DIABETIC PREDICTION USING MULTIPLE MACHINE LEARNING ALGORITHMS

Acknowledgments

I would like to personally thank all those who helped in the accomplishment of this research. 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

The purpose of this research is to examine the 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. Also, K-Means Clustering is applied on the data to find the pattern as well as subgrouping present in the dataset.

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.

Hypothesis testing shows that the clustering of those participants based on their BMI and blood glucose will split them into three different risk zones for diabetes. These findings recommend that whilst current types are apt for general prediction, they need to be tuned to obtain higher sensitivity, especially for detection of diabetic cases. Further studies should look into analyzing more enhanced approaches such as Deep Learning, and it should also seek to integrate more variables with the aim of refining the prediction power. It can be concluded that the use of the models for proper selection and tuning is useful in early identification and control of 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 inability of the body to maintain proper blood glucose levels. 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. Hoping that this research of diabetes prediction based on these algorithms deems fruitful to help enhance the efficient ratio of diabetes prediction and improvement and change of health standards, it is concluded that this type of research is helpful in the global sense.

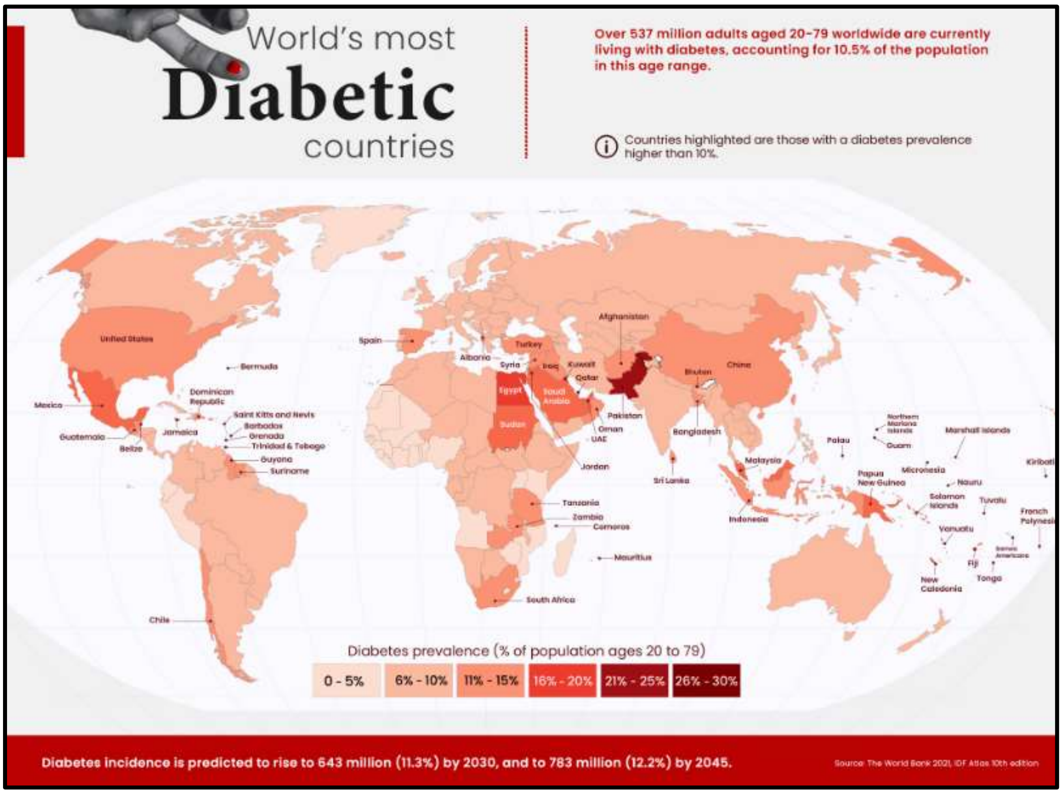


Figure 1.1.1: World's Most Diabetic Country

(Source: visualcapitalist.com, 2023)

1.2 Aim and objectives

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.

Objectives

- 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 Research background

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", "Random Forest Classifier", "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. By analyzing the adequate predictive models, the study assesses effective relationships among demographics along with the medical components, thereby impacting a deeper comprehension of the diabetes risk profiles alongside incorporating the tailored interventions for the at-risk populations.

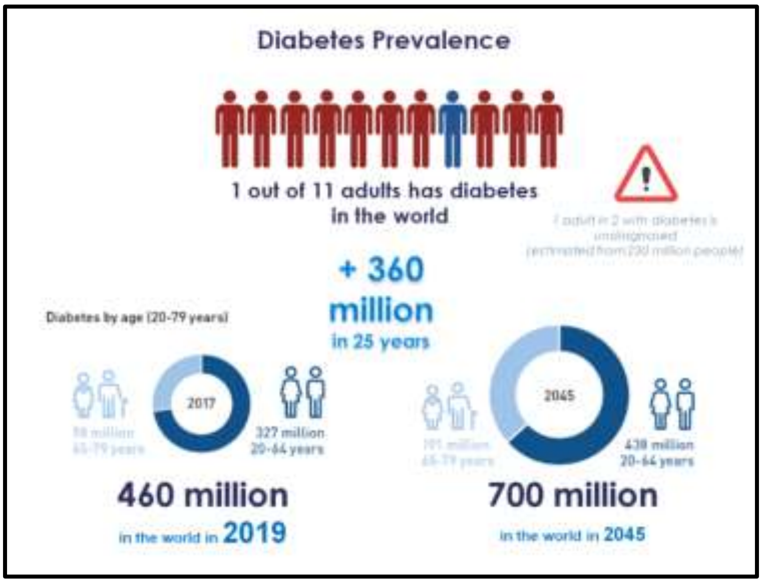


Figure 1.4.1: Diabetes Prevalence

(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). This study will uphold the advancement of greater diagnostic devices alongside preventive techniques, adding to enhanced patient results.

1.6 Research Structure

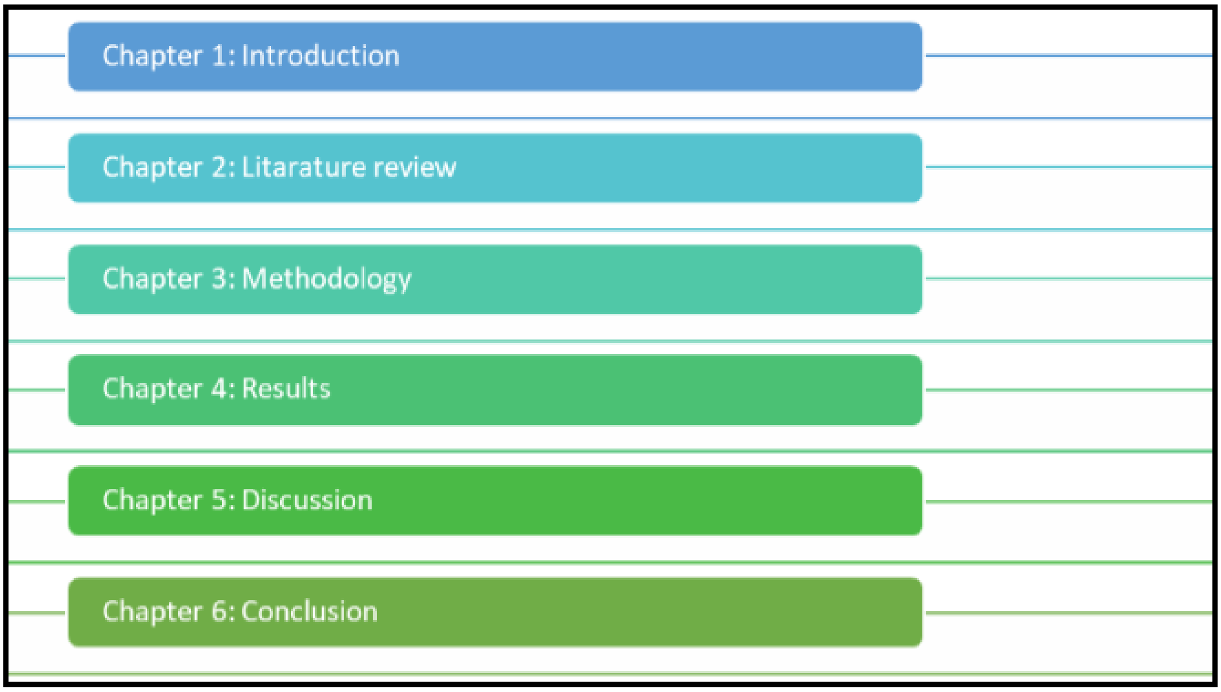


Figure: Research Structure

(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

The following literature canters around the utilization of the machine learning strategies within the prediction and diagnosis of diabetes. Broad research has been directed on using different approaches to improve the accuracy and also productivity of diabetes identification. The respective Machine learning approaches, like the linear regression, logistic regression, Random Forest Classifier, and so on have been broadly evaluated for the significance within the medical diagnostics. Research has demonstrated that logistic regression has been successful within binary classification operations, especially in recognizing diabetic and non-diabetic patients. The utilization of the linear regression for anticipating the continuous factors, for example, blood glucose levels, has been factual. Moreover, the Random Forest has been featured for its capacity to deal with huge datasets and give high predictive accuracy. The particular clustering approaches like K-Means clustering approaches have empowered the identification of crucial indicators of diabetes and assessed with the improvement of customized treatment plans. Ethical contemplations, like data anonymization and consistency with the GDPR, have been assessed for assuring the patient security alongside the data protection. The incorporation of machine learning within diabetes prediction has indicated promising outcomes, adding to the headway of preventive and symptomatic methodologies within the healthcare area. The following literature expects to give an outline of these procedures and their effect on diabetes research.

