9-2023

Using ML to Understand the Factors Impacting Diabetes in Diabetic Patients

Marwa Ahmed Almarzooqi
maa7919@rit.edu

RIT

Using ML to Understand the Factors Impacting Diabetes in Diabetic Patients

By

Marwa Ahmed Almarzooqi

A Graduate Capstone Submitted in Partial Fulfilment of the Requirements for the

Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

September 2023

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Marwa Ahmed Almarzooqi

Paper/Capstone Title: Using ML to Understand the Factors Impacting Diabetes in Diabetic Patients

Graduate Capstone Committee:

Khalil Alhussaeni	**Member of committee (Mentor)**	**Date**
Sanjay Modak	**Chair of committee**	**Date**

Abstract

Diabetes, a well-known medical condition since ancient times, has become a prevalent and significant health concern in recent decades. The rising incidence of diabetes has necessitated early diagnosis and effective treatment. Machine learning (ML) innovations have revolutionized disease prediction and decision-making by utilizing massive datasets. This study aims to develop and compare machine learning (ML) models for diabetes prediction using a preprocessed dataset of 532 instances obtained from Kaggle.

Important variables included in the data set are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. The correlation analysis revealed a strong positive association between Glucose and Outcome, suggesting that elevated glucose levels are associated with an increased risk of diabetes. Similarly, Outcome and Age demonstrated a positive correlation, suggesting that age may be a risk factor.

Six ML models, including Voting, Extra Trees, Bagging, Gradient Boosting, Logistic Regression (LR), and Random Forest Regression (RFR), were trained and optimized using Randomized Search CV for hyperparameter tuning. Using metrics such as Sensitivity, Specificity, Precision, Negative Predicted Value, and Accuracy to evaluate the models revealed the respective models' strengths and weaknesses. Both diabetic and non-diabetic cases had the highest predictive accuracy with the Extra Trees model.

Additional feature significance analysis utilizing SHAP summary plots revealed that "Glucose" and "Age" were the most influential diabetes prediction features. These results highlight the diagnostic value of these characteristics.

This investigation concludes with a thorough comparison of ML models for diabetes prediction. The findings demonstrate the potential of machine learning techniques for early disease detection and decision making. Analyses of the optimized models' performance and the significance of their features contribute valuable insights to the field of diabetes management. Future research may enlarge the dataset and investigate additional potent ML algorithms, thereby potentially improving the accuracy of predictions and facilitating personalized patient care.

KEYWORDS: Diabetes, machine learning, predictive modeling, feature importance, early disease detection

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Introduction

Diabetes, a medical condition known since ancient times, was acknowledged as a significant ailment, but its prevalence was not commonly encountered or well understood by medical practitioners or healers. However, in recent decades, the impact of diabetes on human health and societal development has grown significantly with a rising number of affected individuals. This chronic illness is marked by elevated blood glucose levels and disturbances in the metabolism of fats and proteins. The rise in blood glucose occurs when it cannot be adequately metabolized within the cells due to insufficient insulin production by the pancreas or the cells' reduced ability to utilize the produced insulin effectively(Roglic, 2016). Diabetes, a complex metabolic disorder characterized by disruptions in insulin production and blood sugar regulation (Adhi Tama, S, & Hermansyah, 2011), poses a substantial burden on both individual health and global healthcare systems. Hyperglycemia, a hallmark of diabetes, contributes to severe long-term complications, encompassing eye, kidney, and nerve diseases, as well as circulatory issues and the risk of amputation(Adhi Tama et al., 2011). Beyond its physiological ramifications, diabetes also exerts a significant economic toll, straining national healthcare budgets(Williams et al., 2020).

The insidious nature of diabetes, often devoid of overt clinical symptoms (Adhi Tama et al., 2011), underscores the challenge in achieving accurate diagnoses and appropriate treatments. This predicament has prompted a pressing need for early detection, prevention, and effective disease management strategies. Recognizing this urgency, recent research has delved into interventions to enhance blood sugar control, revealing the profound benefits of maintaining optimal glucose levels in curbing the progression of diabetes-related complications(Adhi Tama et al., 2011). Importantly, holistic diabetes management extends beyond medical interventions, encompassing lifestyle factors such as smoking cessation and weight management (Richard I. G. Holt, 2011).

As the global healthcare landscape grapples with the mounting healthcare expenditures attributed to diabetes(Williams et al., 2020), a comprehensive understanding of the multifaceted impact of the condition is crucial.

Diabetes mellitus can be classified into different types based on its underlying causes and characteristics.

1.1.1 Type 1 Diabetes

It relates to the deterioration of beta cells situated in the pancreas, causing a total absence of insulin. This specific form of diabetes is often linked with an autoimmune response, where the body's immune system erroneously targets and eliminates its own beta cells. In certain situations, the exact cause behind the beta cell degradation remains undisclosed, a condition known as idiopathic.

1.1.2 Type 2 Diabetes

Type 2 diabetes, conversely, encompasses a variety of situations. It spans from a condition primarily marked by resistance to insulin, where the body's cells inadequately react to insulin, to a condition primarily marked by a shortfall in insulin secretion, where the pancreas doesn't generate sufficient insulin, although the cells may still acknowledge it. In type 2 diabetes, both insulin resistance and inadequate insulin secretion may occur simultaneously.

1.1.3 GESTATIONAL DIABETES

Gestational diabetes is a transient kind of diabetes that emerges during pregnancy. It is marked by elevated blood sugar levels resulting from the hormonal shifts and insulin resistance linked to pregnancy. Typically, gestational diabetes resolves following childbirth; however, it heightens the likelihood of developing type 2 diabetes in the future. (Bilous, 2021).

Additionally, there are other specific types of diabetes, such as those caused by genetic defects in beta cell function or insulin action. Diabetes can also be associated with diseases of the exocrine pancreas, which affects the organ's ability to produce digestive enzymes and hormones like insulin. Endocrinopathies (diseases affecting hormone-producing glands), drug-induced or chemical-induced diabetes (caused by certain medications or chemicals like steroids), infections, and uncommon forms of immune-mediated diabetes are among other specific types.

The recent progress in machine learning has greatly improved the capacity of computer systems to recognize and classify images, predict diseases, and improve decision-making through extensive data analysis. The main objective of machine learning applications is to surpass human capabilities by achieving superior performance. This is accomplished by utilizing supervised learning algorithms to train the models, and their effectiveness is assessed using distinct testing data (Healthcare Engineering, 2023).

The digitization of medical records has yielded a wealth of data with multiple dimensions, offering an exceptional platform for the utilization of robust machine learning techniques to recognize intricate patterns and foresee potential outcomes. This phenomenon presents an exclusive occasion for the application of advanced computational methods to discern complex relationships and forecast various results (Anderson et al., 2015). These machine learning methods are well-known for their skill in uncovering complex patterns in large sets of information, making them a strong tool for foreseeing future possibilities. The merging of digital medical records and advanced machine learning techniques opens an exclusive opportunity – the ability to use the latest computational methods to find hidden connections within this data. This is where our study comes in. We're using a machine learning approach to predict and understand diabetes, a common and complex health issue influenced by many factors.

1.2 Research Objectives/ Project Goals

The objective of the current study is to determine how Ml- machine learning can be implemented in highlighting the factors that cause diabetes using the data set of the patient. The study also seeks to expand the knowledge gap that exists in highlighting the importance of the risk factors and predicting the early stages of diabetes. Therefore, the objectives of the study are stated as follows:

- To apply Machine learning to understand the different factors of diabetes in patient
- To evaluate the effectiveness of the ML models in predicting the diabetes factors.
- To recommend a ML-based technique for future studies in this field.

1.3 Research Questions

1. What is the use of machine learning for understanding different factors for diabetes in people?
2. How can machine learning techniques and approaches predict diabetes by analyzing the risk indications from epidemiological information?

Chapter 2

2. Literature review

Diabetes, a chronic and life-threatening disease characterized by elevated blood sugar levels, poses a significant public health challenge worldwide (Healthcare Engineering, 2023). The profound impact of diabetes on cardiovascular health has been extensively investigated, shedding light on its association with an increased risk of various vascular diseases. A comprehensive meta-analysis conducted by the Emerging Risk Factors Collaboration synthesized findings from multiple prospective studies, providing invaluable insights into the magnitude and diversity of these associations (N. Sarwar et al., 2010).

Diabetes, characterized by disturbances in insulin regulation and elevated blood glucose levels, emerges as a pivotal factor contributing to the development of coronary heart disease and major stroke subtypes. The study's adjusted hazard ratios (HRs) underscore the severity of these associations, revealing a twofold excess risk for vascular complications in individuals with diabetes (N. Sarwar et al., 2010). These findings underscore the independent nature of diabetes as a risk factor, transcending other conventional risk determinants. Early diagnosis is pivotal, both for identifying high-risk individuals and for encouraging proactive lifestyle management to mitigate risks (Shan, R. et al., 2019). Timely detection and accurate prediction of diabetes are essential for effective medical decision-making, personalized patient care, and prevention of potential complications. In the past few years, the utilization of machine learning methods in the healthcare sector has demonstrated encouraging outcomes concerning the early prediction and diagnosis of diabetes. While conventional methods of chronic disease management often rely on rule-based systems or predictive scores, they may fall short compared to the potential of machine learning techniques in identifying patient health conditions (Bruen et al., 2017) This extensive examination of relevant literature seeks to delve into numerous research papers that delve into the

prediction and diagnosis of diabetes through the application of diverse machine learning algorithms. By shedding light on their efficacy and potential implications for enhancing healthcare outcomes, this review aims to provide valuable insights into this rapidly evolving field.

