incorporation of other sophisticated methods of machine learning including hybrid and deep learning.

Chapter 3: Methodology

3.1 commencement

Prime focuses of the specific study include using various machine learning models that are used to analyze and model risk of diabetes given various parameters. To determine the effects of predictors such as age, BMI, blood glucose level on diabetes, the application of Linear Regression was made. For relationships that are nonlinear, the Support Vector Regression (SVR) was used in order to better predict and to work within the error bounds. In binary classification, Logistic Regression was used for; For categorical health indicators the cases were categorized as either Diabetic or Non-Diabetic. Random Forest Classifier was used because this algorithm is effective in working with high-dimensional datasets and gives stable predictions by building a forest of decision trees. Besides, to learn the inherent structure of the data, K-Means Clustering was used and contained natural clusters that could be the indication of relationships between subgroups.

Dataset Selection for the Project

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download

In this dataset, there are 9 different columns that are age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This collected dataset can be used for the purpose of building different types of machine leaning models for making the prediction of diabetic patients.

The data set has records for 100,000 people and the data made up of 9 fields including identifiers of the persons. It includes the following columns: sex, years, a binary typically associated with a disease/ situation, the second binary, status categorisation (i. e., presently, previous, never), numeral, the second numeral, evaluation, last mark.

The data is likely to be somehow aggregated in a way that one is likely to study the direction or the pattern of the attributes that are likely to be demographic, status or other related factors. The gender and the age are basic as indicators of demographics; however, the status descriptor may point to processes or the states of the patient as far as the study in question is concerned. The 'numerical values ratings' might be quantitative – it can be linked to performance or other; while the 'binary indicators' may well be conditions or results.

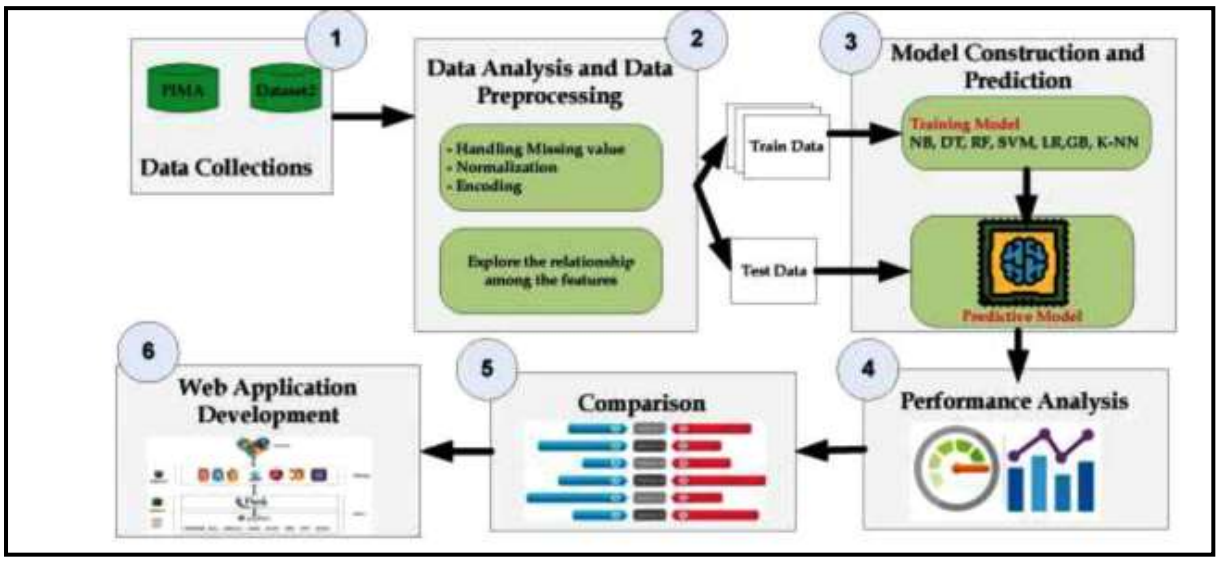


Figure 3.1.1: Workflow Diagram

(Source: Salih and M.S., 2024)

3.2 Exploratory Data Analysis (EDA)

EDA is a crucial phase toward any data driven project, filling the gap among the raw data and significant details. Within this following project, EDA was developed to comprehend the hidden structure of the following dataset and distinguish crucial patterns that could impact the diabetes prediction and also clustering evaluation. The procedure started with an exhaustive evaluation of the specific dataset's key factors, involving the gender, age, BMI, blood glucose levels, alongside the HbA1c levels. Different visualizations, involving scatter plots, histograms, and also count plots, were utilized to address the conveyance and also connections among these factors. For example, histograms were utilized to show the following frequency dissemination of the age and also HbA1c levels, giving a visual comprehension of the data spread and also central tendencies. Count plots were used for displaying the appropriation of the categorical factors like gender, giving experiences into the dataset's demographic structure (Chou *et al.* 2023). Furthermore, EDA involves assessing relationships between various factors to recognize significant indicators of diabetes. The Correlation analysis alongside the scatter plots were utilized to display the connections among the numerical factors, assisting with areas of strength for pinpointing that could be essential for developing the model. EDA likewise recognized the outliers or the anomalies within the following data, which might actually skew the evaluation and also model performance. Tending to these particular outliers guaranteed a more adequate dataset for assuring the modelling endeavours (Afsaneh *et al.* 2022).

3.3 Model Training and Evaluation

Model training alongside the evaluation are critical stages within the analysis, expecting for developing adequate prescient models for the diabetes and also evaluate their performance. Within this following analysis, various machine learning approaches were developed for predicting the diabetes and also categorising individuals in view of significant health indicators, for example, BMI alongside blood glucose levels. The procedure started with splitting the specific dataset into preparing and also testing sets to guarantee that the following model's performance could be assessed on the unseen data. Different approaches were then prepared, involving the Logistic Regression, Random Forest, along with K-Means Clustering. Every approach was fine-tuned for advancing its parameters and also further developing accuracy. For the following classification models, the following performance metrics, for example, Accuracy, Recall, Precision, and also F1-score, "Mean Squared Error (MAE)" and also "Mean Absolute Error (MAE)", also R2 value were determined. These particular metrics gave an extensive perspective on the following models' viability in the prediction of diabetes (Daghistani and Alshammari, 2020). For example, the particular Logistic Regression approach exhibited high accuracy however showed impediments in recognizing specific classes. The K-Means Grouping calculation was utilized to recognize subgroups inside the populace in the view of BMI and age. The particular clustering outcomes, addressed by particular varieties within the visualizations, uncovered significant health-associated trends inside the following dataset.

3.4 Clustering Analysis

The following Clustering evaluation assumes a critical role in distinguishing trends and also subgroups inside the particular dataset, which may illuminate targeted wellbeing health interventions for diabetes management. The specific K-Means clustering technique was used in the analysis that followed to apply it to the population segment based on two important health indicators: Body Mass Index (BMI) in conjunction with age. This strategy assists with uncovering crucial designs within the data, giving experiences that probably won't be evident through adequate statistical evaluation (Hasan *et al.* 2021).

The procedure started by choosing the suitable number of the clusters, guaranteeing adequate separation among the clusters. The following K-Means approach was then applied, parceling the dataset into the distinct clusters. Every cluster addressed the subgroup of the individuals

with comparable BMI and age. Perceptions of the following cluster featured two particular groups inside the populace: one with the lower BMI alongside age , one more with high BMI alongside age (Das *et al.* 2022).

3.5 Research Onion

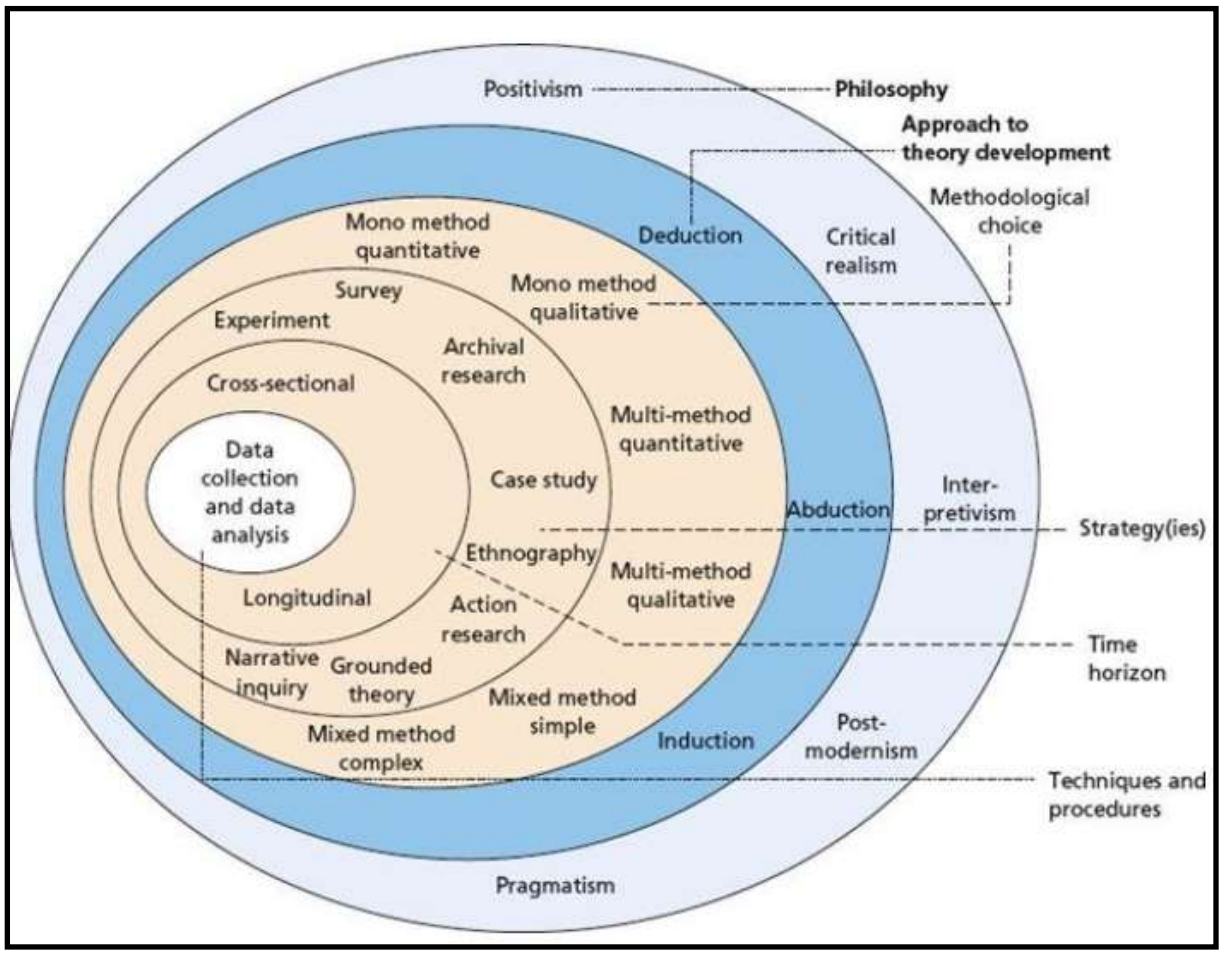


Figure 3.5.1: Research Onion

(Source: Saunders et al.'s Research Onion)

The specific research onion framework directs the systematic strategy to the research. This involves the layers like the research philosophy, strategy, approach, time horizon, choice, alongside the techniques. Assessing the diabetes prediction, this assists in forming the research from defining the particular research paradigm (e.g., positivism) for choosing the methodologies like the regression along with the approach for the data analysis.

3.6 Models Development

Logistic Regression

The statistical approach used for binary classification, Logistic Regression approaches the likelihood of the respective binary outcome, like the diabetes diagnosis, utilizing the logistic function. This is effective for comprehending the relationship among the specific predictor variables along with the likelihood of the particular target event.

Random Forest Classifier

A particular ensemble learning approach, the following “Random Forest Classifier " constructs several decision trees for improving the prediction accuracy. This offers adequate classification by averaging the predictions, mitigating the overfitting, alongside enhancing model reliability within recognizing the diabetic cases. This work makes use of Logistic Regression and Random Forest in assessing the risk factors of diabetes. Logistic Regression is used to analyse the effect

of predictor variables on binary dependent variables in the estimation of prob (probability) of having diabetes. Random Forest being an ensemble learning algorithm results in a better classification by building a number of sub-decision Trees and using the average of the frequencies of the result as the result. Both methods are applied and used to make comparisons with a view to identifying their efficiency in the prediction of susceptibility to diabetes. Further, methods such as Support Vector Machines and K-means clustering are given in detail to understand the ways of increasing the accuracy of the models and to understand the ways of dealing with challenging datasets to get comprehensive techniques for risk assessment.