2.2 Role of machine learning approach for diabetic prediction

Machine learning techniques have changed the domain of the diabetic prediction, providing critical headways over conventional techniques. By utilizing huge datasets and refined approaches, ML approaches can recognize examples and connections that are not clear through the traditional measurable evaluation. These models can possibly change diabetes finding and the executives, giving more precise, convenient, and also customized predictions. An evaluation of the ML approaches has been utilized within the diabetic prediction, with its significance. The following Logistic regression, a broadly utilized classification method, succeeds within binary characterization undertakings, for example, recognizing diabetic and also non-diabetic people in view of different health measurements. Linear regression is frequently used for anticipating continuous results, for example, blood glucose levels, which are effective for diabetes management. The following Random Forest Classifier, known for its usefulness and effective accuracy, manages huge datasets successfully and is proficient at overseeing composite, nonlinear connections within the factors (Ferdous *et al.* 2020).

"Support Vector Regression (SVR)" as well as clustering methods, for example, K-Means clustering, have been effective within diabetes prediction. SVR is utilized for regression issues, giving exact expectations of blood glucose levels. K-Means assists in distinguishing particular patient groups with comparative risk profiles, assessing with designated mediations and customized treatment plans.

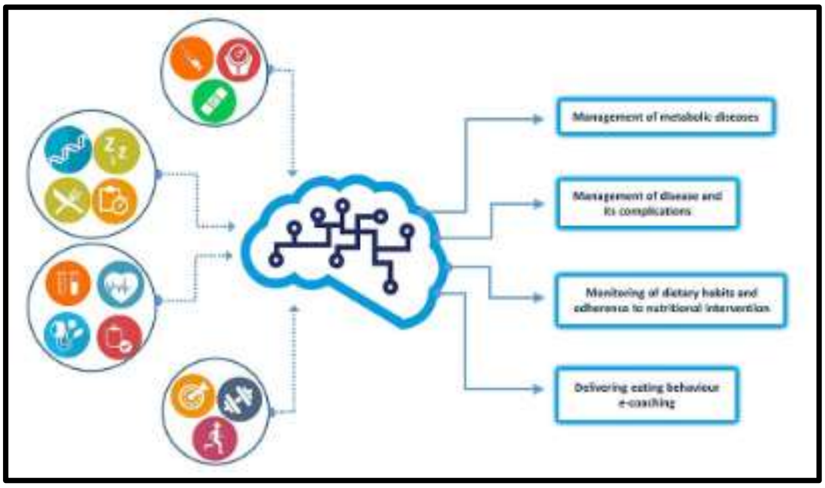

Management of metabolic diseases
Management of disease and its complications
Monitoring of dietary habits and adherence to nutritional intervention
Delivering eating behaviour e-coaching

Figure 2.2.1: Role of machine learning in diabetes prediction

(Source: Ganie and Malik, 2022)

The combination of the following ML strategies into diabetic prediction approaches has various advantages. They upgrade prescient accuracy by concerning different factors and their connections. They additionally support early finding, which is essential for compelling administration and avoidance of diabetes-related inconveniences. Moreover, these models can persistently gain and improve from new details, guaranteeing that expectations stay important and exceptional. Ethical contemplations are principal in the use of ML within medical services (Ganie and Malik, 2022). Guaranteeing patient security through the data anonymization and consenting to guidelines like the GDPR are fundamental to manage with trust and safeguard delicate data. In general, the job of ML in diabetic expectation is significant, offering creative choices for enhancing understanding results and advancing preventive medical services systems.

2.3 Impact of the machine learning on diabetic prediction

The effect of ML on the diabetic prediction is significant, denoting a critical shift from customary effective strategies to further developed, data driven solutions. By outfitting the adequacy of ML approaches, healthcare experts can accomplish higher precision in anticipating diabetes, which is urgent for early mediation and compelling disease management (Deberneh and Kim, 2021).One of the main commitments of the ML is its capacity to proficiently deal with immense measures of data. Clinical records, involving the demographic details, lifestyle components, and clinical estimations, can be dissected to distinguish unobtrusive examples and relationships that may be missed by traditional procedures. This thorough investigation empowers more exact risk evaluation and recognizable proof of high-risk people who might profit from preventive measures. Besides, ML approaches constantly gain and also adjust from new details, enhancing their prescient abilities over some time. This unique growing experience guarantees that the models stay pertinent and can consolidate the most recent clinical exploration and patient details, prompting more exact forecasts.

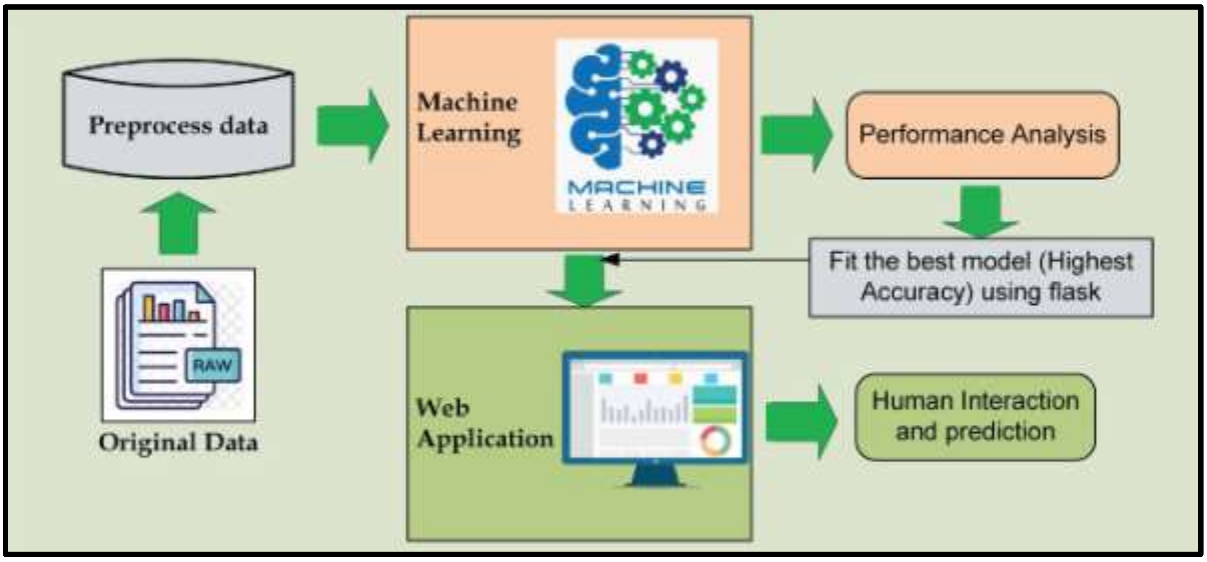


Figure 2.3.1: Impact of machine learning on diabetes prediction

(Source: Deberneh and Kim, 2021)

ML assesses customized medication by fitting expectations and treatment plans to individual patients. By taking into account a large number of elements, ML can assist with making tweaked wellbeing plans that address the remarkable requirements and hazard profiles of every patient. This modified methodology assesses the patient outcomes and lessens the likelihood of complexities. The development of the specific ML approach into the diabetic prediction maintains the resource upgrade inside clinical benefits structures. By unequivocally recognizing individuals at high risk, clinical consideration providers may assign resources effectively, assuring the opportune and assigned interventions. This can cause better organization of diabetes at both the individual and people levels, finally diminishing the load on clinical benefits structures.

2.4 Challenges in developing machine learning in diabetic prediction

Implementing the respective ML approaches for the respective diabetic prediction indicates various crucial difficulties, despite the promising improvements within this particular domain. One difficulty is the particular quality as well as the accessibility of the respective data. Greater quality data that is delegate, finished, and precise is essential for preparing powerful ML models. The healthcare details might be deficient, conflicting, or consist of errors, which can unfavorably influence the model's exhibition. Data protection and security are additionally crucial issues. Medical care details are delicate, and guaranteeing patient security while agreeing with guidelines like the "General Data Protection Regulation (GDPR)" is fundamental. Anonymizing the data adequately to secure identity of the patient without losing the data essential for precise predictions may be an issue of equilibrium to strike (Arumugam *et al.* 2023). One more issue is the interpretability of the ML approaches. While composite approaches, for example, deep learning may give greater accuracy, they frequently serve as "black boxes," making it hard to comprehend how they show up at their predictions. This absence of transparency may be dangerous within a clinical setting, where understanding the thinking behind a finding is fundamental for trust and further clinical direction. The heterogeneity of diabetes itself adds to the intricacy.