In, (Barik et al., 2021) the focus of the researchers was on analyzing the accuracy of prediction of diabetes. It was done by two machine learning algorithms. One was random forest which is a classification-based algorithm. The other one is XGBOOST which is a hybrid algorithm. Diabetes' rapid growth, which also includes youngsters, is making its early detection very crucial for effective medical decision making and preventing the individual from complications.

The aim of the study was to identify a method which can accurately predict diabetes early enough to save lives as well as to prevent organ damage.

The methodology involved the implementation as well as the comparison between the two models. The models were trained on a dataset and then their prediction accuracy was compared. The results showed that the best performing model was xgboost. The mean prediction score of XGBoost was 74.10%. The parallel tree-based approach of xgboost provided better results. The predictions were faster, leveraging hardware and software optimally.

The importance of the use of machine learning algorithms was implied by the findings of this study. The authors acknowledged that there is a potential for improvement by applying other machine learning algorithms and techniques like optimization.

Similarly, model for early diabetes prediction using data mining techniques is proposed by (Alam et al., 2019) Principal component analysis was implemented by the authors to gather information about the significant features. The relationship of features was studied for the prediction of diabetes.

In this study authors used 3 different machine learning techniques named Artificial Neural Network (ANN), Random Forest (RF), and K-means clustering. The ANN provided the highest accuracy of 75.7% among them.

The authors acknowledge the importance of machine learning and data mining techniques in the prediction of diabetes and other diseases.

In (M. A. Sarwar et al., 2018) The authors used six different machine learning algorithms to explore the predictive analysis in healthcare by predicting the diabetes. The best performing algorithms were KNN and SVM. The highest accuracy was 77%. The research highlights the importance of the data driven approaches when it comes to medical-decision-making.

The conventional process of identifying diabetes involves visiting a diagnostic center and consulting a doctor, which can be time-consuming and inefficient. However, in (Sisodia & Sisodia, 2018), the author leverages machine learning approaches to design a model capable of accurately predicting the likelihood of diabetes in patients.

In this study, three machine learning classification techniques were utilized: Decision Tree, SVM, and Naive Bayes, with the objective of early diabetes detection. The research utilized the Pima Indians Diabetes Database (PIDD), obtained from the UCI machine learning repository, for experimentation purposes. Among these three algorithms, Naive Bayes exhibited the highest accuracy rate, achieving a notable 76.30%, surpassing the performance of the other methods.

The study concludes that the designed system, incorporating the Naive Bayes classification algorithm, shows promising results in predicting diabetes at an early stage. This approach can significantly impact real-world medical problems, facilitating timely diagnosis and intervention.

Decision support system using the AdaBoost algorithm with Decision Stump is proposed by (Vijayan & Anjali, 2015). The accuracy of 80.72 was achieved by the system. It outperformed other classifiers like support vector machine, naïve bayes and decision tree. The research emphasizes that the techniques of data mining have great potential in medical data science and decision making.

In this study the techniques of data mining can be useful for the estimation of different disease patterns and extracting medical information. The system that was presented in the paper showed a high accuracy of 80.729%. Similarly (Pradhan & Bamnote, 2014) introduces a Classifier for Diabetes detection utilizing Genetic Programming (GP). This innovative model integrates genetic and immunological features, crucial for predicting autoimmune diabetes. The study stresses that patterns of immune response correspond with disease progression stages, thereby enhancing predictive accuracy. (Tama et al., 2011) reinforces the global significance of diabetes, particularly Type-2 Diabetes Mellitus, estimating its prevalence to reach 438 million within 20

years. The study emphasizes the application of data mining and machine learning for early diabetes detection, leveraging historical patient records for knowledge discovery. The implementation of various learning methods facilitates informed decision-making for clinicians. On the other hand (Jayalakshmi & Santhakumaran, 2010) addresses the challenge of missing data in medical datasets, particularly in the context of Artificial Neural Networks (ANNs). The paper explores diverse missing value techniques to enhance classification accuracy. Furthermore, it highlights the critical impact of preprocessing on classification outcomes, emphasizing the importance of data quality. (Muller et al., 2015) sheds light on the early detection of Gestational Diabetes Mellitus (GDM), a crucial concern in developing nations. The research proposes a Multilayer Neural Network-based decision support system, enabling efficient GDM diagnosis without conventional blood tests. This innovative approach simplifies diagnosis, making it more cost-effective.

(Lesmana et al., 2011) reiterates the escalating global diabetes crisis and the need for precise disease diagnosis. The study focuses on employing back propagation networks, trained by the Levenberg–Marquardt algorithm, to enhance diabetes detection. This utilization of machine learning underscores the potential of advanced technology in medical diagnosis. (Wibawa & Hery Purnomo, 2006) introduces an unconventional method to assess insulin deficiency through iris diagnosis. The paper presents a computerized iris inspection method, harnessing digital image processing techniques. This approach seeks to detect pancreatic abnormalities linked to insulin deficiency, showcasing the integration of medical science and technology. (Zhang et al., 2014) explores the realm of iridology as an alternative mechanism for evaluating internal organ conditions. The study emphasizes the use of real-time video imaging to capture iris images, which then undergo image processing techniques. By detecting broken tissues in specific iris areas, this method aims to provide insights into organ health and as a result helps with the early detection of diabetes. (Durairaj, 2015) revolutionizes diabetes detection with a noninvasive approach based on facial block color features. The study employs a sparse representation classifier to analyze facial images, extracting color features from specific facial blocks. This unique application of image analysis enhances the potential for noninvasive diabetes diagnosis.

2.1 KEY TAKEAWAYS

- The studies showcase the significance of early detection in managing diabetes effectively and preventing complications.
- Various machine learning algorithms, such as Random Forest, XGBoost, Naive Bayes, SVM, KNN, and others, have demonstrated promising results in predicting diabetes accurately.
- The findings underscore the potential for enhancing medical decision-making and patient care through data-driven approaches.
- Future research in this domain should explore the integration of advanced classifiers, automation, and the application of machine learning techniques in other medical domains to further improve patient outcomes and advance the field of healthcare analytics.

2.2 NOVELTY

One crucial aspect is the application of hyperparameter auto-tuning to fine-tune the machine learning models. Hyperparameters are essential parameters in machine learning algorithms that significantly influence the model's performance. Properly optimizing these hyperparameters is essential for achieving the best predictive accuracy and generalization of the model.

In the reviewed papers, while various machine learning algorithms were employed, there was limited emphasis on hyperparameter optimization. This makes the work novel, in our study a systematic approach to automatically search for the best combination of hyperparameters for each algorithm was used. By doing so, the models are more likely to reach their optimal performance, leading to superior predictive accuracy.

Extra Trees is an ensemble learning method that builds multiple decision trees with randomized splits, providing increased diversity among the trees. The diversity in the trees helps to reduce overfitting and enhance the model's generalization ability.

Unlike the studies under review, which utilized various classifiers such as SVM, KNN, LR, DT, RF, and NB, none of them explored the potential of Extra Trees in this context. The application of Extra Trees in our study outperforms the traditional single classifiers used in previous studies.

Based on the findings from our study, the hyperparameter auto-tuning and the inclusion of Extra Trees have significantly improved the predictive accuracy of the diabetes prediction models.

In our study the results demonstrate that the combination of hyperparameter auto-tuning with various machine learning algorithms, including Extra Trees, yields superior accuracy compared to the models presented in the reviewed papers. By carefully optimizing the hyperparameters for each model and leveraging the ensemble power of Extra Trees, this approach surpasses the predictive accuracy achieved by the previously investigated classifiers.

Chapter 3

3. Methodology

The dataset that was used in the study was obtained from Kaggle. The dataset contained 789 datapoints. It had 9 columns in which 8 were the features and were used as input while 1 was the output. The names of the features are as follows:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age

Meanwhile output is the variable that defines whether diabetes is present or not. If the value of output is 1, that means the person has diabetes. Similarly, if the value is 0, it means the induvial is nondiabetic.

Figure 1 presents a detailed flowchart that outlines the step-by-step process of developing and evaluating a diabetes classification model. This visual representation provides a well-structured and clear overview of the essential stages involved in the entire process, facilitating a deeper comprehension of the workflow.