Linear Regression

Utilized for predicting the continuous outcomes, the following Linear Regression approaches the connection between any dependent variable (blood glucose level) as well as independent variables. This assists in recognizing crucial factors impacting the overall blood glucose levels and comprehending the particular linear associations.

Support Vector Regression (SVR)

SVR, the variant of the Support Vector Machines, is utilized for the regression tasks. This aims to search the hyperplane which best fits the specific data into a particular margin, adequately predicting the continuous outcomes like the blood glucose levels while maintaining the non-linear connections.

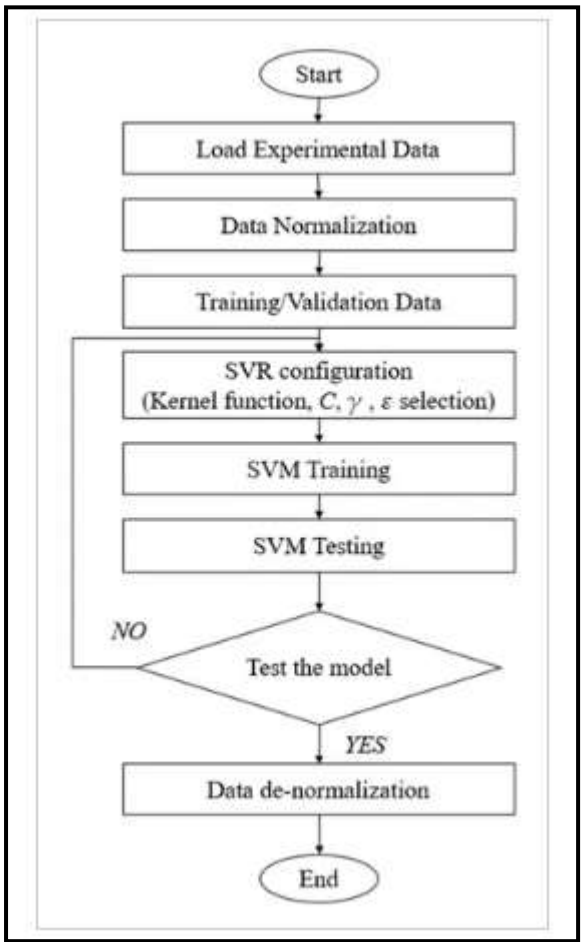


Figure 3.6.1: Flow Chart of Super Vector Regression
(Source: Zhang *et al.* 2022)

The general conception of the framework establishing the SVR model is depicted in the following image. First, data input, namely, loading and normalization, was done. Then, suitable

values for other SVR hyperparameters were identified in order to reduce the error rate. Last, objective assessment of the prediction capability of trained SVR model was done.

K-Means Clustering

An unsupervised learning approach, the following "K-Means Clustering" groups the particular data points into the clusters according to the feature similarities. This assists in recognizing the patterns and also subgroups into the population for the targeted analysis and also interventions.

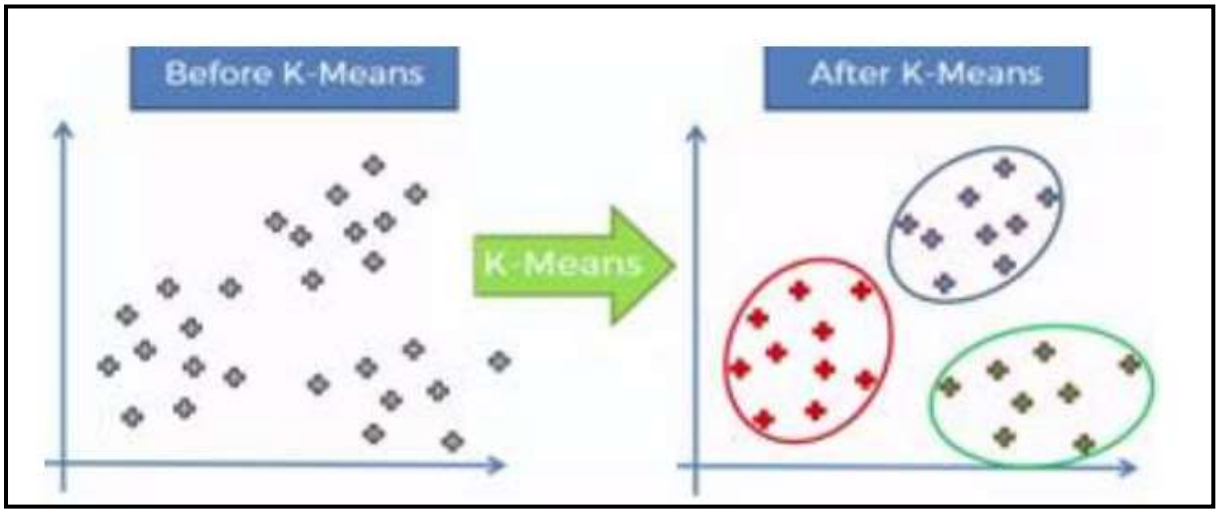


Figure 3.6.2: K-Means Clustering.
(Source: Pranto *et al.* 2020)

Therefore, the determination of the right K, which indicates the number of neighbors, is central to prevent or minimize overfitting or underfitting. In an attempt to deliberate this issue, created the plot of the test and train accuracy from the cluster. Furthermore, the training accuracy is negatively related to the value of K as it increases.

3.7 Ethical Consideration

•It is in this regard that a number of considerations come into play for this study including the manner in which data protection laws including the General Data Protection Regulation (GDPR) are met. EU Regulation 2016/679 is also known as the General Data Protection Regulation and aims at guarding the rights of individuals within the EU and their data. As aforementioned there are no direct identifiers in the dataset described, yet GDPR states that any data that may potentially lead to the identification of individuals must be protected, especially in case of processing of sensitive data which include health data.

•The principles of data processing according to the GDPR include legality, non – discrimination and clarity. Authors have to ascertain that the data has been collected and processed where individuals gave permission in the manner in which their information would be used, how it would be stored or used in research as needed. The information used in this work should conform to these principles, thus guaranteeing the privacy of the individuals whose information has been used in this study.

•GDPR provides data subjects' right to make sure that the necessary organisational and technical safeguards are in place to prevent unintentional or unlawful data destruction, access, disclosure, or alteration. In the light of this study, it implies that the dataset must be secured appropriately and this is only availed to those knowledgeable in administration of the study.

•GDPR also brought about the principles of data minimization as one of the measures to observe. This principle entails that only the data that is relevant to the given purpose of the

research be processed. However, applying the data minimisation principle in this study in a way that only information such as age, gender and health status are collected reduces the probability of putting sensitive information of the patients and clients in the public domain as required by GDPR.

•GDPR empowers the individuals on their data, allowing them to have rights such as an ability to request, correct or erase their data. It is the responsibility of the researchers to respect all these rights in the course of the study thereby offering accountability while enabling control over personal information by the participants. Adherence to such regulations not only makes the research to be ethical but also makes the study to be credible and trustworthy.

Chapter 4: Result

The results that follow stem from the investigation of a dataset that relates to early detection of the disease. The empirical analysis of the study used different mixed method analysis models to examine differences and correlations between important parameters including but not limited to Age, Sex, Hypertension, History of smoking, history of heart diseases, HbA1c, BMI, and Blood glucose levels. The outcomes start with the data set organization analysis, such as completeness analysis and statistic descriptions of data set, in order to check if the dataset is qualified. After this, the study used Linear Regression, Support Vector Regression, Logistic Regression, and Random Forest Classifier models to analyse their capability to prognosticate diabetes. The chapter also applies the K-Means Clustering technique for carrying out a pattern analysis of the dataset and examining the subgroups.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Figure 4.1: Showing top 5 rows of the dataset

The top 5 rows of the dataset are represented here. Crucial features, for example, age, gender, hypertension, smoking history, heart disease, HbA1c level, BMI, and also blood glucose level are incorporated. These factors are crucial indicators for generating the machine learning approaches focused on early diabetes recognition.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
99995	False	False	False	False	False	False	False	False	False
99996	False	False	False	False	False	False	False	False	False
99997	False	False	False	False	False	False	False	False	False
99998	False	False	False	False	False	False	False	False	False
99999	False	False	False	False	False	False	False	False	False

100000 rows × 9 columns

Figure 4.2: Checking Null Values

In this particular step the null values are checked. It is inferred from here that there are no null values presented in this dataset. Within this particular step, the dataset is analyzed for the null values, which may fundamentally affect the accuracy and also dependability of the machine learning approaches. The absence of the specific null values demonstrates that the particular dataset is complete, with all perceptions consisting of values for every variable.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

Figure 4.3: Description of the data

The description of the different variables of this dataset involving count, mean, min, max, standard deviation and so on are represented here. This following figure gives exhaustive statistical details of the dataset's factors, involving fundamental measurements like count, minimum, mean, and so on. The count shows the overall number of the non-null entries for every variable, affirming the dataset's fulfillment. The respective mean provides an average score, providing the sense of the central tendency of the data. The specific standard deviation estimates the dispersion, showing how much the qualities digress from the mean.

```
(100000, 9)
```

Figure 4.4: Shape of the dataset

The shape of the dataset is demonstrated here. It can be identified from this particular figure that this dataset has 100000 rows and 9 columns. It demonstrates that the dataset incorporates 100,000 individual values, every with 9 particular columns or features. Comprehending the particular shape of the dataset is pivotal as this gives details into its size along with structure, which are fundamental for the data preprocessing and evaluation.

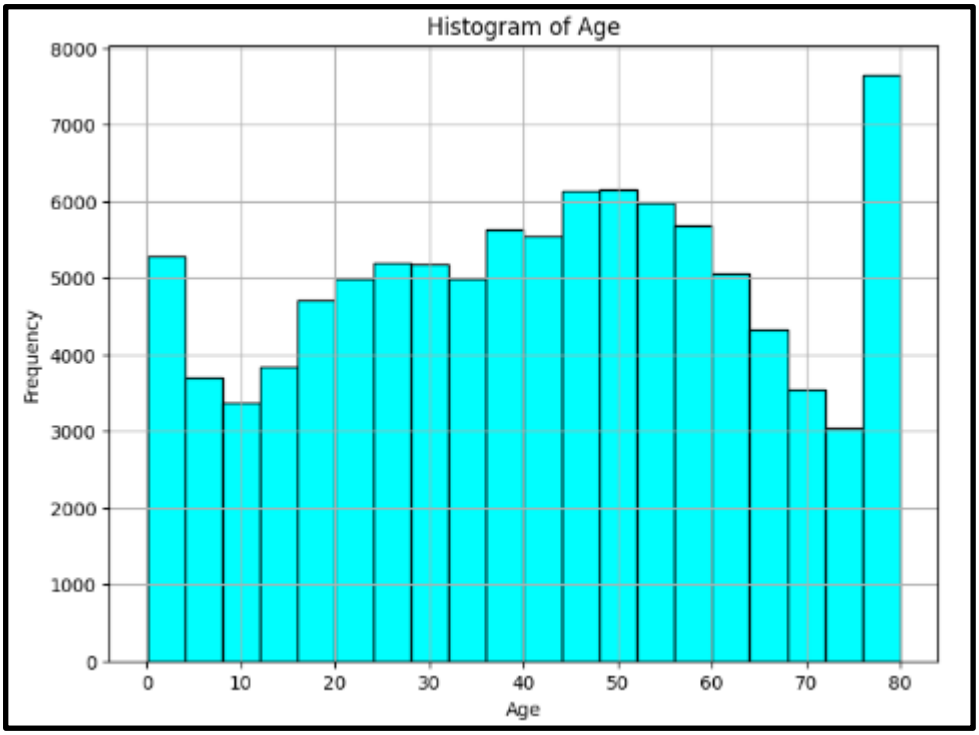


Figure 4.5: Histogram for frequency of the Age

The frequency distribution of the age is illustrated here. This is identified that the frequency is maximum in the age of 80. It shows a critical number of individuals within this following age group. Comprehending the overall age distribution is significant with regards to diabetes prediction, as age is a crucial variable within the risk evaluation for diabetes.