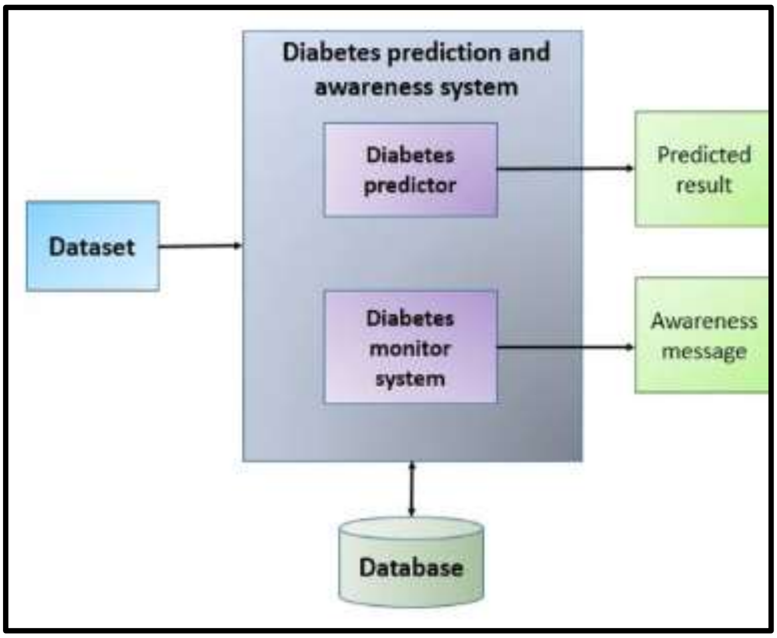


Figure 2.4.1: Challenges in diabetic prediction

(Source: Abaker, and Saeed, 2021)

Diabetes is impacted by the myriads of elements, involving lifestyle, alongside environmental viewpoints, which can shift broadly among people. Catching this inconstancy and precisely foreseeing diabetes beginning or movement across different populaces requires complex displaying strategies and far-reaching datasets. Besides, the incorporation of the ML models into the clinical choice is a huge issue. Clinicians should have the option to utilize these apparatuses flawlessly inside their work process (Abaker, and Saeed, 2021). It needs powerful and easy to use programming as well as preparing and instruction for medical care experts to comprehend and also trust the ML predictions. Predisposition within ML models is another worry. In the event that the training data isn't illustrative of the whole populace, the model might be biased, prompting less exact expectations for underrepresented groups.

2.5 Comparative Analysis of Machine Learning Approach

A comparative evaluation of the (ML) approaches for the diabetic prediction uncovers unmistakable benefits and difficulties related with various approaches. Different evaluations, each with the exceptional qualities, are utilized for predicting the diabetes, and also comprehending their relative performance is fundamental for choosing the most significant model. The Supervised learning approaches are normally utilized within the diabetic prediction (Kodama *et al.* 2022). These following models require marked preparation data to get familiar with the connection among the following input features alongside the following target variable. Several approaches assisting succeed in taking care of the composite datasets and grasping the nonlinear connections within the features. They frequently give greater accuracy and are interpretable, settling on them important for clinical choices. Though, these approaches can be inclined to overfitting, especially having small or the imbalanced datasets. Then again, ensemble approaches consolidate different models to further develop the prediction accuracy alongside Vigor. Approaches like the bagging and also boosting make an assortment of models that vote or average their forecasts, prompting improved performance. These techniques may deal with a huge volume of the data and lessen change; however, they might require critical computational resources alongside longer training times (Ahmed *et al.* 2022).

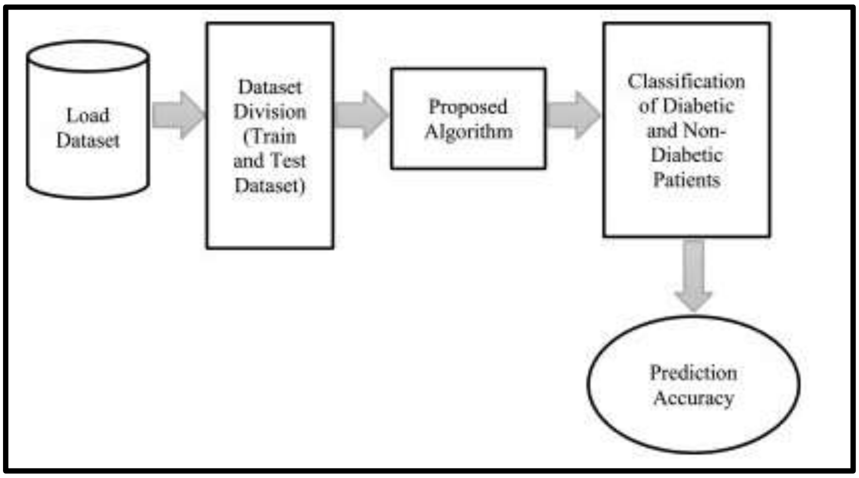

Load Dataset
Dataset Division (Train and Test Dataset)
Proposed Algorithm
Classification of Diabetic and Non-Diabetic Patients
Prediction Accuracy

Figure 2.5.1: Comparative analysis in diabetic prediction)

(Source: Hassan *et al.* 2021)

The respective Unsupervised learning procedures, like clustering, are additionally used within diabetes research. These strategies don't need the labeled data and are valuable for distinguishing examples and groupings inside the dataset. Clustering approaches can uncover hidden structures and also separate patients into the distinct risk groups, helping with customized treatment plans. Be that as it may, their exhibition intensely relies upon the decision of similitude measures and the quantity of clusters, which may not be clear to decide.

2.6 Literature Gap

In spite of critical progressions within the utilization of the ML for the diabetic prediction, various gaps stay within the current literature. One eminent gap is the restricted generalizability of numerous ML models. Most approaches center around unambiguous populaces or datasets, which may not catch the variety of the worldwide populace. Accordingly, models prepared on these datasets may perform ineffectively when applied to various demographic groups. Another gap is the test of coordinating ML models into clinical practice. While many approaches show high prescient precision in controlled settings, there is an absence of exploration on how these models can be successfully sent and used by medical care experts in certifiable situations. This incorporates issues connected with UI design, work process incorporation, alongside the clinician training (Albahli, 2020.).

Furthermore, there is a shortage of the research tending to the interpretability of intricate ML models, like deep learning. The following "black box" feature of these models can ruin their acknowledgment in clinical settings where it is vital to grasp the reasoning behind predictions. Besides, the following ethical ramifications of involving ML within the healthcare services, involving the data security, predisposition, and reasonableness, are not sufficiently evaluated in the ongoing literature. Guaranteeing that ML models don't propagate existing wellbeing variations and that patient details are safeguarded are crucial regions that require more consideration.

2.7 Summary

Within this particular research, the specific role of ML within diabetic prediction has been broadly evaluated. Different ML strategies, involving unsupervised, supervised, alongside deep learning techniques, have shown critical potential in enhancing the following accuracy of the diabetes diagnosis alongside the risk evaluation. Through these improvements, various difficulties persist, like the requirement for the high-quality, different datasets, guaranteeing model interpretability, and incorporating these models into clinical practice. Furthermore, the ethical contemplations, involving the data protection and the potential for predisposition, remain significant worries. A comparative evaluation featured the qualities and limits of various ML techniques, extending the significance of choosing appropriate models in view of explicit requirements and settings. Tending to these difficulties and gaps within the literature is significant for harnessing the maximum capacity of ML within the diabetic prediction, eventually adding to better persistent results and more compelling disease management.

Chapter 3: Methodology

3.1 Data Collection and Preprocessing

The crucial pivotal phase toward this following analysis included the intensive collection and also cleaning of the dataset to guarantee its uprightness and also convenience. This procedure started by collecting an exhaustive dataset, which involved different health associated factors crucial for the diabetes prediction and also clustering evaluation. Guaranteeing that no null values were available within the dataset was crucial. The following missing values could altogether distort the evaluation and model preparation, prompting the inaccurate outcomes. Subsequently, careful cleaning was directed to fill in or eliminate any incomplete data entries, guaranteeing a powerful dataset for additional evaluation (Haque *et al.* 2021).

Then, the change of the categorical factors into the numerical values was performed utilizing the label encoding. Several machine learning evaluations need the numerical input, and also categorical data within their raw structure may impede the following modelling procedure. The following Label encoding successfully changed over these particular categorical properties into the numerical form, assessing adequate incorporation into the particular machine learning pipeline. This specific step was fundamental for factors, for example, gender and also smoking history, which should have been numerical for the model effectiveness. Also, the following "exploratory data analysis (EDA)" was directed to assess and also represent key factors, offering details into the dataset's construction and also patterns. Factors, for example, gender count, age distribution, and also HbA1c levels were inspected. Histograms along with count plots were used to portray the circulation of ages and also the gender composition into the specific dataset, uncovering patterns and also likely inclinations. Likewise, the overall distribution of the HbA1c levels was imagined to grasp its reach and also central tendency, featuring its significance as a predictor for diabetes.