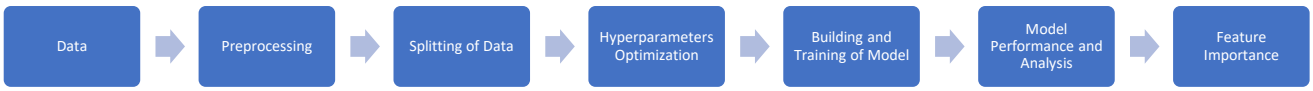

Figure 1 (Flow Chart)

3.1 ILLUSTRATION OF THE STEPS OF FLOWCHART:

3.1.1 Data Gathering

In this step, the data is collected from different sources. This step is very crucial as it is the foundation for training the machine learning model. The data is collected from various sources e.g. medical records.

3.1.2 Data Preprocessing

Data is usually in the raw form. It contains some errors as well as some missing entries. In this stage, the data is cleaned and obtained in the form in which the further analysis can be done. Preprocessing of data includes handling missing values, removing duplicate and handling the outliers.

3.1.3 Data Splitting

Splitting of dataset in 2 parts is the common practice as the evaluation of model's ability to predict is done by training the model on training data set and testing the model's performance on testing dataset.

3.1.4 Hyper-parameters Optimization

This stage includes the fine tuning of the parameters of machine learning algorithms. The best values of parameters are selected in this process.

3.1.5 Model Building and Training

This stage involves the building and training of machine learning models. The machine learning algorithms are used to build the models and then these models are trained on training data set to make predictions.

3.1.6 Model Performance and Analysis

After the building and training of model, their performance is analyzed. The evaluation is done by using the testing set to make predictions and then the evaluation matrix is calculated. In this stage the assessment of how well the model is performing is done. It helps in selecting the most suitable model.

3.1.7 Feature Importance

In this step the importance of each feature is determined. This step plays a very important role in identifying the variables that influence the prediction and accuracy of the model.

Overall, this flowchart offers a systematic and organized approach to developing and evaluating a diabetes classification model using machine learning techniques.

3.2 MACHINE LEARNING MODELS

Machine learning is the type of the artificial intelligence, and it can perform difficult tasks. It uses different techniques and can reduce the human effort (Bhat, 2022)

The models selected for analysis included Voting, Extra Trees, Bagging, Gradient Boosting, Logistic Regression (LR), and Random Forest Regression (RFR).

3.2.1 Random Forest Regression (RFR)

The random forest (RF) algorithm, initially proposed by (Breiman, 2001), is a machine learning algorithm widely used for predictive modeling tasks, including large-scale data prediction. The RF algorithm operates through four key processes: bootstrap resampling, random feature selection, out-of-bag (OOB) error estimation, and the construction of fully grown decision trees.

In the RF algorithm, multiple decision trees, known as weak learners, are generated from the training samples. Each tree is constructed using a randomly selected subset of the training samples obtained through bootstrap resampling, where some samples are left out and designated as out-of-bag samples. The decision trees are grown without pruning, using only the selected subset of training samples. Importantly, the RF algorithm randomly selects a small number of features from the available predictor features, rather than considering all features.

In RF algorithm a forest is constructed by iterating the training process. This results in the ensemble of decision trees. Other models that are based on trees mostly rely on cross validation for error estimation. Random forest on the other hand estimates the out of bag error during the forest's construction. The bagging ensemble is used to make predictions. It is done by collecting the predictions from the weak learners and then using the bagging ensemble method to combine them (Phyo, Byun, & Park, 2022).

In summary, the RF algorithm leverages bootstrap resampling, random feature selection, and an ensemble of decision trees to improve predictive performance. By estimating the out-of-bag error and aggregating predictions from multiple decision trees, the RF algorithm provides robust predictions for a variety of prediction tasks.

3.2.2 Gradient boosting machine

Gradient boosting machine (GBM) algorithms are globally known as prominent ensemble machine learning algorithms. The reason of this recognition is their adaptability and interpretability. It is achieved by the enhancing of weak learners and making them strong learners. In regression and classification tasks, the weights of the original training data are adjusted to iteratively train multiple weak learners. A weighted majority voting scheme is used to combine the predictions of these sequentially trained learners to get the final prediction (Phyo et al., 2022). This approach enables

GBM algorithms to effectively use the collective knowledge of the weak learners and produce accurate predictions.

3.2.3 Logistic Regression

It is a technique of regression analysis which is commonly used for categorical data classification. Even though the name has regression in it, it is still not a true regression model. It is a linear classification model and can be viewed as a log-linear classifier through which the probability of assigning a particular outcome to a given test is estimated. Although logistic regression assumes Gaussian data and independence among learned features, when it comes to classification, it remains a powerful tool. The versatility and effectiveness of logistic regression model are the reasons behind their popularity. In Logistic regression the relationship between one or more independent binary variables is established. In a situation involving binary variables, where linear regression analysis cannot be applied, logistic regressor can be used (Mushtaq et al., 2022).

3.2.4 Voting

A voting ensemble represents a machine learning strategy that doesn't depend solely on a single model but rather combines several models to boost the system's overall performance. This technique is versatile and can be applied to solve both classification and regression problems. It functions by consolidating the predictions generated by numerous methods and employs them collectively to reach a final decision. (Erdebilli & Devrim-İçtenbaş, 2022). By leveraging the collective wisdom of diverse models, the voting ensemble aims to improve the accuracy and robustness of the system's predictions.

3.2.5 Bagging

Bagging is a widely utilized ensemble technique in machine learning. It uses a process of bootstrap resampling to create multiple training sets. In bootstrap resampling, random samples are selected with replacement from the original dataset. It allows the generation of diverse subsets. Individual learning models within the ensemble structure are trained with these diverse subsets. The final prediction is made by combining the predictions obtained from each model. (Taser, 2021). It

enhances the overall predictive capability and robustness of the system by harnessing the collective power of multiple models.

3.2.6 Extra Tree

An element of randomness is added to the process by selecting a random threshold point for splitting during the construction of the trees.

Extra Tree is a type of ensemble classifier that constructs a larger set of binary decision trees. Unlike other ensemble methods, in Extra Tree each tree is built independently without any dependency on other trees. An element of randomness is added to the process by selecting a random threshold point for splitting during the construction of the trees. This randomness contributes to the diversity and variability of the resulting tree ensemble (Kharwar & Thakor, 2022).

3.3 DATA PREPROCESSING AND VISUALIZATION

The data quality plays a crucial role in the reliability of the results. The real-world data is raw and the presence of missing values, noise and inconsistencies are common. When the data quality is compromised, it can lead to unreliable results. Data preprocessing is an important step in machine learning. The high-quality data allows the generation of accurate and meaningful outcomes. The two main steps in data cleaning are as follows.

- Dealing with missing values
- Dealing with noisy data

As shown in the summary of dataset in table 1, the minimum values of some features is 0. It is indicating the presence of missing values. A thorough examination of the dataset showed that the number of zero values in the dataset was high and it had to be removed or dealt with.

Table 1 (Summary)

Column	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.845	3.370	0	1	3	6	17
Glucose	768	120.894	31.973	0	99	117	140.25	199
Blood Pressure	768	69.105	19.356	0	62	72	80	122
Skin Thickness	768	20.536	15.952	0	0	23	32	99
Insulin	768	79.799	115.244	0	0	30.5	127.25	846
BMI	768	31.993	7.884	0	27.3	32	36.6	67.1
Diabetes Pedigree Function	768	0.472	0.331	0.078	0.244	0.372	0.626	2.42
Age	768	33.241	11.760	21	24	29	41	81
Outcome	768	0.349	0.477	0	0	0	1	1

To address this issue, the rows containing zero values in the following columns were removed.

- Glucose
- Blood Pressure
- Skin Thickness
- BMI

The removal of rows with zero values helped ensure the quality and reliability of the dataset for subsequent analysis. This preprocessing step aimed to mitigate any potential bias or inaccuracies that could arise from including data points with zero values in columns such as Glucose, Blood Pressure, Skin Thickness and BMI.

After removing the rows with zero values, the dataset was reduced to 532 rows. The summary statistics of the preprocessed dataset is shown in table 2.

Table 2 (Summary)

Column	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	532	3.517	3.312	0	1	2	5	17
Glucose	532	121.030	31.999	56	98.75	115	141.25	199
Blood Pressure	532	71.506	12.310	24	64	72	80	110
Skin Thickness	532	29.182	10.524	7	22	29	36	99
Insulin	532	114.989	123.008	0	0	91.5	165.25	846
BMI	532	32.890	6.881	18.2	27.875	32.8	36.9	67.1
Diabetes Pedigree Function	532	0.503	0.345	0.085	0.259	0.416	0.659	2.42
Age	532	31.615	10.762	21	23	28	38	81
Outcome	532	0.333	0.472	0	0	0	1	1

The summary statistics highlight the central tendency, spread, and distribution of variables related to diabetes. By analyzing the summary, it was observed that the average pregnancy among the participants is 3.51. It indicated that the average number of pregnancies is around 3,4. The mean glucose level in the dataset is around 121.03. It provides an estimate blood glucose concentration. Further it was analyzed that the participants' average blood pressure reading is approximately 71.51 mmHg. The mean of skin thickness which provides the information about the skinfold measurements is 29.18 mm. These summary statistics help in the understanding of dataset’s characteristics, through which the relationship between different features can be explored.