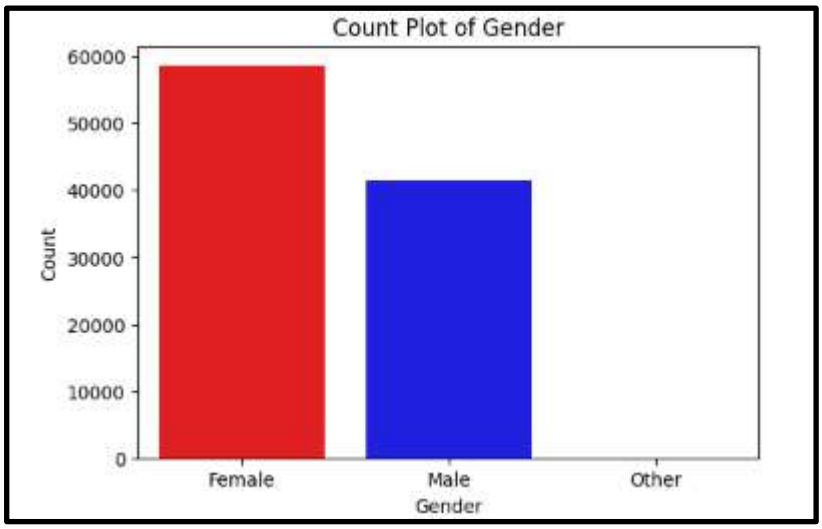


Figure 4.6: Count plot of gender

The mentioned figure represents the count plot of the gender. There is a maximum count of females which is nearly 60000. The following gender uniqueness is huge with regards to diabetes prediction, as this features a possible requirement for the gender-explicit examination and mediations. Comprehending the gender conveyance helps in fitting prescient models to address the subtleties of diabetes risk factors that might differ among the males, females, and also different genders.

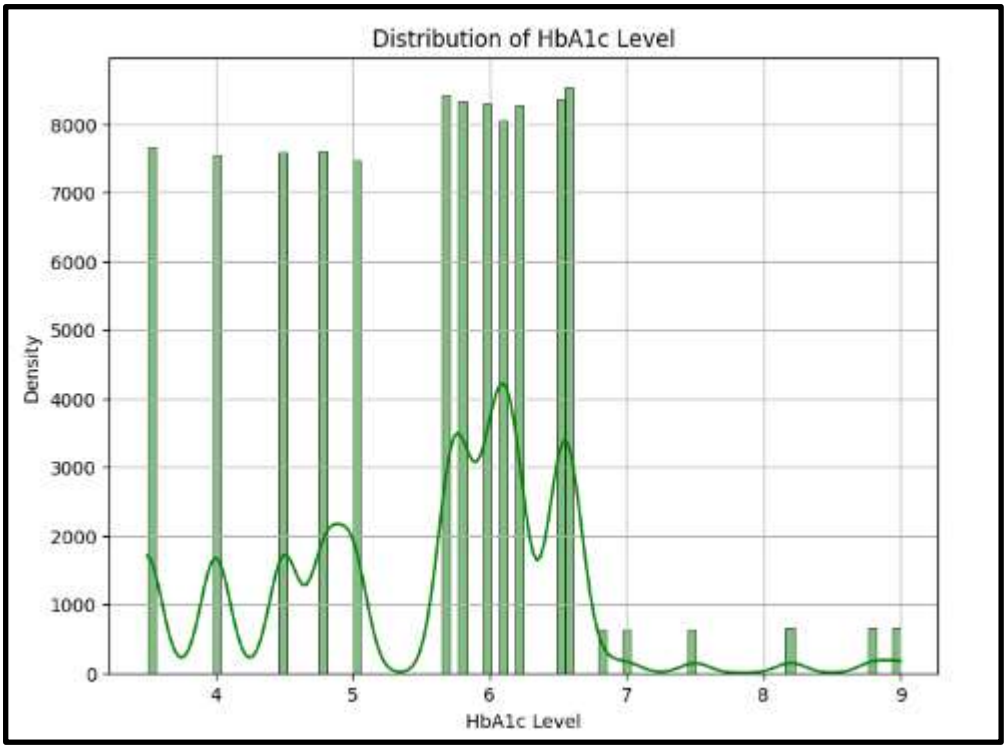


Figure 4.7: Distribution of the HbA1c level

The overall distribution of the HbA1c level is illustrated here. It can be inferred from this mentioned figure that the maximum density of HbA1c level is in the range between 6 and 7. It shows the highest density within the specific interval. This particular range is urgent as it frequently indicates prediabetes or early diabetes, highlighting the significance of checking and overseeing HbA1c levels for preventing disease progression.

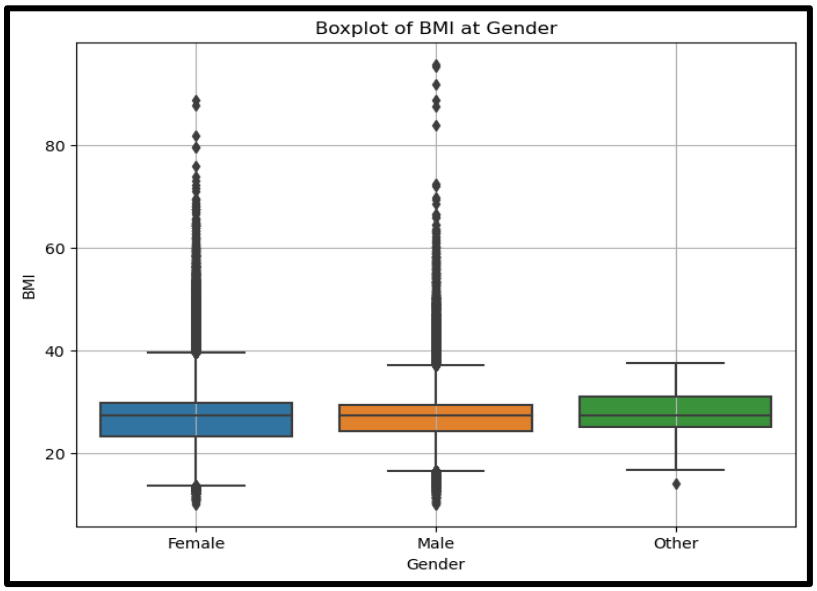


Figure 4.8: Box plot of BMI by gender

This following box plot gives a visual synopsis of the dissemination of the BMI values for various gender groups. This indicates the quartiles, median, and also likely outliers for the

females, males and also potentially other genders. The specific median BMI is addressed by the line within every case, while the actual box shows the "interquartile range (IQR)".

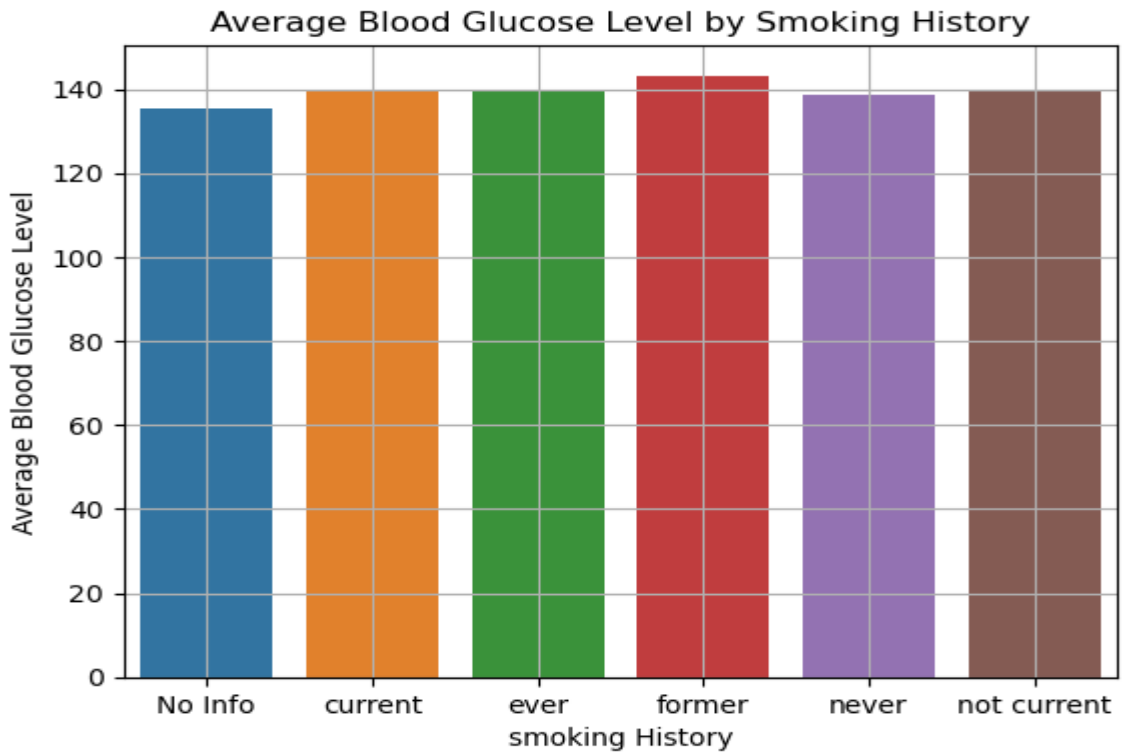


Figure 4.9: Average Blood Glucose Level by the Smoking History

The mentioned figure shows the average blood glucose level by smoking history. The maximum average blood glucose level found in the former type of smoking history is slightly greater than 140. This result is crucial as it recommends a significant connection among the past smoking habits as well as raised blood glucose levels, which is a crucial variable within the diabetes management alongside the risk evaluation.

```
from sklearn.preprocessing import LabelEncoder

#transforming descriptive to numerical columns
data ['gender'] = LabelEncoder().fit_transform(data ['gender'])
data ['smoking_history'] = LabelEncoder().fit_transform(data ['smoking_history'])
```

Figure 4.10: Transforming categorical to numerical variables

In this particular step the descriptive column is transformed into numerical columns using the label encoder. It is an important step toward preparing the specific dataset for the machine learning approaches, which need the numerical input. By changing the categorical variables like the gender, smoking history, or further categorical features into the numerical values, the dataset becomes viable with different approaches.

```
value of Mean Absolute Error (MAE): 30.7237
value of Mean Squared Error (MSE): 1367.3011
R^2 Score: 0.1717
```

Figure 4.11: Linear Regression Result

The specific figure indicates crucial performance measurements for the specific Linear Regression approach. The specific predictions of the model are off by around 30.72 units from the following true values. The particular "Mean Squared Error (MSE)" of about 1367.3011 assesses the average squared discrepancy among predicted alongside the actual qualities, having a lower MSE proposing a superior fit; ,for this situation, the score is moderately high, showing significant prediction errors. The particular R^2 value of 0.1717 uncovers that the particular model assesses only 17.17% of the specific variance within the dataset, highlighting the weak explanatory performance. Overall, the particular approach shows critical prediction errors alongside a restricted capacity for grasping the data fluctuation, proposing the requirement for additional refinement or elective ways for dealing with improved accuracy and prescient performance.

```
Mean Absolute Error (MAE): 30.7237
Mean Squared Error (MSE): 1367.3011
R^2 Score:0.1717
```

Figure 4.12: Result of Support Vector Regression

The particular figure shows the performance measurements for the predictive approach. The specific "Mean Absolute Error (MAE)" of 30.0247 implies that the model's expectations deviate by nearly 30.02 units from the following actual values on the average. The particular "Mean Squared Error (MSE)" of 1367.301 shows the average of the squared contrasts among the predicted and also actual values, with a greater value recommending a poorer fit. The following R^2 value of 0.0866 uncovers that the model assesses 8.66% of the particular variance within the dataset, showing the weak fit and restricted explanatory power.

```
Accuracy of Logistic Regression: 0.9584

Report on classification:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     22850
           1       0.87      0.61      0.72      2150

    accuracy                           0.96     25000
   macro avg       0.92      0.80      0.85     25000
weighted avg       0.96      0.96      0.96     25000


Confusion Matrix:
[[22647   203]
 [  838  1312]]
```

Figure 4.13: Performance Metrics of Logistic regression

The mentioned figure gives an outline of the implemented performance metrics of the Logistic Regression approach. Having the accuracy of 95.84%, the specific model exhibits overall forecast accuracy. The specific classification report uncovers that while Class 0 has the high precision alongside recall, Class 1 shows lower recall, expected difficulties in distinguishing this class.