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)", and 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 also blood glucose levels. 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. Within this following analysis, the particular K-Means clustering approach was applied to the segment of the populace in view of two crucial health indicators: BMI (Body Mass Index) alongside the blood glucose levels. 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 also blood glucose levels. The specific centroids, indicated by the red 'X' indications within this visualization, demonstrated the essential issues of these clusters,limiting the overall distance among the data points and also their relegated cluster. Perceptions of the following cluster featured three particular groups inside the populace: one with the lower BMI alongside the blood glucose levels, one more with moderate qualities, alongside a third with the greater BMI and also the blood glucose levels (Das *et al.* 2022). These following clusters give a reasonable segmentation of the populace, assessing the recognition of the high-risk groups along with empowering more engaged health interventions.

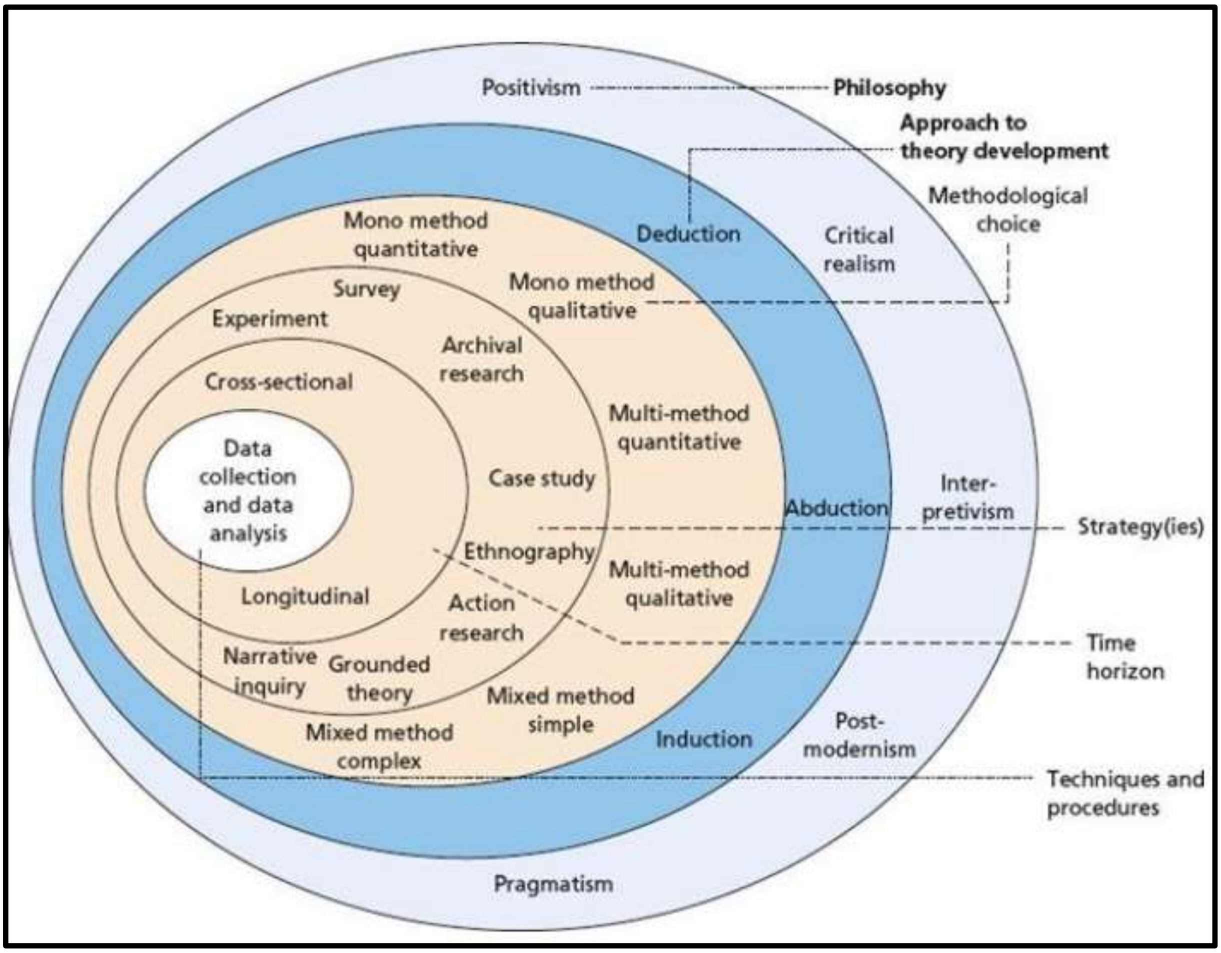


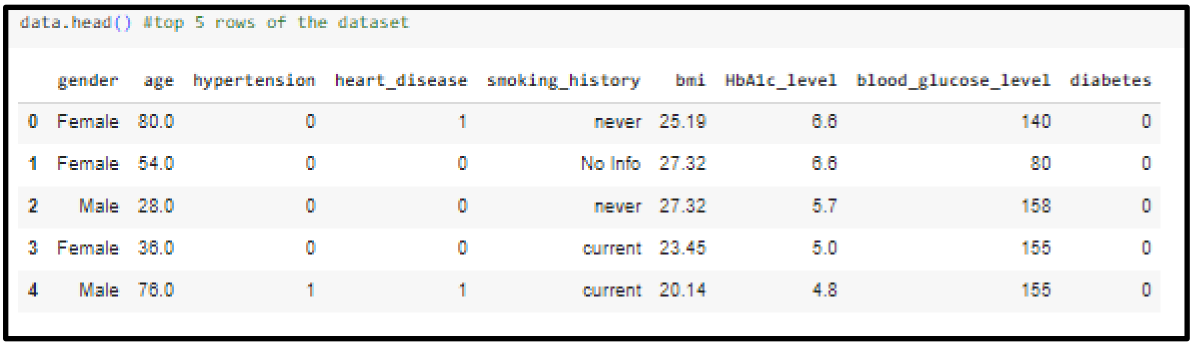
Figure 3.5.1: Research Onion

(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 clustering for the data analysis.

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. In addition, each section is accompanied by histograms, box plots, and confusion matrices to present the performance of the models and to shed light on possible directions for enhancement in the diabetes risk prognosis space.

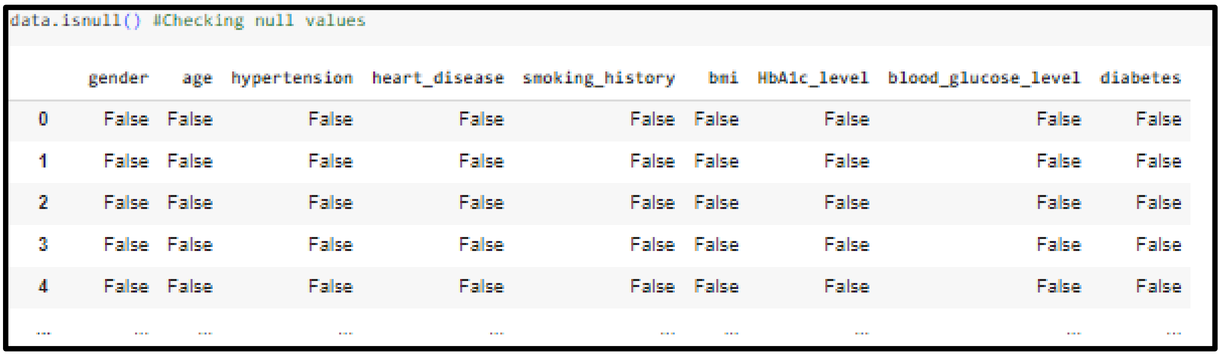


```
data.head() #top 5 rows of the dataset
```

	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. Assessing these underlying columns assists in figuring out the dataset's design and the connections between factors, assessing the making of additional precise and compelling prescient models for recognizing individuals at risk of diabetes.

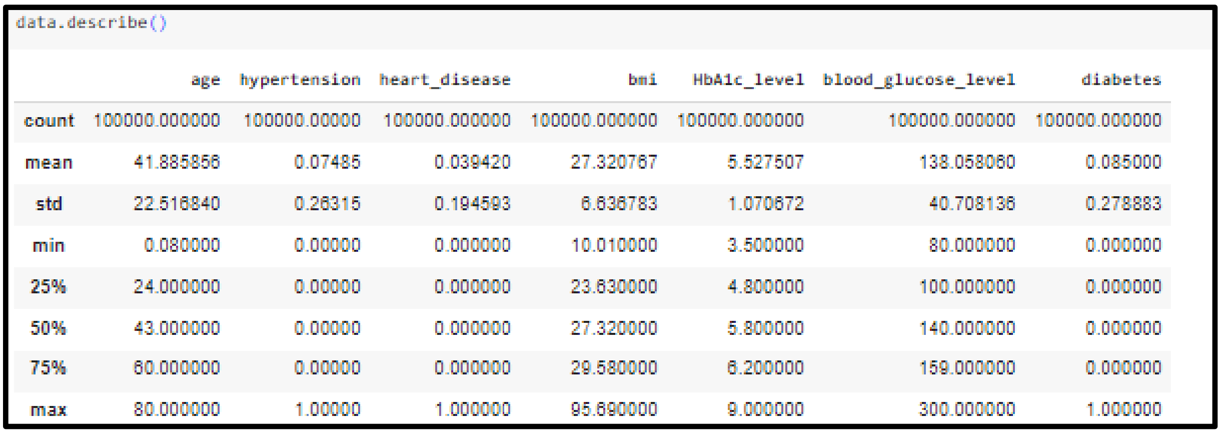


```
data.isnull() #Checking null values
```

	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
...	...	...	...	...	...	...	...	...	...