The data after preprocessing was used for further analysis and building of model to predict the outcome of diabetes in individuals. After preprocessing the data consisted of 532 data points. The data was then divided into two following data sets.

- Training data set
- Testing data set

The training dataset is used to train the machine learning model. Once the model is trained, it is used to make predictions on the testing dataset.

3.4 DATA VISUALIZATION

3.4.1 CORRELATION

The valuable information about the relationship between different features in the dataset can be obtained by the analysis of correlation plot in figure 2. Understanding these correlations will give insights into potential feature importance for predicting whether an individual has diabetes or not.

Firstly, we observe a relatively strong positive correlation of approximately 0.50 between the feature "Glucose" and the target variable "Outcome." This suggests that higher levels of glucose in the blood are associated with a higher likelihood of having diabetes (Outcome=1). Glucose levels could be an essential indicator for diabetes prediction in this dataset.

Additionally, we find a moderately positive correlation of about 0.64 between the features "Age" and "Pregnancies." This indicates that as the age of individuals increases, the number of pregnancies they have tends to increase as well.

Furthermore, we observe moderate positive correlations between some other feature pairs. For instance, there is a correlation of approximately 0.25 between "Glucose" and "BMI" (Body Mass Index). This suggests some association between higher glucose levels and higher BMI values. Additionally, "BMI" and "Skin Thickness" exhibit a moderate positive correlation of approximately 0.65, indicating that individuals with higher BMI values tend to have increased skin thickness.

The correlation between "Age" and "Outcome" is also of interest, with a value of about 0.32. This suggests that older individuals may have a higher likelihood of having diabetes (Outcome=1). Age could be a relevant feature for diabetes prediction in this dataset.

While some features show weak positive correlations with each other, there is no strong correlation between any two features in the dataset. This absence of strong correlations implies that multicollinearity may not be a significant issue, making the features relatively independent of each other.

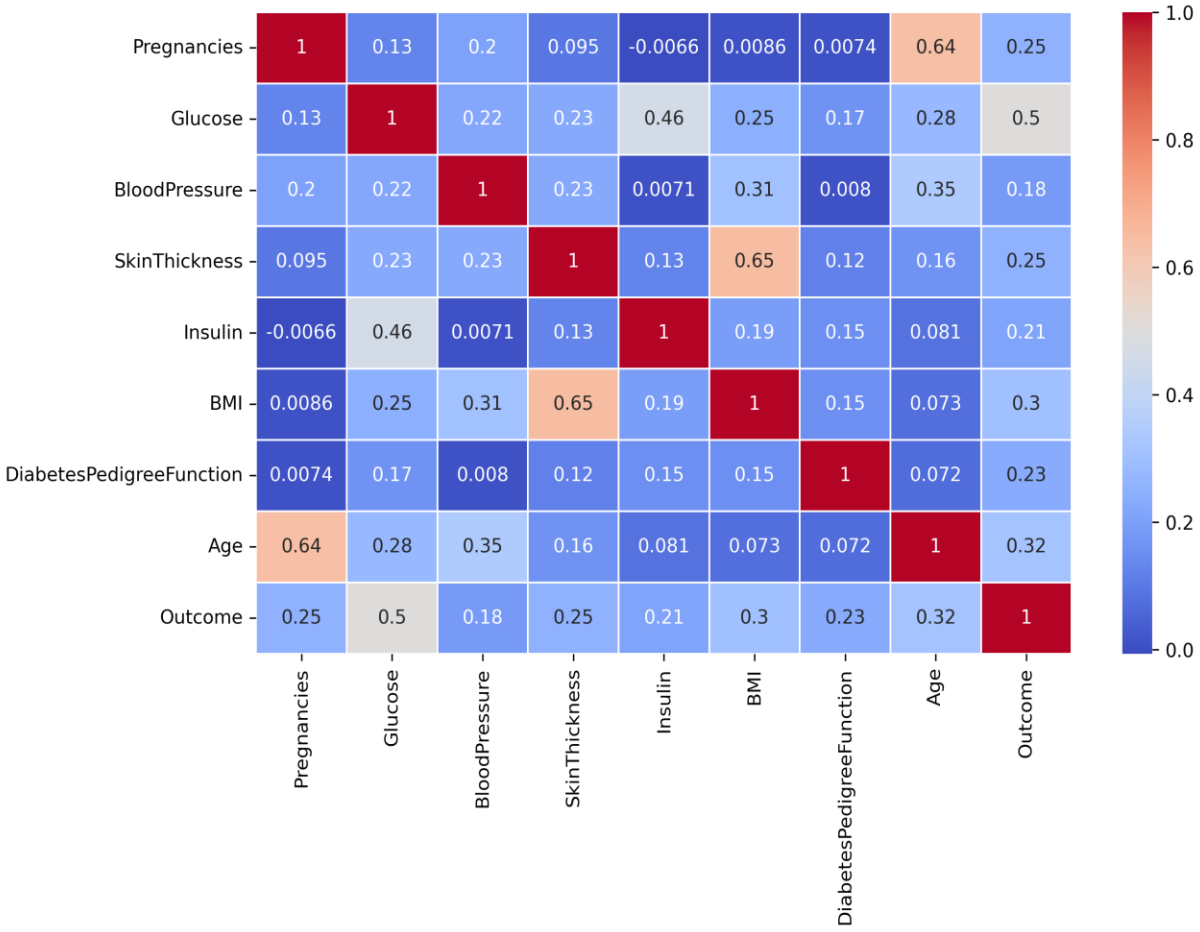

Figure 2 (Correlation Plot)

3.4.2 HISTOGRAM

The following histograms show the distribution of features in our data.

Pregnancies

The histogram for the Pregnancies features in figure 3 shows that most individuals in the dataset have a relatively low number of pregnancies, with the highest frequency observed in the range of 0 to 4 pregnancies. As the number of pregnancies increases, the frequency gradually decreases, indicating that the dataset contains a diverse range of individuals with different pregnancy counts.

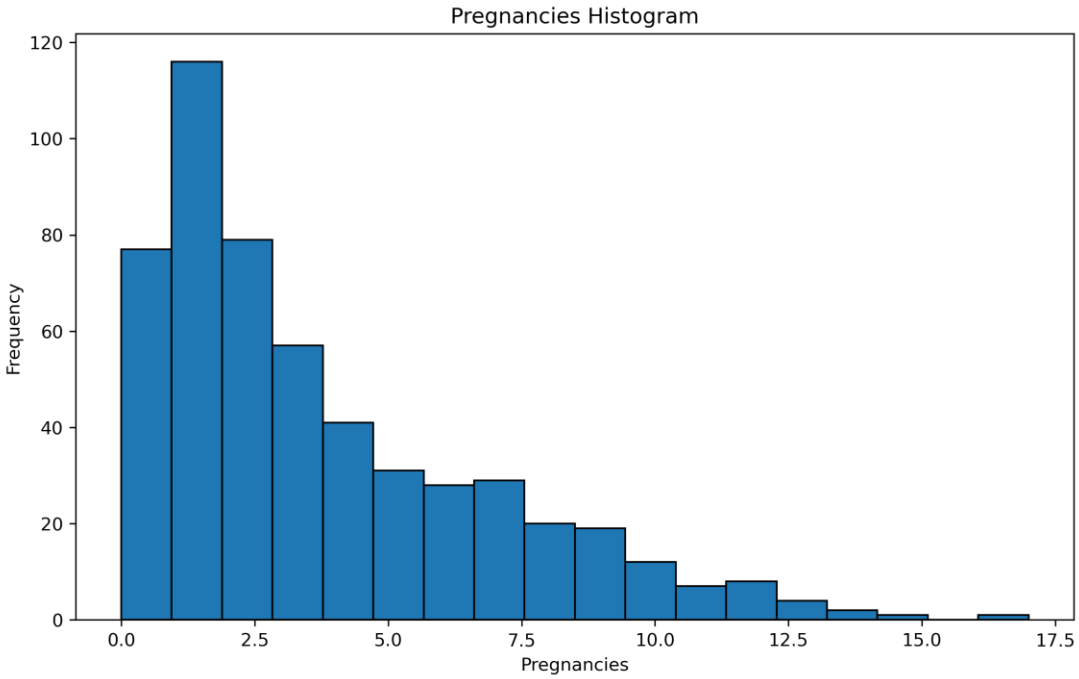

Figure 3 (Pregnancies Histogram)

Glucose

The histogram in figure 4 demonstrates the distribution of glucose levels in the dataset. It reveals that the most common glucose levels fall within the range of approximately 90 to 120. The peak around 100 suggests that a significant number of individuals have glucose levels near this value.