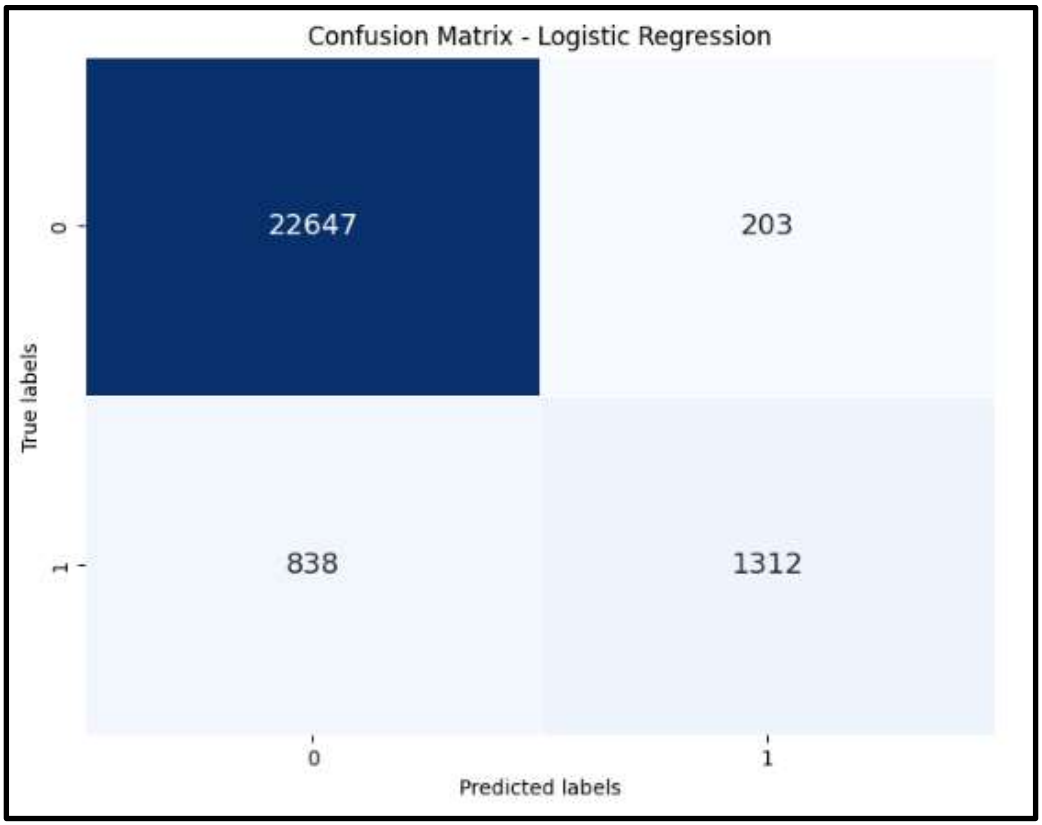


Figure 4.14: Confusion Matrix of Logistic regression

This is the visualization of the implemented confusion matrix of the Logistic regression. The respective confusion matrix indicates 203 False Positives (FP), 22,647 True Negatives (TN), 838 False Negatives (FN), along with 1,312 True Positives (TP). It addresses the correct predictions for the diabetic cases.

```
Random Forest Classifier Accuracy: 0.9700

Report on classification:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     22850
           1       0.96      0.68      0.80      2150

    accuracy                           0.97     25000
   macro avg       0.96      0.84      0.89     25000
weighted avg       0.97      0.97      0.97     25000


Confusion Matrix:
[[22786    64]
 [  687  1463]]
```

Figure 4.15: Performance Metrics of Random Forest Classifier

The specific figure portrays the specific performance measurements for the Random Forest approach. The particular model accomplishes a high accuracy of about 97.00%, accurately classifying most cases. Regardless of the high accuracy, the specific model's lesser recall for the class 1 proposes that further refinement is expected to enhance its awareness towards identifying diabetic cases.

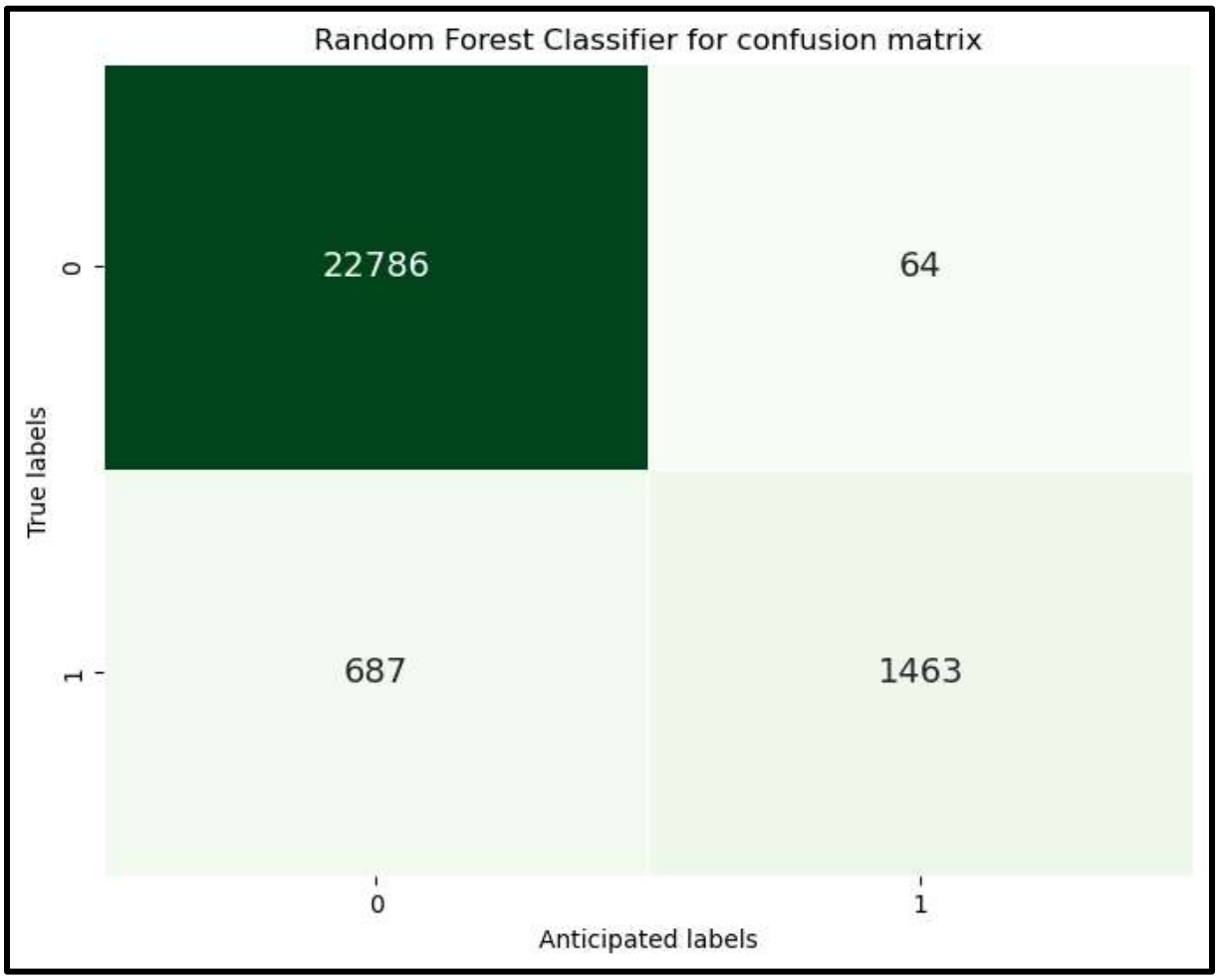


Figure 4.16: Confusion Matrix of Random Forest Classifier

In the following confusion matrix, there are about 22,786 True Negatives (TN) alongside 64 False Positives (FP) (FP), assessing the model's effective performance in distinguishing non-diabetic occasions. However, there are about 687 False Negatives (FN) alongside 1,463 True Positives (TP), demonstrating some trouble in accurately recognizing diabetic cases. The particular confusion matrix of this approach uncovers the model's strengths and also weaknesses in predicting the diabetic alongside non-diabetic cases. With 22,786 True Negatives (TN) along with 64 False Positives (FP), the approach shows greater accuracy in recognizing non-diabetic cases. Though, the presence of 687 False Negatives (FN) along with 1,463 True Positives (TP) shows difficulties in accurately distinguishing diabetic cases.

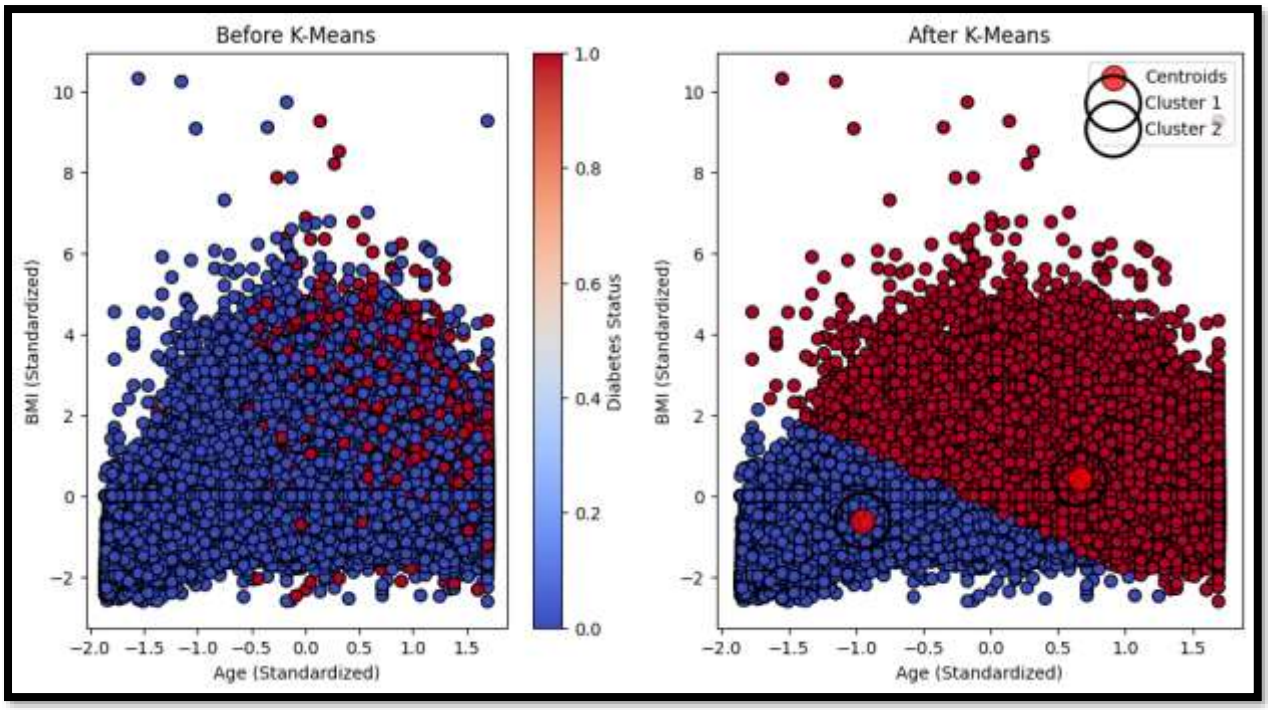


Figure 4.17: Showing the K means Clustering Graph

The visualization depicts the clustering of data points based on standardized age and BMI using the K-Means algorithm.

Before K-Means Clustering

- **Left Plot**:
  - The scatter plot shows the data distribution before K-Means clustering.
  - X-axis: standardized age; Y-axis: standardized BMI.
  - Color indicates diabetes status: blue (non-diabetic) to red (diabetic).
  - No clear separation or clustering of points based on diabetes status.

After K-Means Clustering

- **Right Plot**:
  - The scatter plot displays data after applying K-Means clustering.
  - K-Means algorithm identifies two clusters, depicted by distinct colors.
  - Cluster centroids are marked with larger red dots encircled by black outlines.

Observations

- Post-clustering, data points are grouped into two clusters.
  - Cluster 1: Predominantly blue (non-diabetic individuals).
  - Cluster 2: Predominantly red (diabetic individuals).
- The algorithm effectively distinguishes between individuals based on age, BMI, and diabetes status.

Conclusion

- K-Means clustering has successfully grouped individuals with similar age, BMI, and diabetes characteristics.
- This method aids in identifying patterns in the data and can be useful for predicting diabetes status based on these features.

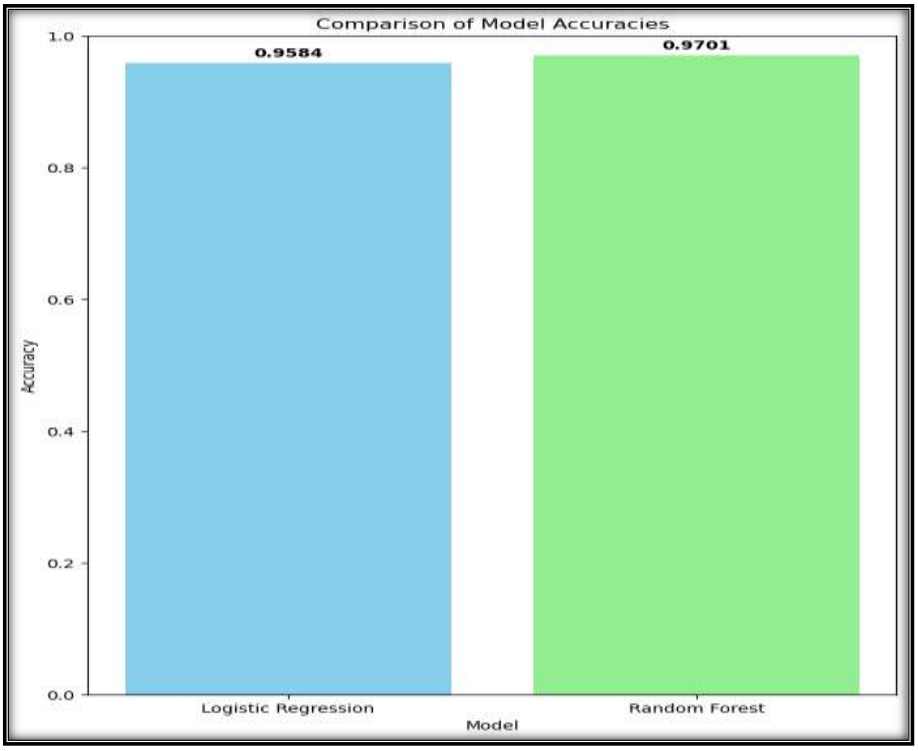


Figure 4.18: Comparison graph of the classification models

The comparison graph of the implemented classification models shows the highest accuracy score identified in the Random Forest model.