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. This fulfillment guarantees that the following machine learning approaches may use the whole dataset without requiring attribution or evacuation of rows, prompting more strong and exact prescient models. By affirming the absence of the null values, the uprightness of the specific data can be improved and the resulting examination for anticipating diabetes, guaranteeing dependable and noteworthy details.

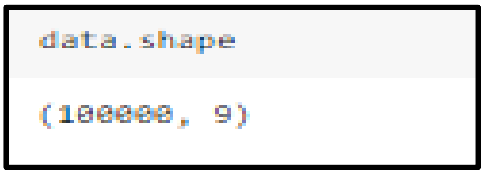


```
data.describe()
```

	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
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. The minimum alongside maximum values feature the scope of the data for every variable. These illustrative measurements are pivotal for grasping the dataset's circulation and fluctuation, supporting the advancement of precise diabetes prediction approaches.



```
data.shape
```

```
(100000, 9)
```

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. The enormous number of rows proposes a significant measure of the data, which can upgrade the vigour and precision of the prescient models for diabetes identification. The specific 9 columns address various factors that will be assessed to recognize trends and relationships connected with diabetes risk factors.

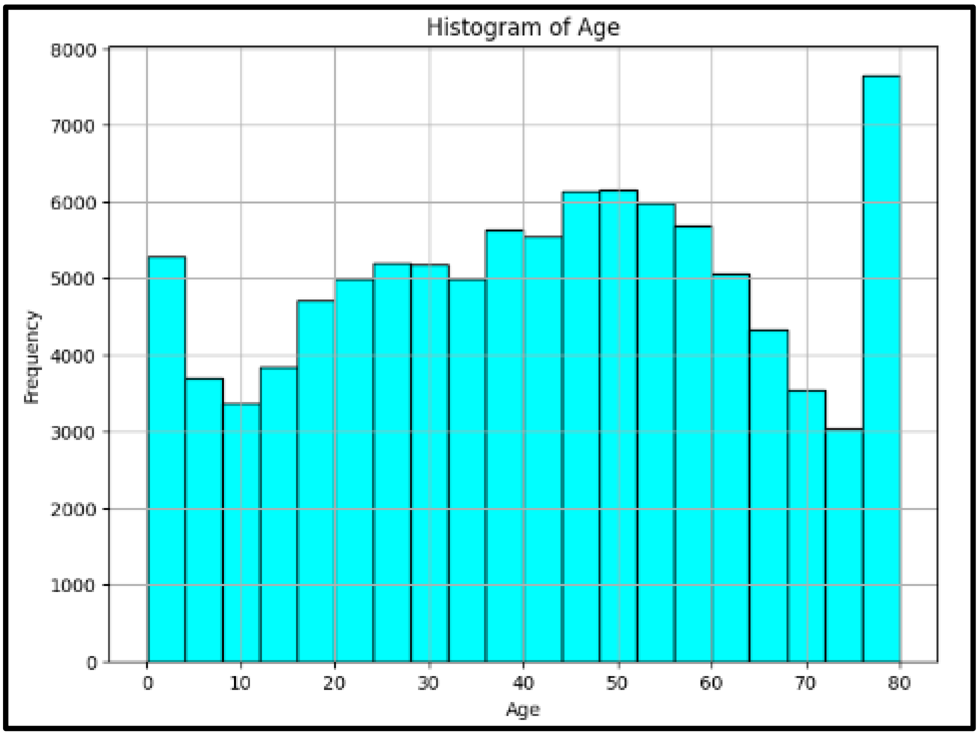


Figure 4.5: Histogram for frequency of the Age

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. The histogram assists in recognizing the following age groups that are more crucial within the dataset, which can impact the advancement of designated methodologies for early location and counteraction of diabetes within various age demographics. This visual portrayal supports getting a handle on the age dispersion and pinpointing expected age-related patterns within diabetes occurrence.

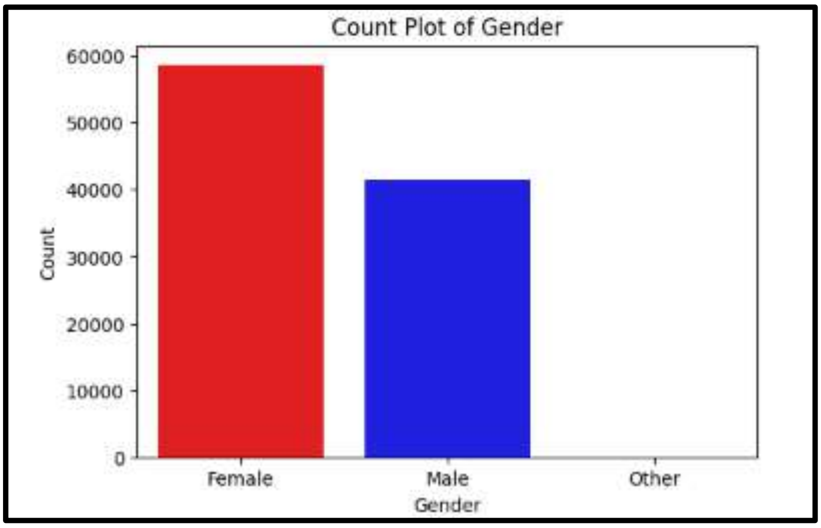


Figure 4.6: Count plot of gender

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. The particular count plot serves a crucial tool within envisioning the demographic development of the dataset, guaranteeing that the evaluation represents the possible effect of the gender on diabetes prevalence alongside progression. The particular visualization highlights the significance of concerning gender as a variable within generating precise and successful prescient models for diabetes.

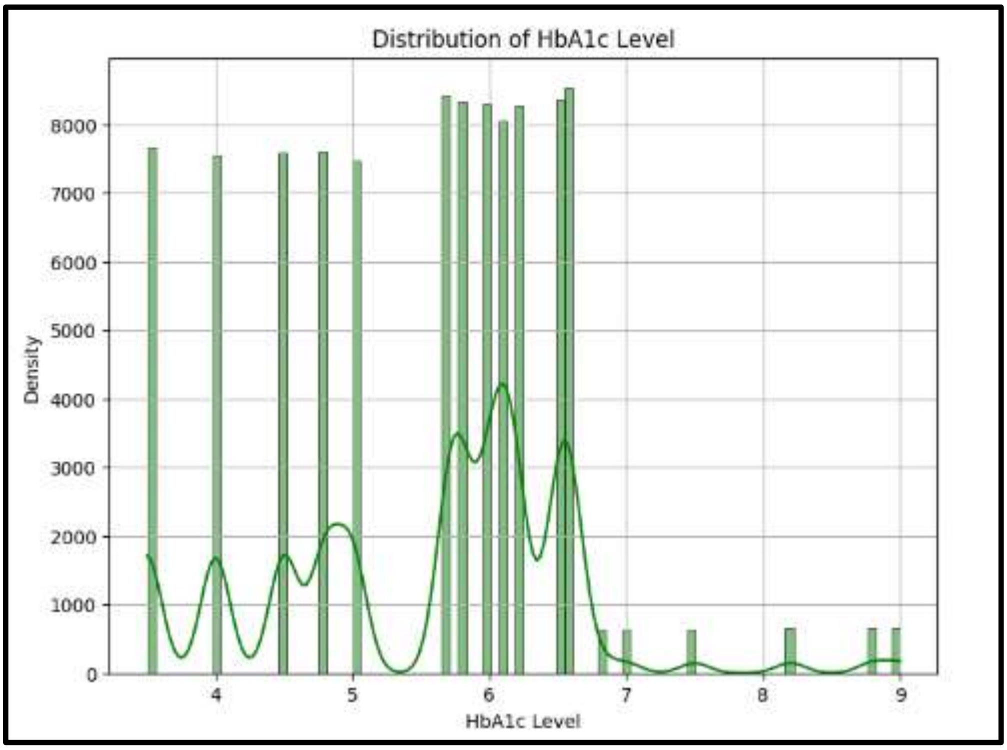


Figure 4.7: Distribution of the HbA1c level

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. Assessing the appropriation assists in comprehending the prevalence of various HbA1c levels within the populace, supporting the improvement of the designated interventions and treatment methodologies. This representation is fundamental for distinguishing patterns within HbA1c levels, which can illuminate better clinical choices and also enhance the prescient model accuracy for the diabetes identification and also management.