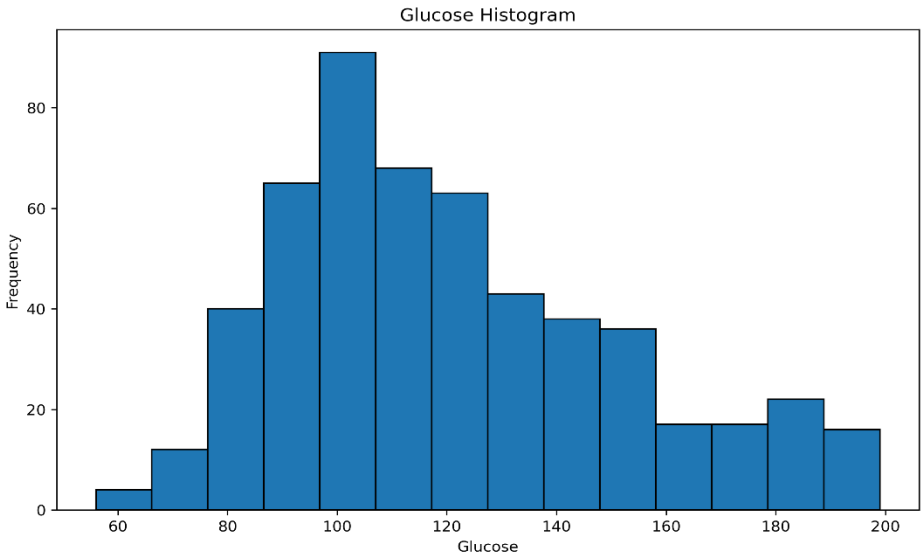

Figure 4 (Glucose histogram)

Blood Pressure

The histogram in figure 5 exhibits a relatively normal distribution, with a concentration of values around the 60 to 80 range. This indicates that the majority of individuals in the dataset have blood pressure readings in this range.

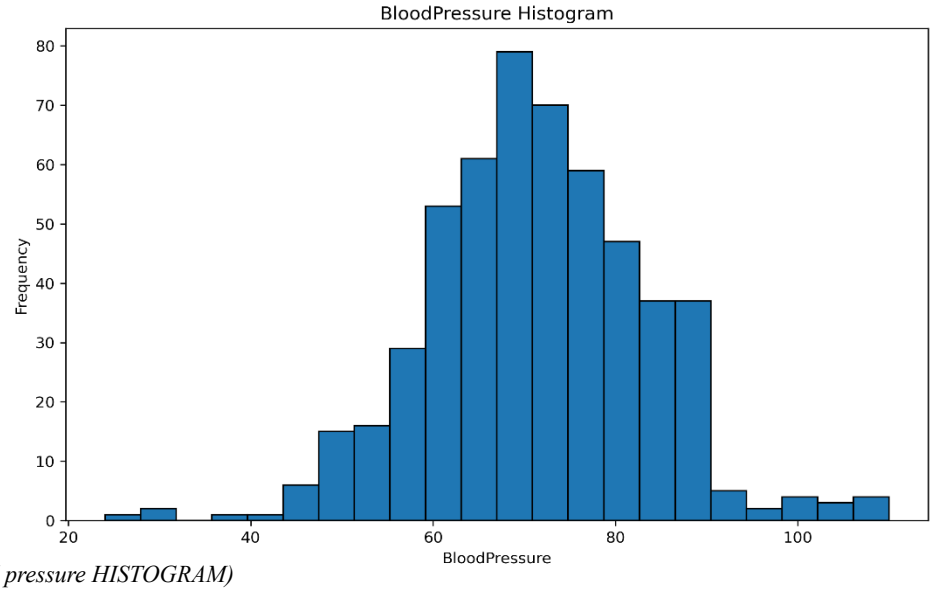

Figure 5 (blood pressure HISTOGRAM)

Skin Thickness

The histogram in figure 6 shows the frequency of different skin thickness measurements in the dataset. The data appears to be concentrated around the range of 20 to 40, with a slight decrease in frequency for higher skin thickness values.

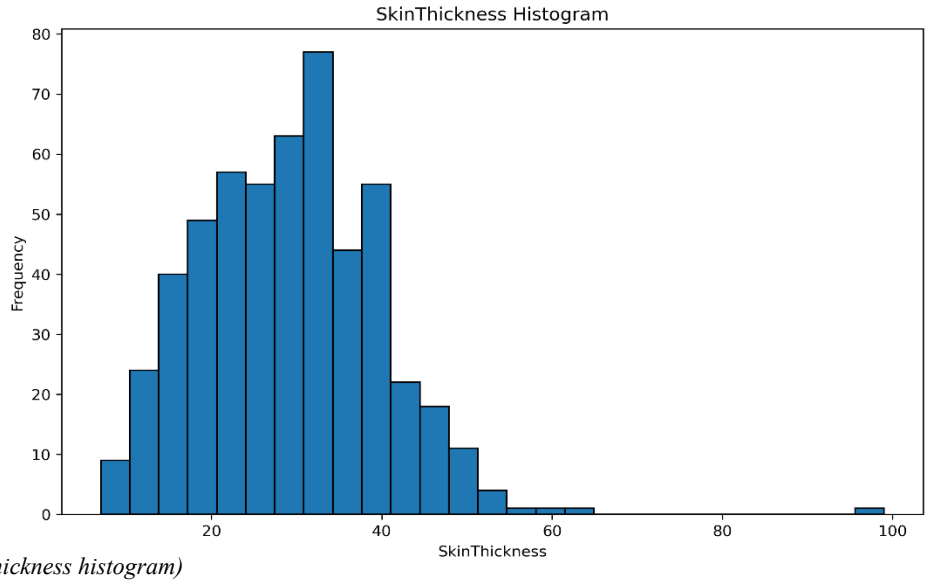

Figure 6 (skin thickness histogram)

Insulin

The histogram in figure 7 reveals that insulin levels are most commonly found in the range of 0 to 100. This suggests that a significant portion of the dataset consists of individuals with relatively low insulin levels.

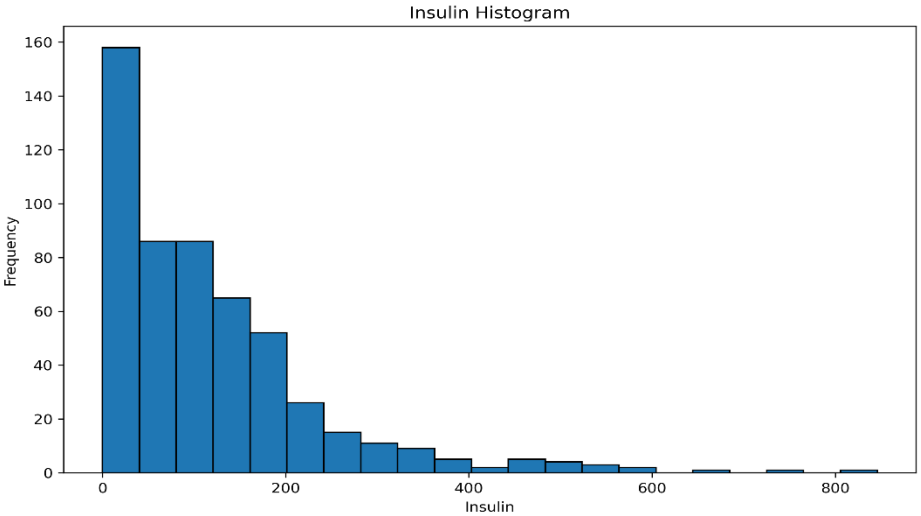

Figure 7 (insulin histogram)

BMI

The histogram in figure 8 displays the distribution of Body Mass Index values in the dataset. It shows a peak around 35, indicating that a considerable number of individuals have BMI values close to this range.

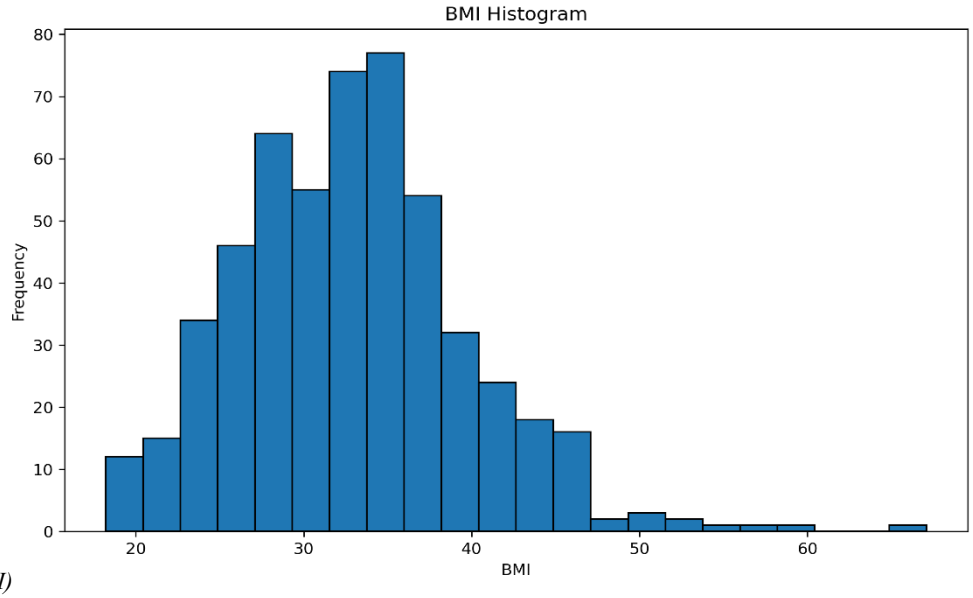

Figure 8 (BMI)

Diabetes Pedigree Function

The histogram in figure 9 shows a skewed distribution of values, with higher concentrations in the range of 0.1 to 0.4.