Chapter 5: Discussion of Result

5.1 Overview of Findings

The analysis from Chapter 4 shows that various machine learning models are useful in the prediction of diabetes but have their weaknesses. Data contained in the tables included 100, 000 entries, with the qualities like age, gender, hypertension, smoking, the history of the heart diseases, HbA1c, BMI, and blood glucose. Among the novel discoveries, the presence of no null values means that data completeness is quite commendable, which will be important in the subsequent modeling.

5.2 Performance of Predictive Models

The development of a simulation model to test the robustness of predictive models becomes crucial in this process.

The Linear Regression model resulted quite imprecise, presenting a high MSE and a low R^2 value; this means high prediction errors and the ability to explain only minor data variation. The Support Vector Regression is also not appropriate for this data as the high value of MSE combined with the low coefficient of determination of elaboration data indicates.

As for the general performance, Logistic Regression was shown to have a high accuracy (95. 84%), yet it poorly classified samples belonging to Class 1 (diabetic cases), where False Negative count was high. This renders the necessity to increase sensitivity levels for accurate identification of diabetes cases apparent.

With a 97% accuracy level Random Forest Classifier was the best performing model to the other models. 81% accuracy, this method is quite efficacious for classifying the non-diabetic cases. Nevertheless, it performed less effectively on the recall aspect for diabetic cases and this factor was identified as a possible area for improvement. When analyzing the confusion matrix of Random Forest, a high number of true negative instances and fewer false positive instances were obtained, but, at the same time, a significant number of false negative instances which depict the cases of diabetes have been missed out were also identified.

5. 3 Key Findings of Clustering

To further investigate the research question, the secondary analysis using the K-Means Clustering yielded two clusters primarily according to the BMI and age categories. Such clusters which are commonly classified into low and high-risk can go a long way in helping to provide data to better help design further various health patterns based intervention efforts.

5. 4 Implications of Findings and Future Research

Therefore, the choice of a model, as well as model tuning has been highlighted as critical in the prediction of diabetes in this study. However, Random Forest and other models displayed only a moderate performance, thereby revealing the requirements for the model's increased sensitivity and actuality, especially in the identification of cases of diabetes. Future improvements can be made on these models, and more features can be added to these models and very efficient methodologies such as deep learning should be employed to obtain better performance.

The work discusses the possible use of machine learning models in the early detection of diabetes and identifies the model's strengths, limitations, and future development needs.

Chapter 6: Conclusion

6.1 Interpretation of the findings

This exhaustive evaluation of the particular diabetes dataset, involving the data collection, preprocessing, EDA, model training, assessment, and also clustering evaluation, gives significant details into the health designs and also prescient variables related with the diabetes. The underlying advances guaranteed a perfect and very much organized dataset, changing the categorical factors into the numerical forms and envisioning crucial distributions like the gender, age, and also HbA1c levels. Through the model preparation and also evaluation, the developed logistic regression along with random forest approaches exhibited changing levels of accuracy , with the following logistic regression accomplishing the accuracy of 95.84% alongside random forest accomplishing 97.00%. In spite of high accuracy, the following two models showed difficulties in foreseeing the minority class (diabetic cases), as shown by their particular confusion matrices and also order reports. This underlines the requirement for additional refinement to improve prescient power and decrease the false negatives, which are crucial within a medical services setting. The particular K-Means cluster evaluation gave extra layers of understanding by portioning the populace in view of BMI and age. The following Cluster 1 blue included individuals with the lower BMI and also age, the respective cluster 2 red addressed demonstrating a higher risk for the diabetes entanglements.

6.2 Limitation of the research

While this specific exploration gives significant details into diabetes prediction along with the management, this isn't without the limitations. One essential limit is the intrinsic bias within the particular dataset, which may not be representative of the more extensive populace (Qiao *et al.* 2020). The following dataset's demographic and also clinical features might impact the overall model's generalizability, restricting its appropriateness to various populations with several genetic, lifestyle, along with the environmental variables. Also, the following data preprocessing steps, for example, changing the categorical factors into the respective numerical values utilizing the label encoding, could oversimplify the composite connections among specific categorical variables and also diabetes risk. The particular models trained within this analysis, involving the logistic regression as well as random forest, displayed high accuracy however struggled with the prediction of the minority class of the diabetic cases, prompting higher false-negative rates (Gupta *et al.* 2022). It shows a requirement for additional modern approaches or the balanced datasets to further develop the prediction of the minority class. One more impediment is the dependence on the BMI along with age for the clustering evaluation, which might neglect other crucial elements impacting the overall diabetes risk, like the physical activity, diet, alongside genetic inclinations. Besides, the following clustering evaluation utilizing K-Means, while viable in recognizing particular subgroups, expects clusters are spherical alongside similarly measured, which could not precisely assess the true dissemination of the following population. Moreover, the following research didn't evaluate the following temporal parts of the diabetes progression, passing up adequate patterns along with trends after some time. Ultimately, the evaluation findings are constrained by the static characteristic of the specific dataset and don't represent longitudinal data that could give further experiences into the development and also the management of the diabetes (Shahriare Satu *et al.* 2020).

6.3 Recommendation for future

For the future evaluation within diabetes prediction alongside the management, various recommendations may improve the effectiveness and also relevance of the findings. Initially, extending the dataset to incorporate a more different and delegate sample will enhance the generalizability of the particular approaches, tending to expected biases and also better grasping the fluctuation within the diabetes risk throughout various populations. Consolidating

further elements, for example, genetic markers, lifestyle variables, and also dietary propensities, could give a more far reaching comprehension of diabetes risk and enhance model performance (Dong *et al.* 2022). Improved modeling approaches demonstrating strategies, like the ensemble techniques or the deep learning methods, ought to be evaluated to more readily deal with the class imbalances along with improving the prediction accuracy, especially for minority classes. Besides, incorporating longitudinal details to follow changes within the health metrics over the long period may offer details into the overall progression of the diabetes and also the viability of the interventions. Incorporating the predictive evaluation with the continuous data from the wearable equipment could likewise prompt more customized and also convenient administration procedures (Alfian *et al.* 2020).

Reference :

Afsaneh, E., Sharifdini, A., Ghazzaghi, H. and Ghobadi, M.Z., 2022. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetology & Metabolic Syndrome, 14(1), p.196. https://link.springer.com/content/pdf/10.1186/s13098-022-00969-9.pdf

Albahli, S., 2020. Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. Journal of Medical Imaging and Health Informatics, 10(5), pp.1069-1075. https://www.researchgate.net/profile/Saleh-Albahli/publication/341082446_Type_2_Machine_Learning_An_Effective_Hybrid_Prediction_Model_for_Early_Type_2_Diabetes_Detection/links/5f4801d5299bf13c50428816/Type-2-Machine-Learning-An-Effective-Hybrid-Prediction-Model-for-Early-Type-2-Diabetes-Detection.pdf

Alfian, G., Syafrudin, M., Fitriyani, N.L., Anshari, M., Stasa, P., Svub, J. and Rhee, J., 2020. Deep neural network for predicting diabetic retinopathy from risk factors. Mathematics, 8(9), p.1620. https://www.mdpi.com/2227-7390/8/9/1620/pdf

Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare applications. Journal of healthcare engineering, 2021(1), p.9930985. https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/9930985

Chen, Z. and Liu, W., 2020. An efficient parameter adaptive support vector regression using K-means clustering and chaotic slime mould algorithm. *Ieee Access*, *8*, pp.156851-156862.

Chou, C.Y., Hsu, D.Y. and Chou, C.H., 2023. Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), p.406. https://www.mdpi.com/2075-4426/13/3/406/pdf

Daghistani, T. and Alshammari, R., 2020. Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. Journal of Advances in Information Technology Vol, 11(2), pp.78-83. https://www.jait.us/uploadfile/2020/0417/20200417070739281.pdf

Das, D., Biswas, S.K. and Bandyopadhyay, S., 2022. A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning. Multimedia Tools and Applications, 81(18), pp.25613-25655. https://link.springer.com/content/pdf/10.1007/s11042-022-12642-4.pdf

Dong, Z., Wang, Q., Ke, Y., Zhang, W., Hong, Q., Liu, C., Liu, X., Yang, J., Xi, Y., Shi, J. and Zhang, L., 2022. Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. Journal of translational medicine, 20(1), p.143. https://link.springer.com/content/pdf/10.1186/s12967-022-03339-1.pdf

Gupta, H., Varshney, H., Sharma, T.K., Pachauri, N. and Verma, O.P., 2022. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex & Intelligent Systems, 8(4), pp.3073-3087.https://link.springer.com/content/pdf/10.1007/s40747-021-00398-7.pdf

Hasan, D.A., Zeebaree, S.R., Sadeeq, M.A., Shukur, H.M., Zebari, R.R. and Alkhayyat, A.H., 2021, April. Machine learning-based diabetic retinopathy early detection and classification systems-a survey. In 2021 1st Babylon International Conference on Information Technology and Science (BICITS) (pp. 16-21). IEEE. https://www.researchgate.net/profile/Dathar-Hasan/publication/353860904_Machine_Learning-based_Diabetic_Retinopathy_Early_Detection_and_Classification_Systems-_A_Survey/links/6116f6b21ca20f6f861e5b3f/Machine-Learning-based-Diabetic-Retinopathy-Early-Detection-and-Classification-Systems-A-Survey.pdf

Hassan, M.M., Peya, Z.J., Mollick, S., Billah, M.A.M., Shakil, M.M.H. and Dulla, A.U., 2021, July. Diabetes prediction in healthcare at early stage using machine learning approach. In 2021

12th International conference on computing communication and networking technologies (ICCCNT) (pp. 01-05). IEEE. https://www.researchgate.net/profile/Swarnali-Mollick-2/publication/355899609_Diabetes_Prediction_in_Healthcare_at_Early_Stage_Using_Machine_Learning_Approach/links/61b61273fd2cbd7200965288/Diabetes-Prediction-in-Healthcare-at-Early-Stage-Using-Machine-Learning-Approach.pdf

Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M. and Kabir, M.N., 2020. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*, pp.1-14.

Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 2021 May;77(5):5198-219.

Jaiswal, V., Negi, A. and Pal, T., 2021. A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), pp.435-443. https://ir.vignan.ac.in/705/1/25-21.pdf

Joshi, R.D. and Dhakal, C.K., 2021. Predicting type 2 diabetes using logistic regression and machine learning approaches. *International journal of environmental research and public health*, *18*(14), p.7346.

Lynam, A.L., Dennis, J.M., Owen, K.R., Oram, R.A., Jones, A.G., Shields, B.M. and Ferrat, L.A., 2020. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagnostic and prognostic research*, *4*, pp.1-10.

Muneer, A. and Fati, S.M., 2020. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, *12*(11), p.187.

Nahzat, S. and Yağanoğlu, M., 2021. Diabetes prediction using machine learning classification algorithms. Avrupa Bilim ve Teknoloji Dergisi, (24), pp.53-59. https://dergipark.org.tr/en/download/article-file/1648927

Pranto, B., Mehnaz, S.M., Mahid, E.B., Sadman, I.M., Rahman, A. and Momen, S., 2020. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information*, *11*(8), p.374. Available at: https://www.mdpi.com/2078-2489/11/8/374

Qiao, L., Zhu, Y. and Zhou, H., 2020. Diabetic retinopathy detection using prognosis of microaneurysm and early diagnosis system for non-proliferative diabetic retinopathy based on deep learning algorithms. IEEE Access, 8, pp.104292-104302. https://ieeexplore.ieee.org/iel7/6287639/8948470/09091167.pdf

Shahriare Satu, M., Atik, S.T. and Moni, M.A., 2020. A novel hybrid machine learning model to predict diabetes mellitus. In Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2019 (pp. 453-465). Springer Singapore. https://www.researchgate.net/profile/Md-Satu/publication/335727823_A_Novel_Hybrid_Machine_Learning_Model_To_Predict_Diabetes_Mellitus/links/5d77ee434585151ee4adeb1d/A-Novel-Hybrid-Machine-Learning-Model-To-Predict-Diabetes-Mellitus.pdf

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P. and Homayouni, S., 2020. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, pp.6308-6325.