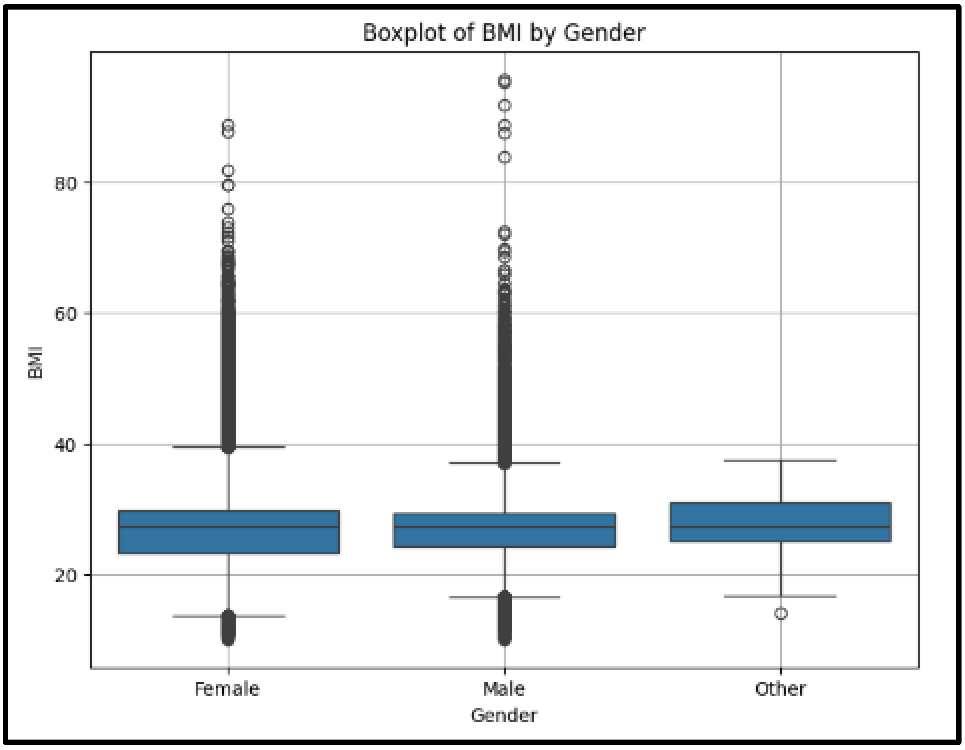


Figure 4.8: Box plot of BMI by gender

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)". Whiskers stretch out to show the range of the particular data, elevating the outliers, which are plotted as individual points. This representation recognizes contrasts in BMI conveyances across the gender categories and features crucial varieties or the anomalies.

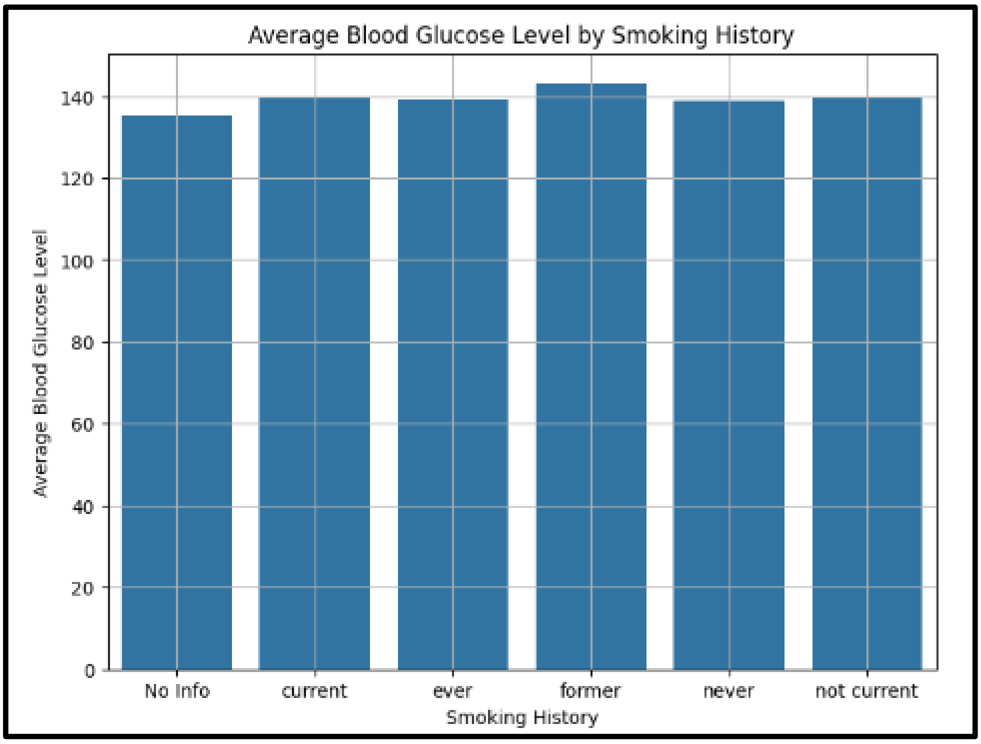
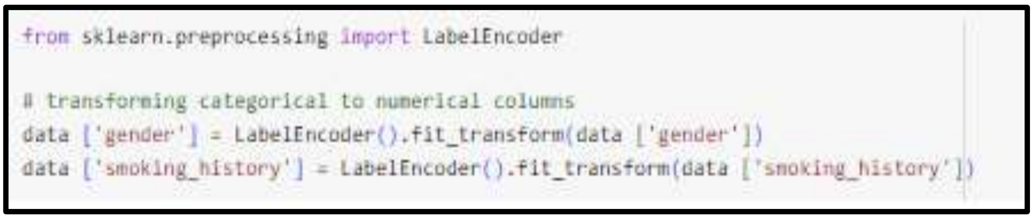


Figure 4.9: Average Blood Glucose Level by the Smoking History

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. Comprehending these examples assists in distinguishing high-risk gatherings and fitting precaution measures or mediations appropriately. The particular figure highlights the significance of considering lifestyle factors like the smoking history within diabetes forecast and also the management approaches.

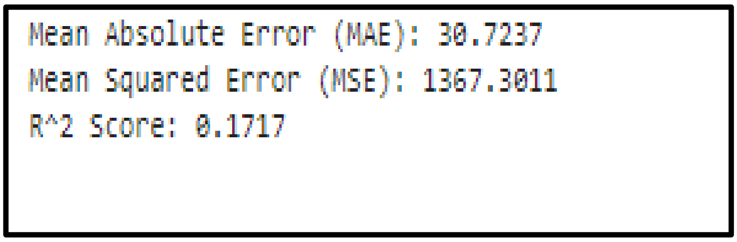


```
from sklearn.preprocessing import LabelEncoder

# transforming categorical 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

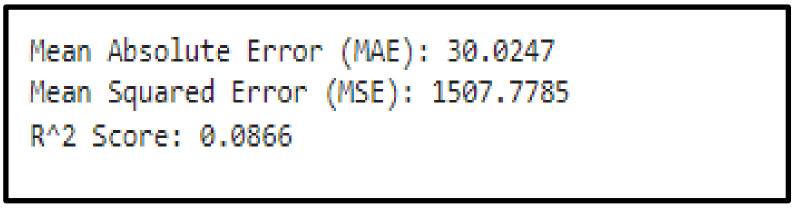
In this particular step the categorical column is transformed into numerical columns using the label encoder. It is an important step toward preparing the specific dataset for the machine learning approaches, which need the numerical input. By changing the categorical variables like the gender, smoking history, or further categorical features into the numerical values, the dataset becomes viable with different approaches. This change assesses with better data evaluation and also modelling, empowering the specific machine learning approaches to successfully process and gain from the details. The figure epitomizes how every class is relegated an exceptional numerical value, guaranteeing an organized and uniform dataset for the subsequent evaluation.



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

Figure 4.11: Linear Regression Result

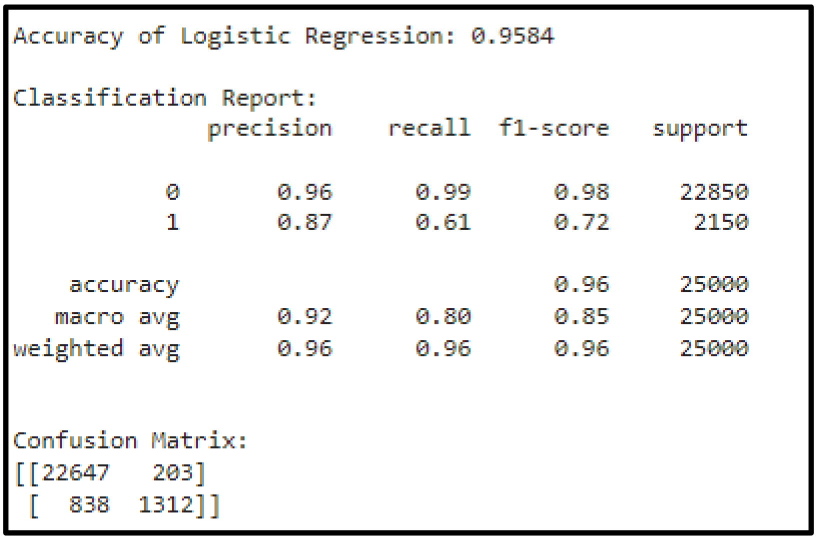
The specific figure indicates crucial performance measurements for the specific Linear Regression approach. The specific predictions of the model are off by around 30.72 units from the following true values. The particular "Mean Squared Error (MSE)" of about 1367.3011 assesses the average squared discrepancy 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 R² 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): 30.0247
Mean Squared Error (MSE): 1507.7785
R^2 Score: 0.0866
```