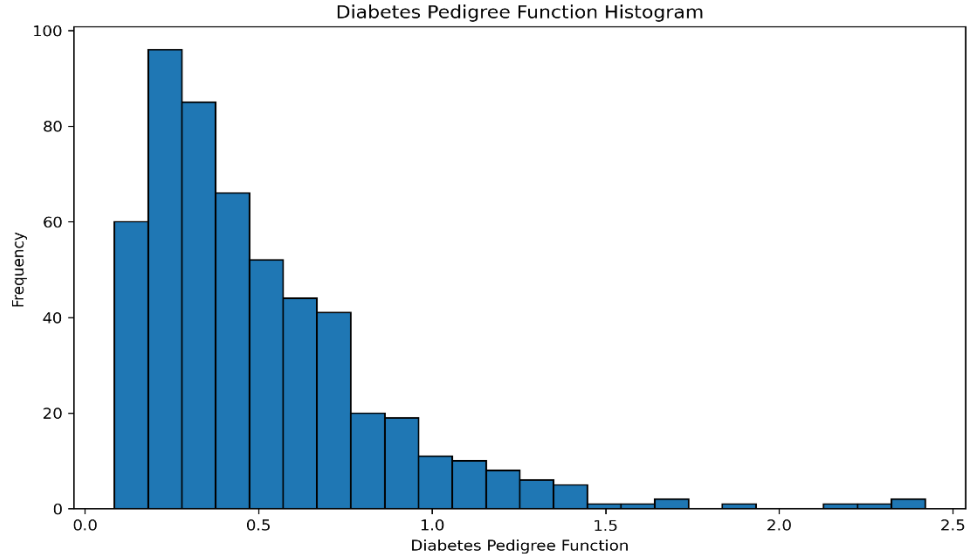

Figure 9 (Diabetes Pedigree function histogram)

Age

The histogram in figure 10 displays the distribution of individuals' ages in the dataset. It shows a relatively even distribution, with a peak in the range of 20 to 40, followed by a gradual decrease in frequency for older ages.

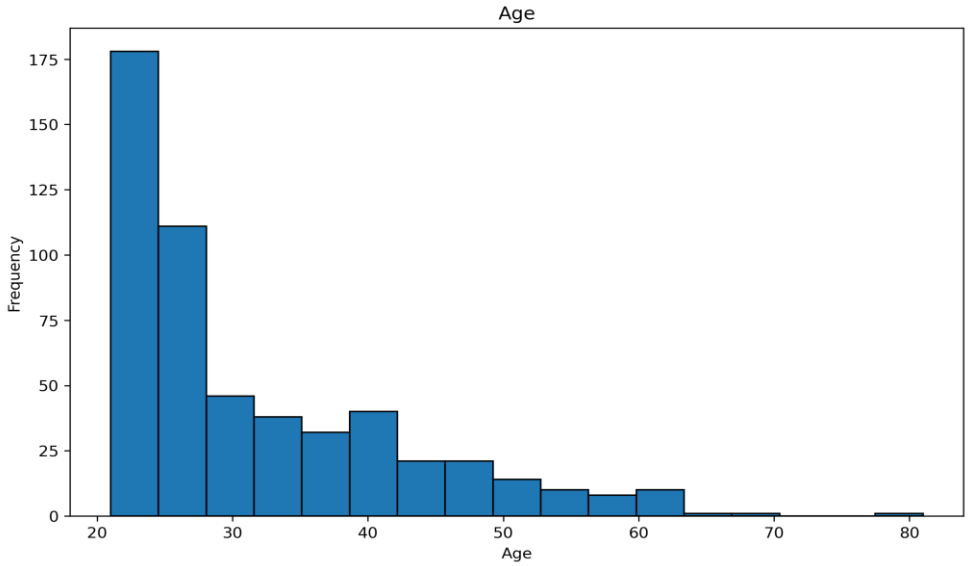

Figure 10 (AGE HISTOGRAM)

3.5 HYPERPARAMETER TUNING

Hyper-parameter tuning is a crucial step in evaluating machine learning models. It involves selecting the optimal values for certain parameters before initiating the machine learning task. These parameters are not learned from the data itself and must be fixed beforehand. The process of hyper-parameter tuning aims to find the best settings that result in the most accurate model fit. Since choosing the ideal hyper-parameter values can be challenging, grid search and random search algorithms are commonly employed. By fine-tuning the hyper-parameters, the accuracy of the machine learning classifier is enhanced, leading to improved performance in various tasks (Krishnamoorthi et al., 2022).

3.5.1 Randomized search CV

Randomized Search CV was utilized to optimize the hyperparameters for each model. Randomized Search CV is a technique that automates the process of hyperparameter tuning by randomly sampling hyperparameter combinations from a defined search space.

Several papers support the use of random search as an effective approach for hyperparameter optimization. (J. a. B. Bergstra, Yoshua, 2012)suggests that random search is a step towards formalizing hyperparameter optimization. (J. Bergstra & Bengio, 2012) provides empirical and theoretical evidence demonstrating the superiority of random search over grid search.

According to the findings of (J. Bergstra & Bengio, 2012) randomly chosen trials in hyperparameter optimization are more efficient than trials on a grid. Random search can find models that are as good as or even better than those obtained through grid search, while requiring only a fraction of the computation time. Even when granted the same computational budget, random search outperforms grid search by effectively exploring a larger, less promising configuration space. By leveraging random search, the study achieved efficient hyperparameter optimization, enabling the selection of optimal hyperparameters for each model.

The table 3 presents the best hyperparameters obtained through Randomized Search CV for each model used in the study. For the voting ensemble, the best parameters for the Random Forest, Gradient Boosting, and Bagging models are shown. Additionally, the optimal parameters for Extra Trees, Bagging, Logistic Regression, Random Forest Regression, and Gradient Boosting are provided. These hyperparameters were selected based on their performance in optimizing the respective models. By fine-tuning these parameters, the models were able to achieve better performance and improve the accuracy of diabetes prediction.

Table 3 (HYPER-PARAMETERS)

Model	Best Parameters
Voting (Random Forest)	'n_estimators': 100, 'min_samples_split': 4, 'max_depth': 4
Voting (Gradient Boosting)	'n_estimators': 50, 'min_samples_split': 4, 'max_depth': 3, 'learning_rate': 0.05
Voting (Bagging)	'n_estimators': 100, 'max_samples': 0.7, 'max_features': 0.5, 'base_estimator__max_depth': 3
Extra Trees	'n_estimators': 300, 'min_samples_split': 6, 'min_samples_leaf': 3, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': False
Bagging	'n_estimators': 200, 'max_samples': 0.9, 'max_features': 0.7
Logistic Regression	'solver': 'lbfgs', 'penalty': 'l2', 'C': 1
Random Forest Regression	'n_estimators': 200, 'min_samples_split': 4, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 5, 'bootstrap': False
Gradient Boosting	'n_estimators': 50, 'min_samples_split': 4, 'max_depth': 3, 'learning_rate': 0.05

3.6 EVALUATION METRICS

A comprehensive understanding of the performance of the different machine learning models can be gained through evaluation metrics.

3.6.1 Precision

The accuracy of the positive predictions made by the model are evaluated through precision. If the rate of false positive is low, it means the precision is high. It means that a larger proportion of instances predicted as positive are actually positive.

3.6.2 Recall (sensitivity)

Recall, also referred to as sensitivity or the true positive rate, and quantifies the fraction of real positive cases that are accurately identified as positive by the models.

3.6.3 Confusion matrix

A tabular representation that provides a detailed breakdown of the model's performance by comparing the predicted class labels with the actual class labels is known as the confusion matrix. It consists of four following components:

- True positives
- True negatives
- False positives
- False negatives.

Through confusion matrix a comprehensive analysis of the models' predictive capabilities can be done. It highlights the correct and incorrect predictions for each class.

3.6.4 Specificity

Specificity, in the context of these models, represents their capacity to accurately recognize negative instances. It gauges the ratio of correctly predicted negative cases relative to all the genuine negative cases.

3.6.5 Negative predicted value

Negative Predicted Value evaluates the model's precision in predicting negative outcomes. It quantifies the ratio of accurately predicted negative instances to the total instances predicted as negative.

3.6.6 Accuracy

Accuracy is a widely used metric that measures the overall correctness of the models' predictions. It calculates the proportion of correct predictions out of the total instances. In diabetes prediction,

accuracy provides a general assessment of the models' performance, indicating their ability to predict both positive and negative instances correctly.

3.7 COMPARISON OF DIFFERENT MODELS

3.7.1 Voting

The confusion matrix in table 4 represents the performance of the machine learning model on the test dataset. It shows the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. The Sensitivity, also known as recall for class 1 (diabetic instances), is 0.57576, indicating that the model correctly identified 57.58% of the actual positive instances. On the other hand, the Specificity is 0.93243, demonstrating that the model accurately predicted 93.24% of the actual negative instances (non-diabetic instances).