Suresh, K., Obulesu, O. and Ramudu, B.V., 2020. Diabetes prediction using machine learning techniques. Helix-The Scientific Explorer| Peer Reviewed Bimonthly International Journal, 10(02), pp.136-142. https://helixscientific.pub/index.php/Home/article/download/123/123

Zhang, Y., Wang, Q., Chen, X., Yan, Y., Yang, R., Liu, Z. and Fu, J., 2022. The prediction of spark-ignition engine performance and emissions based on the SVR algorithm. *Processes*, *10*(2), p.312. Available at: https://www.mdpi.com/2227-9717/10/2/312

Appendix:

```
# importing required modules
import pandas as ps
import numpy as ny
import matplotlib.pyplot as plot
import seaborn as sns
data=ps.read_csv(r"D:\WORK\Diabetic-prediction-using-multiple-machine-learning-algorithms\diabetes_prediction_dataset.csv")

# top 5 rows of the dataset

data.head()
```

	gender	**age**	**hypertension**	**heart_disease**	**smoking_history**	**bmi**	**HbA1c_level**	**blood_glucose_level**	**diabetes**
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

```
# checking null values
data.isnull()
```

	gender	**age**	**hypertension**	**heart_disease**	**smoking_history**	**bmi**	**HbA1c_level**	**blood_glucose_level**	**diabetes**
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
99995	False	False	False	False	False	False	False	False	False
99996	False	False	False	False	False	False	False	False	False
99997	False	False	False	False	False	False	False	False	False
99998	False	False	False	False	False	False	False	False	False
99999	False	False	False	False	False	False	False	False	False

# describe data set
data.describe()

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000

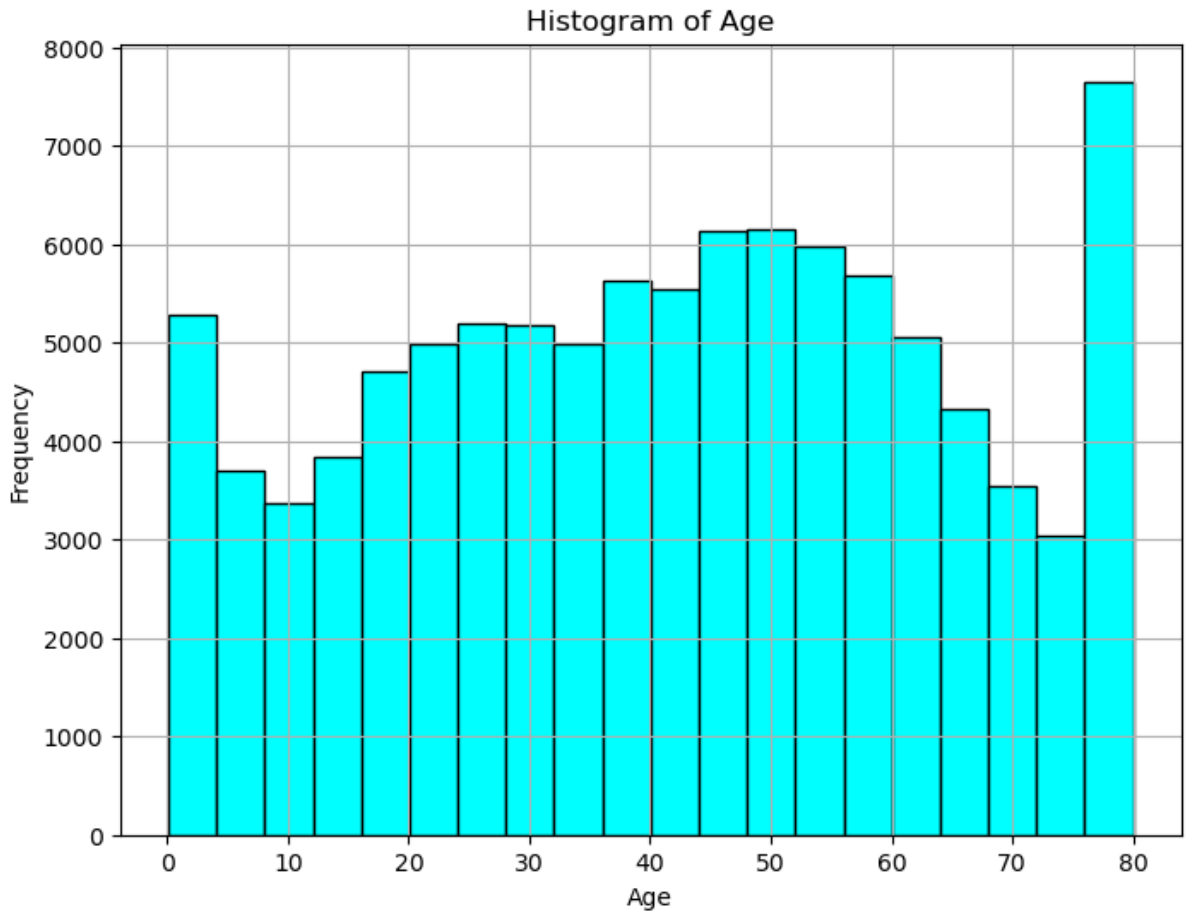
	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

data.shape

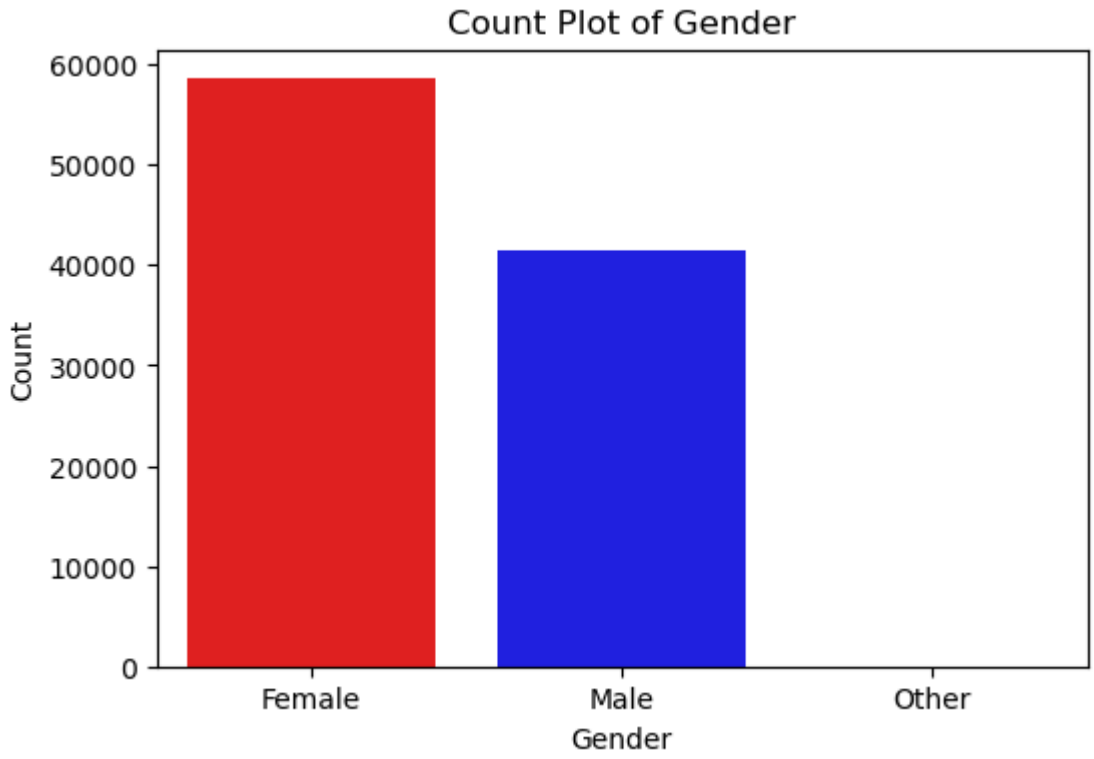
(100000, 9)

Exploratory Data Analysis (EDA)

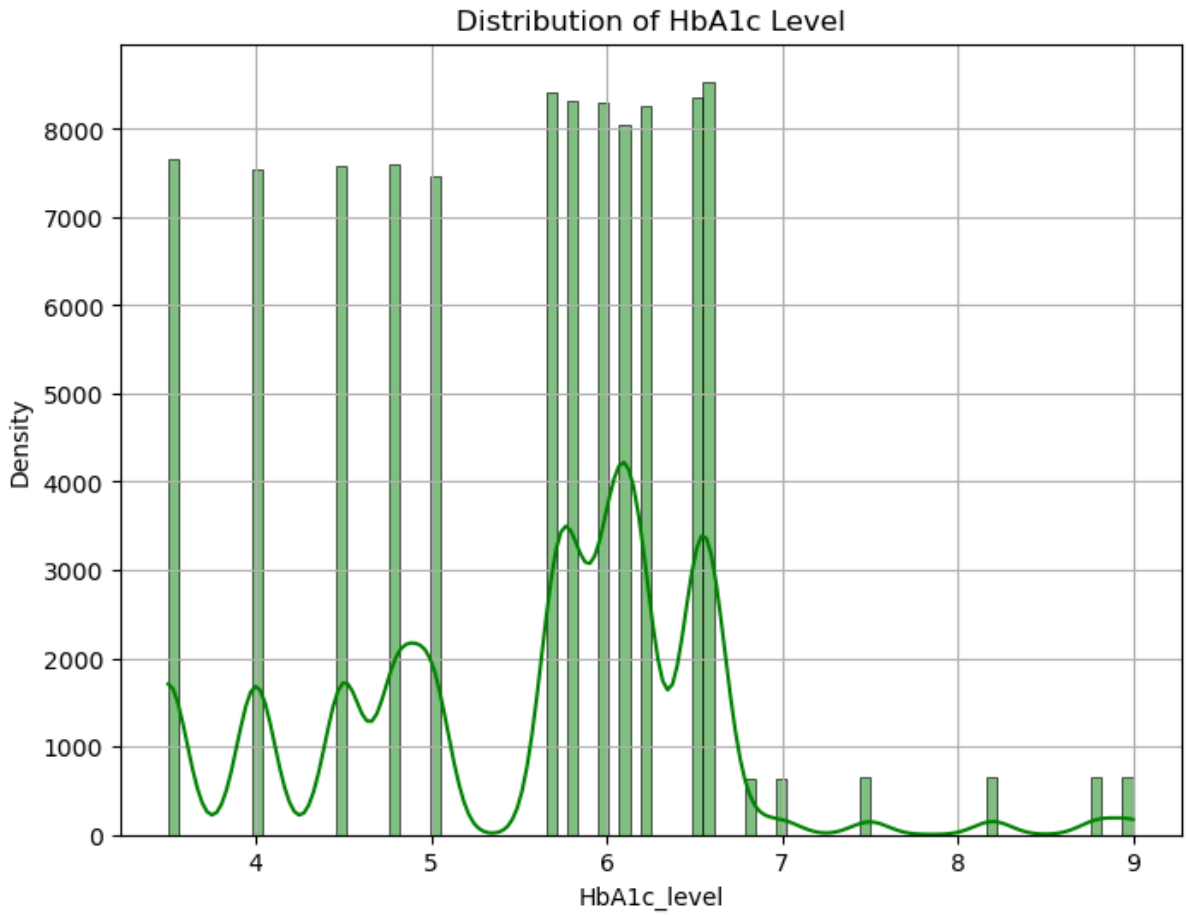
```
plot.figure(figsize=(8,6))
plot.hist(data['age'], bins=20, edgecolor="black", color="cyan")
plot.title('Histogram of Age')
plot.xlabel('Age')
plot.ylabel('Frequency')
plot.grid(True)
plot.show()
```