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 1507.7785 shows the average of the squared contrasts among the predicted and also actual values, with a greater value recommending a poorer fit. The following R² 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. These measurements recommend that the model struggles with precision and also requires further refinement to enhance its exhibition and prescient capacity.



```
Accuracy of Logistic Regression: 0.9584

Classification Report:
              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. In spite of the great accuracy, the approach struggles with the following Class 1 identification, as proven by a significant number of the False Negatives. A more profound evaluation of the class-explicit measurements is important to address these constraints and enhance the model performance.

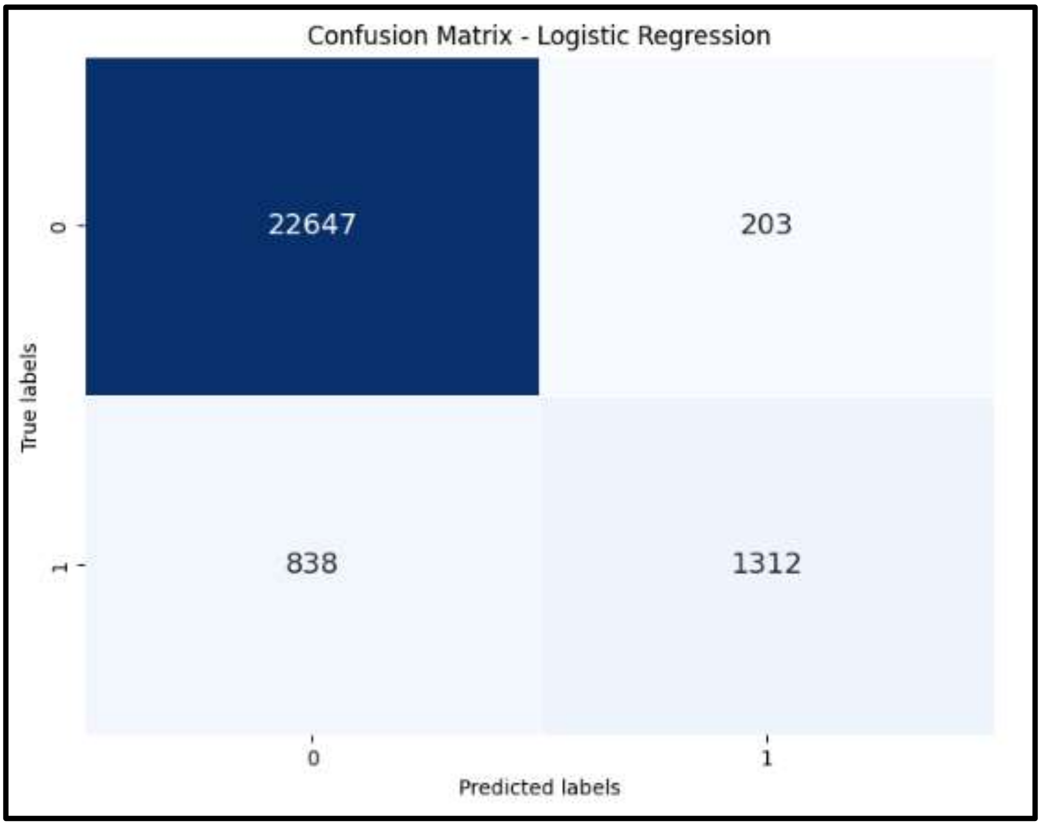
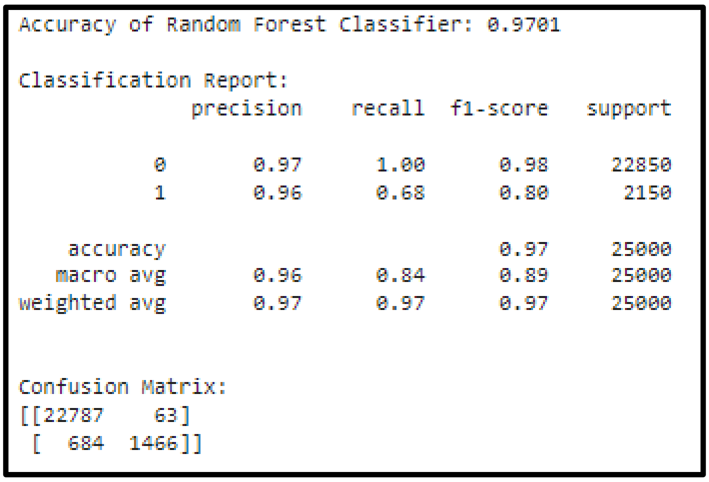


Figure 4.14: Confusion Matrix of Logistic regression

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. This particular visualization features the model's presentation in recognizing diabetic as well as non-diabetic instances, underscoring domains where the model performs well and where the enhancements are required.