The Precision, also referred to as the positive predicted value for class 1, is 0.79167. This means that among all instances predicted as positive (class 1), approximately 79.17% were indeed true positive cases. The Negative Predicted Value is 0.83133, showing that around 83.13% of instances predicted as negative (class 0) were true negative cases.

The overall Accuracy of the model is 0.82243, indicating the proportion of correctly predicted instances out of the total instances in the dataset. An accuracy of 82.24% demonstrates the model's ability to make accurate predictions on the diabetes classification task.

In conclusion, the machine learning model demonstrated a good balance between sensitivity and specificity, achieving reasonably accurate predictions for both diabetic and non-diabetic instances.

Table 4(CONFUSION MATRIX VOTING)

	0	**1**
0	69	5
1	14	19

Table 5 (COMPARISON RATIOS VOTING)

Sensitivity (Recall for class 1)	**Specificity**	**Precision (Positive Predicted Value for class 1)**	**Negative Predicted Value**	**Accuracy**
0.57576	0.93243	0.79167	0.83133	0.82243

3.7.2 Logistic regression

The model achieved 66 correct predictions for non-diabetic instances (class 0) and 21 correct predictions for diabetic instances (class 1). However, it misclassified 8 non-diabetic instances as diabetic and 12 diabetic instances as non-diabetic.

The Sensitivity, also known as recall for class 1, is 0.63636, indicating that the model accurately identified 63.64% of the actual diabetic cases. On the other hand, the Specificity is 0.89189, demonstrating that the model correctly predicted 89.19% of the actual non-diabetic cases.

The Precision is 0.72414, signifying that approximately 72.41% of the instances predicted as diabetic (class 1) were indeed true positive cases. The Negative Predicted Value is 0.84615, showing that around 84.62% of instances predicted as non-diabetic (class 0) were true negative cases.

The overall Accuracy of the model is 0.81308, indicating that 81.31% of the instances in the dataset were correctly predicted by the model.

In summary, the machine learning model demonstrated moderate to good performance, with relatively high sensitivity and specificity, showcasing its ability to correctly identify both diabetic and non-diabetic cases. The precision and negative predicted value further validate the model's accuracy in making predictions for the diabetes classification task.

Table 6 (confusion matrix Logistic regression)

	0	**1**
0	66	8
1	12	21

Table 7 (Comparison ratios logistic regression)

Sensitivity (Recall for class 1)	Specificity	Precision	Negative Predicted Value	Accuracy
0.63636	0.89189	0.72414	0.84615	0.81308

3.7.3 Gradient boosting

The confusion matrix shows the model's performance in predicting non-diabetic (class 0) and diabetic (class 1) instances. Out of 74 non-diabetic instances, the model correctly predicted 67 instances (True Negatives), but it misclassified 7 instances as diabetic (False Positives). Similarly, out of 33 diabetic instances, the model correctly predicted 19 instances (True Positives), but it misclassified 14 instances as non-diabetic (False Negatives).

The Sensitivity (Recall) value of 0.57576 indicates that the model correctly identified 57.58% of the actual positive instances (diabetic cases). The Specificity value of 0.90541 demonstrates that the model accurately predicted 90.54% of the actual negative instances (non-diabetic cases).

The Precision value of 0.73077 signifies that among all instances predicted as positive (class 1), approximately 73.08% were indeed true positive cases. The Negative Predicted Value of 0.82716 shows that around 82.72% of instances predicted as negative (class 0) were true negative cases.

The overall Accuracy of the model is 0.80374, indicating the proportion of correctly predicted instances out of the total instances in the dataset. An accuracy of 80.37% demonstrates the model's ability to make accurate predictions on the diabetes classification task.

In summary, the machine learning model exhibited moderate performance with relatively balanced sensitivity and specificity, suggesting its ability to identify both diabetic and non-diabetic cases to some extent. The precision and negative predicted value further support the model's capability in making accurate predictions for the diabetes classification task. However, there is room for improvement in sensitivity to enhance the detection of true positive cases.

Table 8 (confusion matrix Gradient boosting)

	0	**1**
0	67	7
1	14	19

Table 9 (comparison ratios gradient boosting)

Sensitivity (Recall for class 1)	**Specificity**	**Precision**	**Negative Predicted Value**	**Accuracy**
0.57576	0.90541	0.73077	0.82716	0.80374

3.7.4 Bagging

Out of 74 non-diabetic instances, the model correctly predicted 67 instances (True Negatives), while it misclassified 7 instances as diabetic (False Positives). Similarly, out of 33 diabetic instances, the model correctly predicted 21 instances (True Positives), but it misclassified 12 instances as non-diabetic (False Negatives).

The Sensitivity (Recall) value of 0.63636 indicates that the model correctly identified approximately 63.64% of the actual positive instances (diabetic cases). The Specificity value of 0.90541 demonstrates the model's high accuracy in predicting approximately 90.54% of the actual negative instances (non-diabetic cases).

The Precision value of 0.75 signifies that among all instances predicted as positive (class 1), approximately 75% were indeed true positive cases. The Negative Predicted Value of 0.8481 shows that around 84.81% of instances predicted as negative (class 0) were true negative cases.

The overall Accuracy of the model is 0.82243, indicating the proportion of correctly predicted instances out of the total instances in the dataset. An accuracy of 82.24% demonstrates the model's ability to make accurate predictions on the diabetes classification task.

Table 10 (confusion matrix bagging)

	0	**1**
0	67	7
1	12	21

Table 11 (comparison ratios bagging)

Sensitivity	**Specificity**	**Precision**	**Negative Predicted Value**	**Accuracy**
0.63636	0.90541	0.75	0.8481	0.82243

3.7.5 Extra trees

The model exhibited a sensitivity (recall for class 1) of 0.6060, indicating that it correctly identified approximately 60.60% of the actual diabetic cases. Moreover, the specificity value of 0.93243 demonstrated the model's high accuracy in predicting non-diabetic instances, achieving approximately 93.24% correctness in this regard. The model's precision stood at 0.80, indicating that about 80% of instances predicted as positive (diabetic cases) were indeed true positive cases. Additionally, the negative predicted value of 0.8414 revealed that around 84.14% of instances predicted as negative (non-diabetic cases) were true negative cases. Furthermore, the model's overall accuracy of 0.8317 highlighted its ability to make accurate predictions, with an 83.17% correctness rate across all instances in the dataset.

From the confusion matrix, we can observe that the model made 69 correct predictions for non-diabetic instances (true negatives) and 20 correct predictions for diabetic instances (true positives). However, it also misclassified 5 instances of class 0 as diabetic cases (false positives) and 13 instances of class 1 as non-diabetic cases (false negatives).

Overall, these results demonstrate the model's potential in identifying both diabetic and non-diabetic cases, with a good balance between sensitivity and specificity. The high precision score indicates that the model can accurately classify positive instances, while the negative predicted value reflects its ability to predict negative instances correctly.

Table 12 (Confusion matrix extra trees)

	0	**1**
0	69	5
1	13	20

Table 13 (comparison ratios extra trees)

Sensitivity	**Specificity**	**Precision**	**Negative Predicted Value**	**Accuracy**
0.6060	0.93243	0.80	0.84146	0.8317

3.7.6 RFR

The model achieved a sensitivity (recall for class 1) of 0.60606, indicating that it correctly identified around 60.61% of the actual diabetic cases. Moreover, the specificity value of 0.90541 demonstrated the model's high accuracy in predicting non-diabetic instances, achieving approximately 90.54% correctness in this regard. The precision score of 0.74074 reflected that about 74.07% of instances predicted as positive (diabetic cases) were indeed true positive cases. Additionally, the negative predicted value of 0.8375 revealed that around 83.75% of instances predicted as negative (non-diabetic cases) were true negative cases. Furthermore, the model's overall accuracy was 0.81308, indicating its ability to make accurate predictions, achieving an 81.31% correctness rate across all instances in the dataset.

From the confusion matrix, we observe that the model made 67 correct predictions for non-diabetic instances (true negatives) and 20 correct predictions for diabetic instances (true positives). However, it also misclassified 7 instances of class 0 as diabetic cases (false positives) and 13 instances of class 1 as non-diabetic cases (false negatives).

Overall, these results highlight the model's potential in identifying both diabetic and non-diabetic cases, with good specificity and precision.