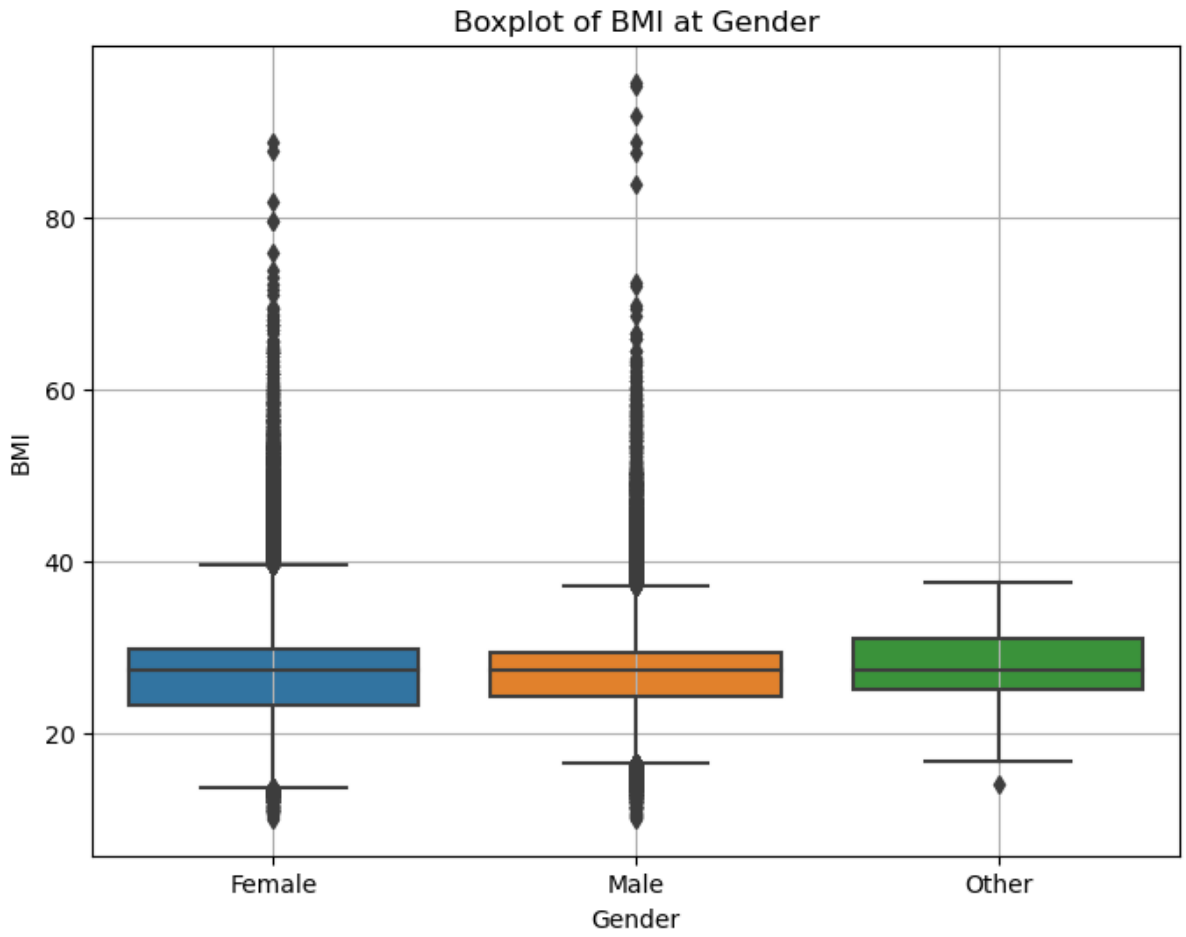

```
# Defining colors to every gender catagory
colors = {'Male': 'blue', 'Female': 'red', 'Other': 'green'}

plot.figure(figsize=(6, 4))
sns.countplot(x='gender', data=data, palette=colors)
plot.title('Count Plot of Gender')
plot.xlabel('Gender')
plot.ylabel('Count')
plot.show()
```



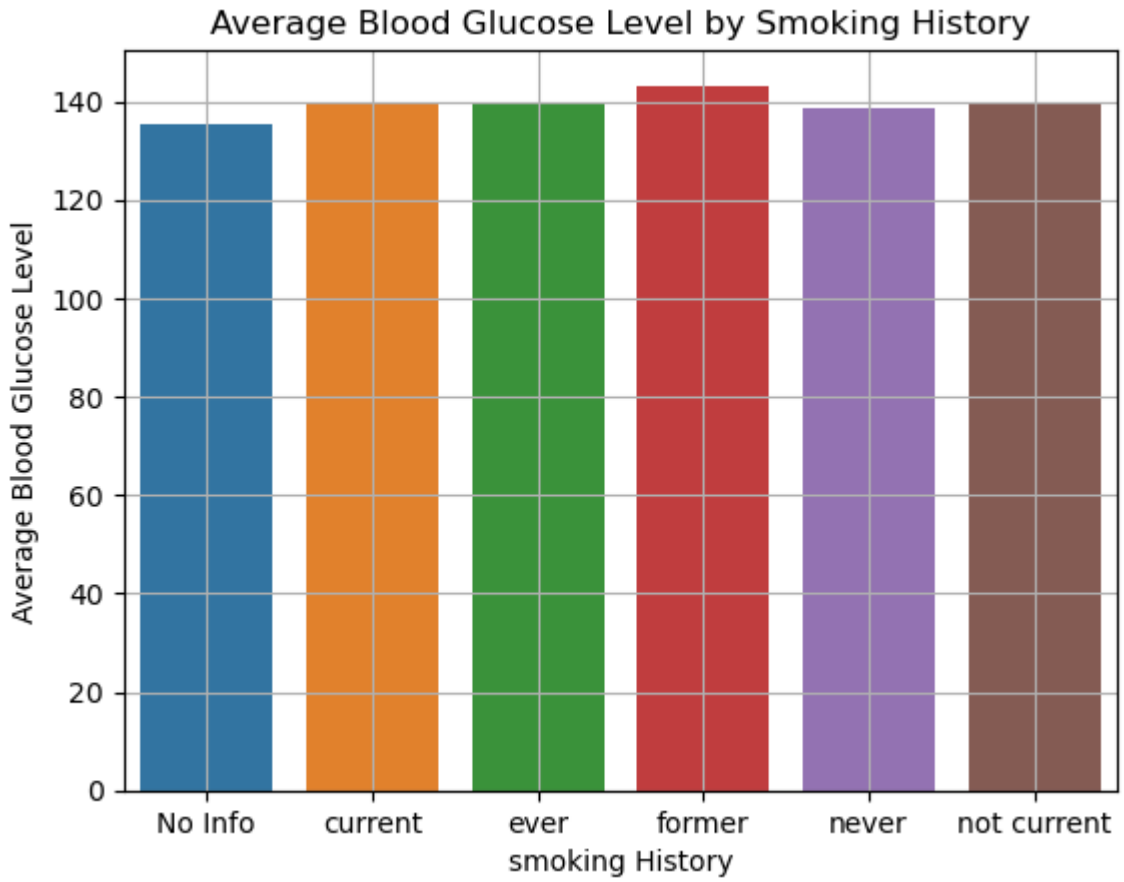

```
plot.figure(figsize=(8, 6))
sns.histplot(data['HbA1c_level'],kde=True, color= 'green')
plot.title('Distribution of HbA1c Level')
plot.xlabel('HbA1c_level')
plot.ylabel('Density')
plot.grid(True)
plot.show()
```




```
plot.figure(figsize=(8, 6))
sns.boxplot(x='gender', y= 'bmi', data=data)
plot.title('Boxplot of BMI at Gender')
plot.xlabel('Gender')
plot.ylabel('BMI')
plot.grid(True)
plot.show
```




```
# Defineing color palatte for several smoking history categories
avg_glucose_by_smoking                                                    =
data.groupby('smoking_history')['blood_glucose_level'].mean().reset_index()
sns.barplot(x='smoking_history', y='blood_glucose_level', data=avg_glucose_by_smoking)
plot.title('Average Blood Glucose Level by Smoking History')
plot.xlabel('smoking History')
plot.ylabel('Average Blood Glucose Range')
plot.grid(True)
plot.show()
```



data.head()

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Linear Regression

```
from sklearn.preprocessing import LabelEncoder

#transforming descriptive to numerical columns
data ['gender'] = LabelEncoder().fit_transform(data ['gender'])
data ['smoking_history'] = LabelEncoder().fit_transform(data ['smoking_history'])
```

data.head()

	gender	**age**	**hypertension**	**heart_disease**	**smoking_history**	**bmi**	**HbA1c_level**	**blood_glucose_level**	**diabetes**
0	0	80.0	0	1	4	25.19	6.6	140	0
1	0	54.0	0	0	0	27.32	6.6	80	0
2	1	28.0	0	0	4	27.32	5.7	158	0
3	0	36.0	0	0	1	23.45	5.0	155	0
4	1	76.0	1	1	1	20.14	4.8	155	0

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler

# separating characteristics and target variable
char = data.drop('blood_glucose_level', axis=1) # characteristics
target_var = data['blood_glucose_level'] # target variable

#splitting the dataset into train dataset and testing datasets
x_train, x_test, y_train, y_test = train_test_split(char, target_var, test_size=0.3, random_state=42)

# standardizing featurs
scaler = StandardScaler()
x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)
```

```
# linear regresion model training
model = LinearRegression()
model.fit(x_train_scaled, y_train)

# predictions on the test set
y_pred = model.predict(x_test_scaled)

# Linear regression model evaluation
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'value of Mean Absolute Error (MAE): {mae:.4f}')
print(f'value of Mean Squared Error (MSE): {mse:.4f}')
print(f'R^2 Score: {r2:.4f}')

value of Mean Absolute Error (MAE): 30.7237
value of Mean Squared Error (MSE): 1367.3011
R^2 Score: 0.1717
```

Support Vector Regression (SVR)

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from sklearn.metrics import mean_absolute_error, mean_squared_error,r2_score

# seperating characteristics and target variable
char = data.drop('blood_glucose_level', axis=1) #characteristics
target_var = data['blood_glucose_level'] #target variable

#splitting the dataset into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(char, target_var, test_size=0.3, random_state=42)

# standardizing features (recommended for SVR)
scalar = StandardScaler()
x_train_scale = scaler.fit_transform(x_train)
x_test_scale = scaler.transform(x_test)

# Training Support vector regression model
svrmodel = SVR(kernel='rbf')
model.fit(x_train_scale, y_train)

# making predictions on testdata set
y_pred = model.predict(x_test_scaled)

# Evaluating the svr model
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
```

```
r2 = r2_score(y_test, y_pred)

print(f'Mean Absolute Error (MAE): {mae:.4f}')
print(f'Mean Squared Error (MSE): {mse:.4f}')
print(f'R^2 Score:{r2:.4f}')

Mean Absolute Error (MAE): 30.0247
Mean Squared Error (MSE): 1367.3011
R^2 Score:0.1717
```

Logestic regression

```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Separating characteristics  and target variable
char = data.drop('diabetes', axis=1)  # characteristics
target_var = data['diabetes']  # Target variable

# Splitting dataset into training dataset and testing dataset
X_train, X_test, y_train, y_test = train_test_split(char, target_var, test_size=0.25, random_state=42)

# Standardizing  features (recommended for Logistic Regression)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Logistic Regression model training
model = LogisticRegression(max_iter=1000)
model.fit(X_train_scaled, y_train)

# predictions on testing set
y_pred = model.predict(X_test_scaled)

# Logestic regression model evaluating
accuracy_lr = accuracy_score(y_test, y_pred)
print(f'Accuracy of Logistic Regression: {accuracy_lr:.4f}')

# classification report
print('\nReport on classification:')
print(classification_report(y_test, y_pred))

# confusion matrix
print('\nConfusion Matrix:')
print(confusion_matrix(y_test, y_pred))

Accuracy of Logistic Regression: 0.9584
```