```
Accuracy of Random Forest Classifier: 0.9701

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:
[[22787    63]
 [  684  1466]]
```

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 97.81%, 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 towardsidentifying diabetic cases. This exhibits adequate accuracy, assessing major areas of strength for predicting the non-diabetic occasions. Though, its performance on classifying the diabetic cases uncovers opportunities for enhancement. While the following classifier succeeds in recognizing non-diabetic cases, the lesser recall for the diabetic cases proposes that the particular model might profit from further tuning or elective ways for dealing with the sensitivity and guarantee more exact distinguishing recognition of the diabetic individuals.

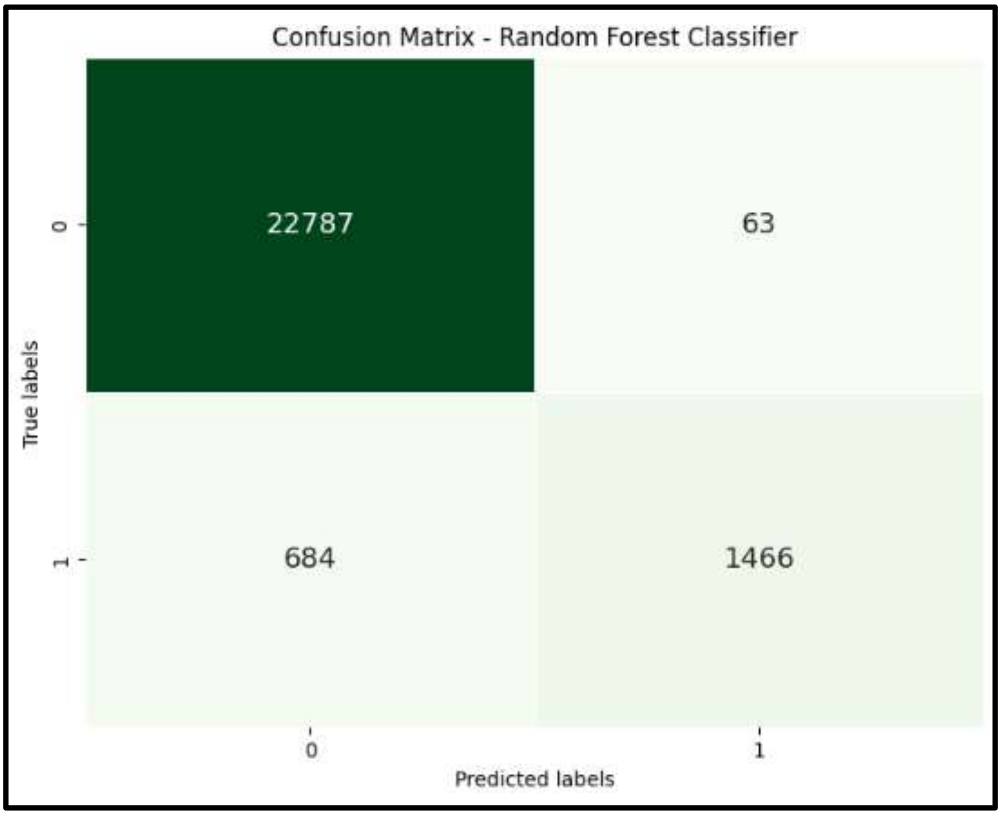


Figure 4.16: Confusion Matrix of Random Forest Classifier

In the following confusion matrix, there are about 22,787 True Negatives (TN) alongside 63 False Positives (FP) (FP), assessing the model's effective performance in distinguishing non-diabetic occasions. However, there are about 684 False Negatives (FN) alongside 1,466 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 22,787 True Negatives (TN) along with 63 False Positives (FP), the approach shows greater accuracy in recognizing non-diabetic cases. Though, the presence of 684 False Negatives (FN) along with 1,466 True Positives (TP) shows difficulties in accurately distinguishing diabetic cases. This disparity proposes the requirement for additional model refinement, especially to further develop responsiveness and lessen the quantity of missed diabetic determinations. Such changes could improve the model's exhibition and unwavering quality in clinical predictions.

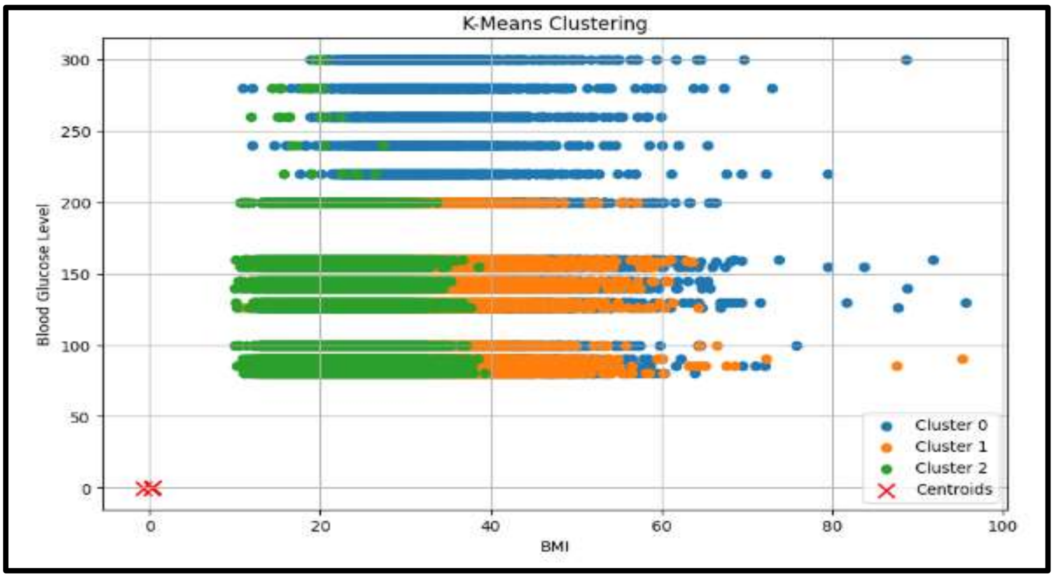


Figure 4.17: Showing the K means Clustering Graph

The following K-Means Clustering chart shows the conveyance of the individuals in the view of BMI alongside the Blood Glucose Level, with three distinct clusters distinguished. Cluster 0 (low BMI as well as blood glucose), Cluster 1 (moderate levels), alongside Cluster 2 (high levels) are outwardly addressed by various colors. Red 'X' markers demonstrate the centroids of every cluster. This evaluation features likely subgroups, providing details for designated health interventions alongside further evaluation of the health-associated patterns.

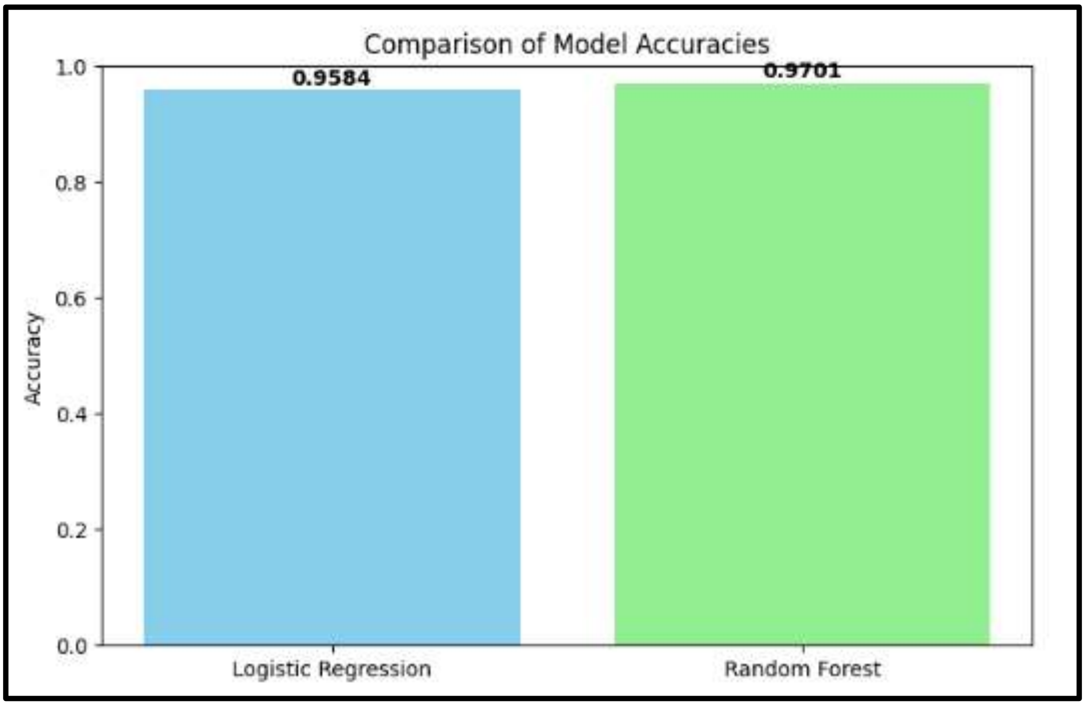


Figure 4.18: Comparison graph of the classification models

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 different 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 Réd 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 three clusters primarily according to the BMI and blood glucose categories. Such clusters which are commonly classified into low, moderately, 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 97.81%. 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 also blood glucose levels. The representation of the clusters and centroids uncovered three unmistakable subgroups, featuring potential high-risk bunches requiring designated mediations. The following Cluster 0 included individuals with the lower BMI and also blood glucose levels, the respective cluster 1 addressed those with moderate levels, and also group 2 incorporated individuals with more elevated levels, demonstrating a higher risk for the diabetes entanglements. These details highlight the significance of customized medical services procedures, as the clustering evaluation recognized explicit clustering that might profit from various sorts of interventions.

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 blood glucose levels 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). Tending to these restrictions in future evaluation can improve the strength and appropriateness of the details, eventually adding to more powerful diabetes predictions and the management procedures.

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). Furthermore, tending to the impediments of the clustering approaches by exploring further procedures, like the DBSCAN or the particular hierarchical clustering, may yield more precise subgroupings.

Reference :

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

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

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

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

Ferdous, M., Debnath, J. and Chakraborty, N.R., 2020, July. Machine learning algorithms in healthcare: A literature survey. In 2020 11th International conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE. https://www.researchgate.net/profile/Narayan-Chakraborty/publication/346238330_Machine_Learning_Algorithms_in_Healthcare_A_Literature_Survey/links/60a3500c92851cc80b61036c/Machine-Learning-Algorithms-in-Healthcare-A-Literature-Survey.pdf

Ferdous, M., Debnath, J. and Chakraborty, N.R., 2020, July. Machine learning algorithms in healthcare: A literature survey. In 2020 11th International conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE. https://www.researchgate.net/profile/Narayan-Chakraborty/publication/346238330_Machine_Learning_Algorithms_in_Healthcare_A_Liter

ature_Survey/links/60a3500c92851cc80b61036c/Machine-Learning-Algorithms-in-Healthcare-A-Literature-Survey.pdf

Deberneh, H.M. and Kim, I., 2021. Prediction of type 2 diabetes based on machine learning algorithm. International journal of environmental research and public health, 18(6), p.3317. https://www.mdpi.com/1660-4601/18/6/3317/pdf

Arumugam, K., Naved, M., Shinde, P.P., Leiva-Chauca, O., Huaman-Osorio, A. and Gonzales-Yanac, T., 2023. Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings, 80, pp.3682-3685. https://www.researchgate.net/profile/Mohd-Naved/publication/353651472_Multiple_disease_prediction_using_Machine_learning_algorithms/links/61279a342b40ec7d8bc8275c/Multiple-disease-prediction-using-Machine-learning-algorithms.pdf

Abaker, A.A. and Saeed, F.A., 2021. A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. Informatica, 45(1). https://www.informatica.si/index.php/informatica/article/viewFile/3111/1499

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

Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A., Ghazal, T.M. and Ahmad, M., 2022. Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, pp.8529-8538. https://ieeexplore.ieee.org/iel7/6287639/9668973/09676634.pdf

Kodama, S., Fujihara, K., Horikawa, C., Kitazawa, M., Iwanaga, M., Kato, K., Watanabe, K., Nakagawa, Y., Matsuzaka, T., Shimano, H. and Sone, H., 2022. Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. Journal of diabetes investigation, 13(5), pp.900-908. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/jdi.13736

Ganie, S.M. and Malik, M.B., 2022. Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus. International Journal of Medical Engineering and Informatics, 14(6), pp.473-483. https://www.researchgate.net/profile/Shahid-Ganie/publication/349858232_Comparative_analysis_of_various_supervised_machine_learni ng_algorithms_for_the_early_prediction_of_type-II_diabetes_mellitus/links/632451e170cc936cd311caf7/Comparative-analysis-of-various-supervised-machine-learning-algorithms-for-the-early-prediction-of-type-II-diabetes-mellitus.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

Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R. and Saba, T., 2019. Current techniques for diabetes prediction: review and case study. Applied Sciences, 9(21), p.4604. https://www.mdpi.com/2076-3417/9/21/4604/pdf

Haque, F., Bin Ibne Reaz, M., Chowdhury, M.E.H., Srivastava, G., Hamid Md Ali, S., Bakar, A.A.A. and Bhuiyan, M.A.S., 2021. Performance analysis of conventional machine learning algorithms for diabetic sensorimotor polyneuropathy severity classification. Diagnostics, 11(5), p.801. https://www.mdpi.com/2075-4418/11/5/801/pdf

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

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

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

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

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

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

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

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

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

Syed, A.H. and Khan, T., 2020. Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study. IEEE Access, 8, pp.199539-199561. https://ieeexplore.ieee.org/iel7/6287639/6514899/09245498.pdf