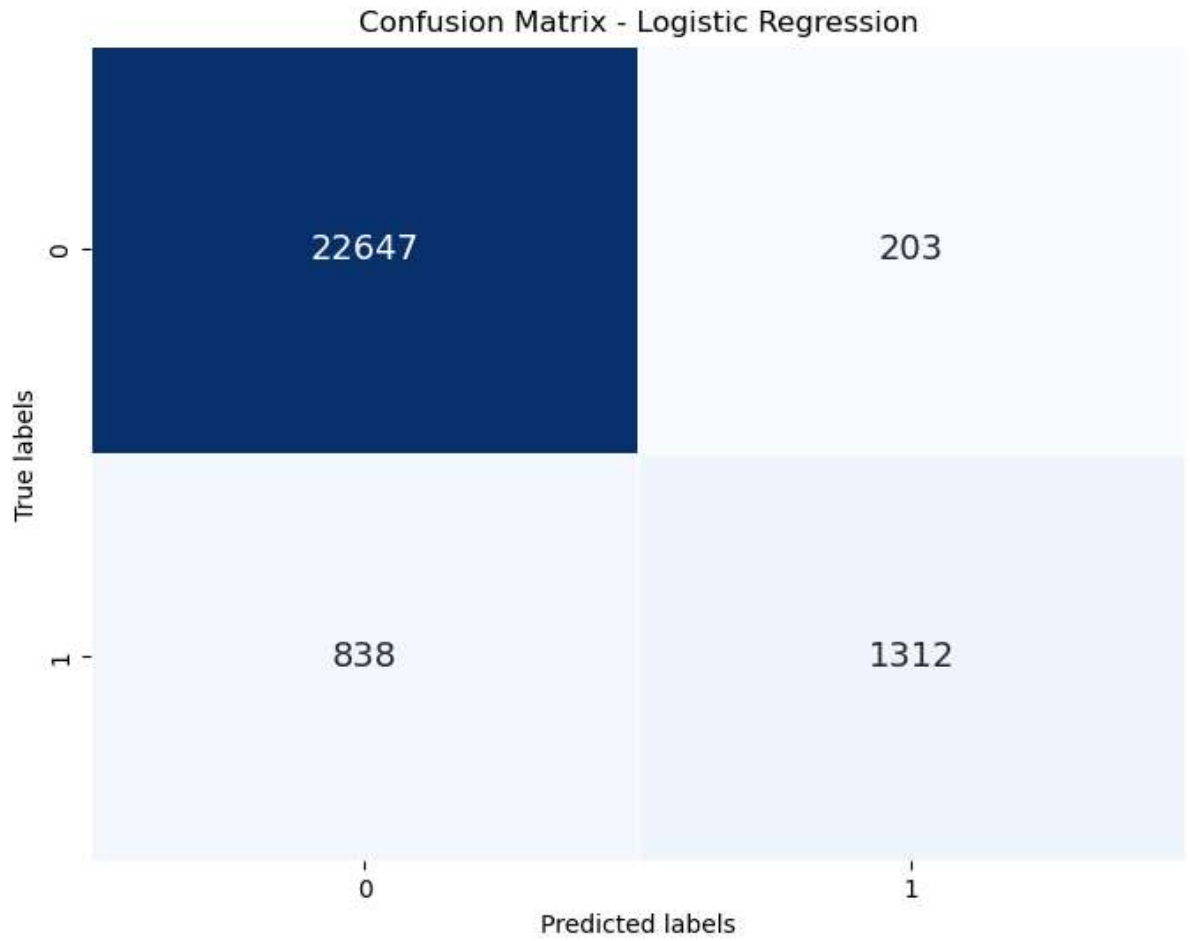
Report on Classification:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	22850
1	0.87	0.61	0.72	2150
accuracy			0.96	25000
macro avg	0.92	0.80	0.85	25000
weighted avg	0.96	0.96	0.96	25000

Confusion Matrix:

```
[[22647   203]
 [  838  1312]]
```

```
from sklearn.metrics import confusion_matrix

# Computing the confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Plotting confusion matrix heatmap
plot.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False,
            annot_kws={'fontsize': 14}, linewidths=0.5)
plot.title('Confusion Matrix - Logistic Regression')
plot.xlabel('Predicted labels')
plot.ylabel('True labels')
plot.show()
```



Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Separating features (X) and target variable (y)
features = data.drop('diabetes', axis=1)  # Features
target_var = data['diabetes']  # Target variable

# dividing the dataset into training set and testing set
X_train, X_test, y_train, y_test = train_test_split(features, target_var, test_size=0.25, random_state=42)

# Model training for the Random Forest Classifier
model = RandomForestClassifier(random_state=50)
model.fit(X_train, y_train)

# Making predictions on testing set
y_pred = model.predict(X_test)

# Model Evaluation
accuracy_rfc = accuracy_score(y_test, y_pred)
print(f'Random Forest Classifier Accuracy: {accuracy_rfc:.4f}')
```

```
# Printing classification report
print('\nClassification Report:')
print(classification_report(y_test, y_pred))

# Printing confusion matrix
print('\nConfusion Matrix:')
print(confusion_matrix(y_test, y_pred))
```

Random Forest Classifier Accuracy: 0.9700

Classification Report:

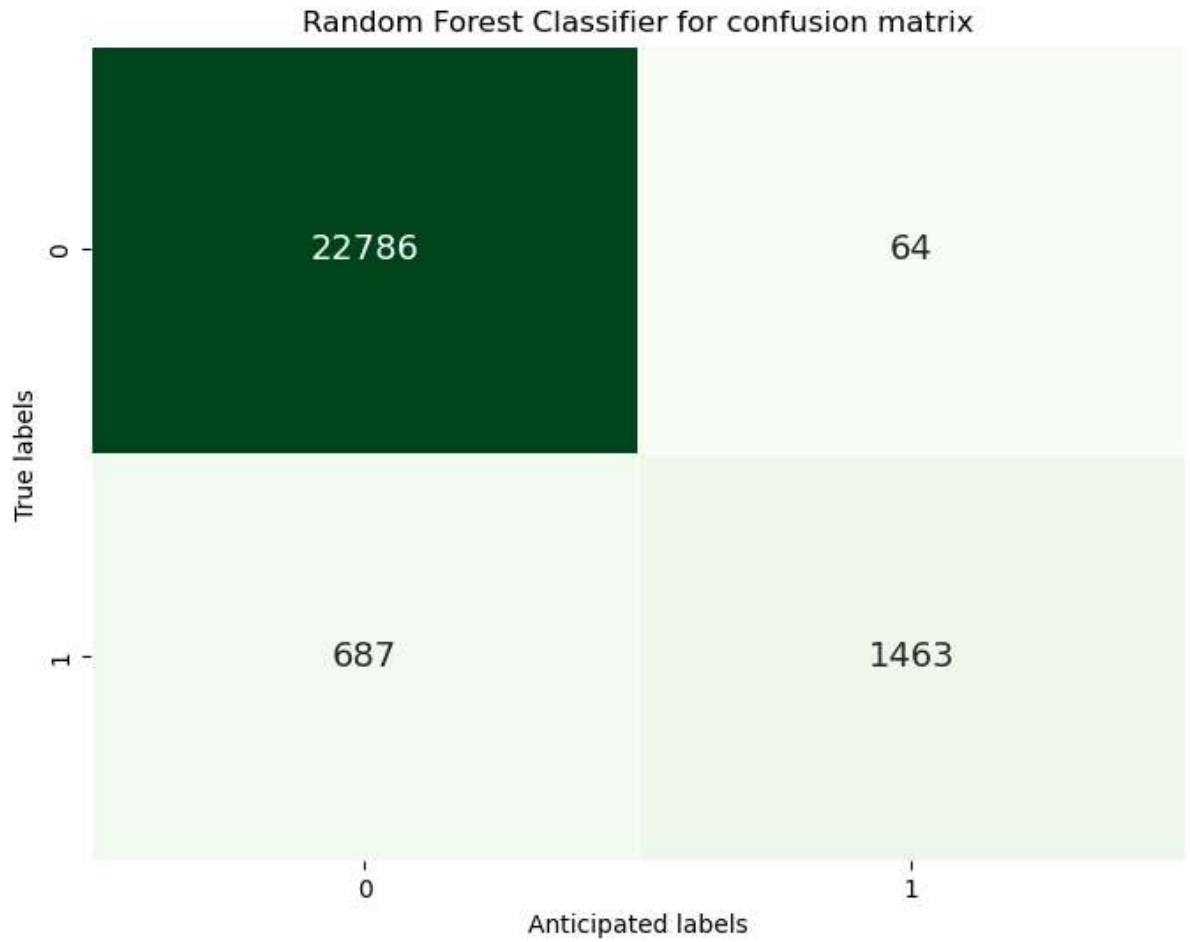
	precision	recall	f1-score	support
0	0.97	1.00	0.98	22850
1	0.96	0.68	0.80	2150
accuracy			0.97	25000
macro avg	0.96	0.84	0.89	25000
weighted avg	0.97	0.97	0.97	25000

Confusion Matrix:

[[22786 64]

[687 1463]]

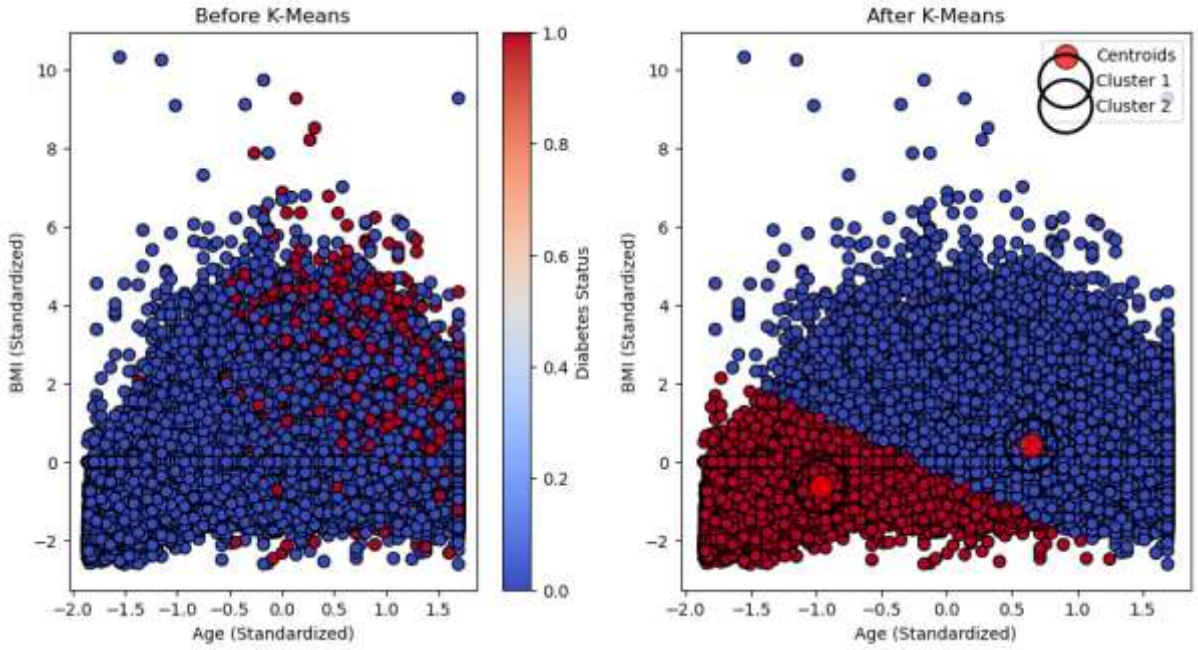
```
# Computing confusion matrix
cm_r = confusion_matrix(y_test, y_pred)

# Plotting confusion matrix heatmap for Random Forest Classifier
plot.figure(figsize=(8, 6))
sns.heatmap(cm_rf, annot=True, fmt='d', cmap='Greens', cbar=False,
            annot_kws={'fontsize': 14}, linewidths=0.5)
plot.title('Random Forest Classifier for confusion matrix')
plot.xlabel('Anticipated labels ')
plot.ylabel('True labels')
plot.show()
```



K-means Clustering

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Selecting the features for X and Y axes
X = data[['age', 'bmi']]  # Features: 'age' for X-axis, 'bmi' for Y-axis
y = data['diabetes']  # Target: 'diabetes'

# Standardizing the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Before K-Means
plot.figure(figsize=(12, 6))
plot.subplot(1, 2, 1)
plot.scatter(X_scaled[:, 0], X_scaled[:, 1], s=50, c=y, cmap='coolwarm', edgecolor='k')
plot.title('Before K-Means')
plot.xlabel('Age (Standardized)')
plot.ylabel('BMI (Standardized)')
plot.colorbar(label='Diabetes Status')

# Apply K-Means
```

```
kmeans = KMeans(n_clusters=2, random_state=0)  # We expect 2 clusters: diabetic and non-diabetic
kmeans.fit(X_scaled)
y_kmeans = kmeans.predict(X_scaled)

# After K-Means
plot.subplot(1, 2, 2)
plot.scatter(X_scaled[:, 0], X_scaled[:, 1], c=y_kmeans, s=50, cmap='coolwarm', edgecolor='k')
centers = kmeans.cluster_centers_
plot.scatter(centers[:, 0], centers[:, 1], c='red', s=200, alpha=0.75, edgecolor='k', label='Centroids')

# Draw circles around clusters to highlight them
for i, center in enumerate(centers):
        plot.scatter(center[0], center[1], s=1000, facecolors='none', edgecolors='black', linewidths=2, label=f'Cluster {i+1}')
plot.title('After K-Means')
plot.xlabel('Age (Standardized)')
plot.ylabel('BMI (Standardized)')
plot.legend()
plot.show()
```




```
# Comparing accuracies of respective models
model = ['Logistic Regression', 'Random Forest']
accuracie = [accuracy_lr, accuracy_rfc]

# Plotting the accuracies
import matplotlib.pyplot as plot

plot.figure(figsize=(8, 10))
plot.bar(model, accuracies, color=['skyblue', 'lightgreen'])
plot.title('Comparison of Model Accuracies')
```

```
plot.xlabel('Model')
plot.ylabel('Accuracy')
plot.ylim(0, 1)

for i, v in enumerate(accuracies):
    plot.text(i, v+0.01 , f'{v:.4f}', ha='right', fontweight='bold')

plot.show()
```