Table 14 (confusion matrix rfr)

	0	**1**
0	69	5
1	13	20

Table 15 (comparison ratios RFR)

Sensitivity	**Specificity**	**Precision**	**Negative Predicted Value**	**Accuracy**
0.6060	0.93243	0.80	0.84146	0.8317

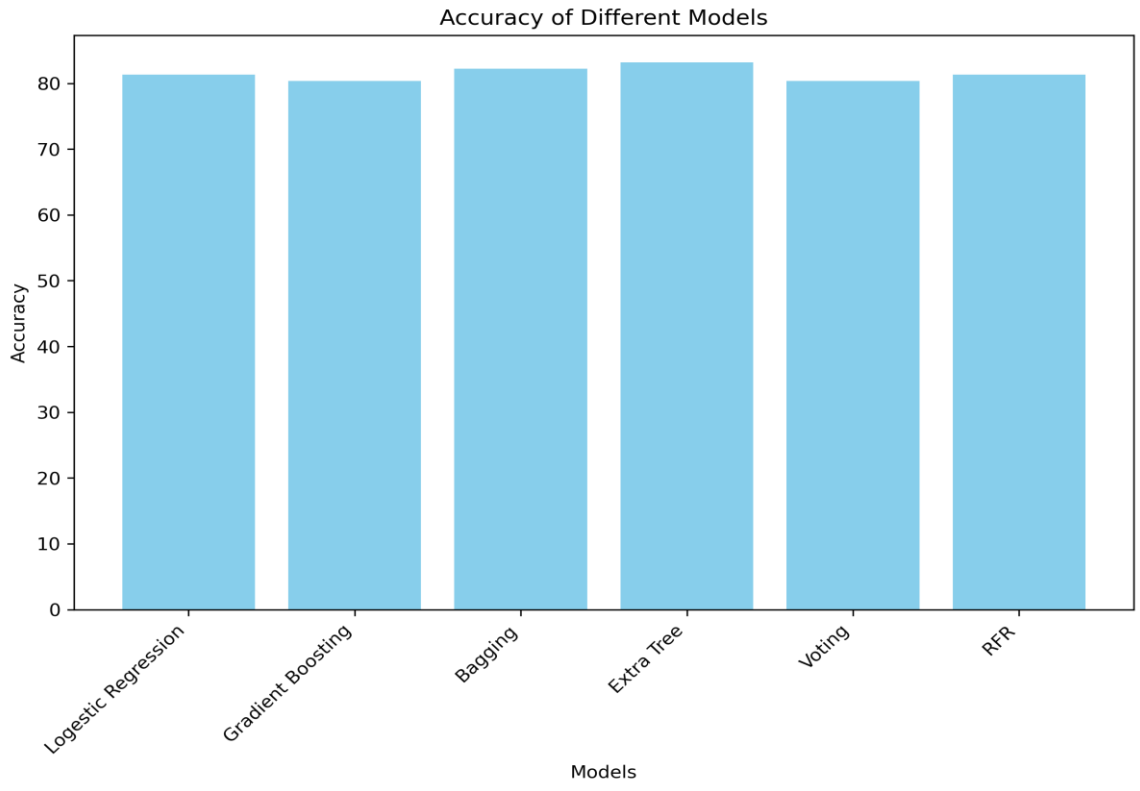

Figure 11 comparison of accuracies between different models)

Chapter 4

4. Results Discussion

Comparing these metrics can help us understand how well each model performs on the task of diabetes prediction.

For instance, we can observe that Bagging and linear regression models have the highest Sensitivity, indicating they are better at correctly identifying diabetic instances. However, Extra Trees model has the highest Specificity, indicating it is better at correctly identifying non-diabetic instances.

In terms of Precision, the Extra tree model has the highest score, suggesting that among all instances predicted as positive, a higher percentage were true positive cases. On the other hand, the Bagging model has the highest Negative Predicted Value, indicating its accuracy in predicting non-diabetic instances.

Regarding Accuracy, Extra Tree models have the highest score, indicating it's overall correctness in predicting diabetes. Moreover, all the models have accuracy in the range of 80%-83%. Overall, these results highlight the model's potential in identifying both diabetic and non-diabetic cases with good accuracy.

Chapter 5

5. Feature Importance

Feature importance is a crucial aspect of predictive modeling, especially in classification tasks. It involves assigning scores to input features to understand their significance in predicting the target variable. This technique provides valuable insights into how each feature contributes to the model's performance. By enumerating feature importance, the effectiveness and efficiency of the predictive model can be enhanced, leading to more accurate and informed predictions (Debjit et al., 2022)

5.1 SHAP

SHAP was introduced by Lundberg and Lee (Scott Lundberg, 2017). It is an interpretable method grounded in game theory and Shapley value(Shapley, 1953). The incremental impact of each input feature on the model's predictions is donated by The Shapley value. A weighted linear regression model is established by Utilizing the Shapley values (Molnar, 2020):

$$\widehat{y}_i = y_{\text{base}} + \sum_{j=1}^{\text{M}} \phi_j^{(i)}$$

Here, y_base represents the baseline prediction of the entire model, typically set as the mean of all predicted values.

To assess the overall significance of each input feature, we compute the average absolute Shapley value across all instances (n) in the dataset (Molnar, 2020):

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|$$

This allows us to quantify the relative importance of each feature in influencing the model's predictions.

In summary, SHAP provides an interpretable approach to understand the contribution of individual features in predicting the target variable. By incorporating Shapley values and analyzing feature importance, the model's behavior becomes more transparent, enhancing trust and interpretability in the predictive modeling process.

In this study, we employed the SHAP summary plot to analyze the features' importance. This plot offers two main advantages: feature ranking and the impact of each feature. The y-axis represents the feature ranking, with higher importance placed at the top and lower importance at the bottom. On the other hand, the x-axis displays the SHAP values, which indicate the effect of each feature on the model's predictions. This plot provides valuable insights into the relative importance of features.

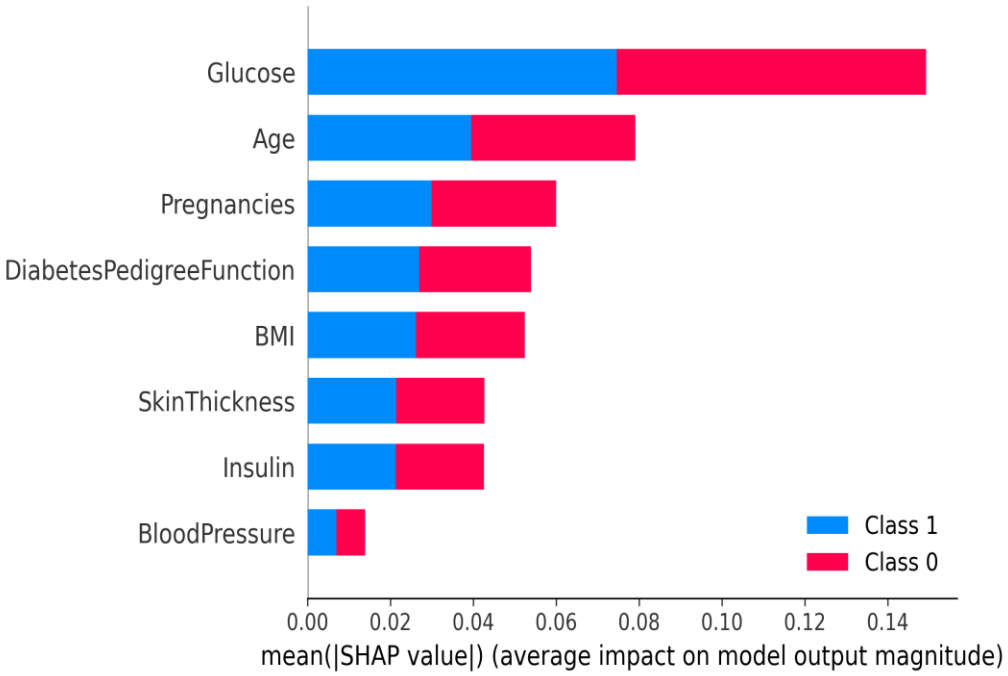

Figure 12 (SHAP FEATURE IMPORTANCE)

According to (Vaishali, Sasikala, Ramasubbareddy, Remya, & Nalluri, 2017), when medical experts were consulted for decision-making on diabetes, they identified Glucose as the most critical feature. In SHAP feature importance analysis, Glucose was also identified as the top influential feature, followed by Age. These findings highlight the consensus between domain experts and the model's interpretability, both recognizing Glucose's significant role in predicting diabetes outcomes. This alignment enhances the credibility and relevance of the model's insights for medical decision-making.

Chapter 6

6. Conclusion

Six machine learning models, namely Voting, Extra Trees, Bagging, Gradient Boosting, Logistic Regression (LR), and Random Forest Regression (RFR), were implemented. Random search CV was used to optimize these models. The performance of models in prediction of diabetes is evaluated based on various metrics.

Extra trees model performed the best among all the models. It displayed the highest accuracy when it comes to predicting diabetic and non-diabetic cases. Good balance between sensitivity and specificity was demonstrated by the model. It showcased the ability of the model in identifying both positive and negative instances.

SHAP summary plots were used for the feature importance analysis. It revealed that "Glucose" and "Age" were the features with most influence in the diabetic prediction. Its significance as a crucial risk factor was highlighted by the strong correlation between glucose and the outcome of diabetes. Potential role of" age" in the development and progression of disease is also suggested by the positive correlation of age with the outcome of diabetes.

The importance of these features in the diagnostic process was highlighted by these findings. This knowledge can be used by the medical practitioners to prioritize the monitoring and management of these critical factors in individuals suffering from diabetes. It can lead to improves patient outcomes.

In conclusion, the findings of this study contribute valuable knowledge to the field of healthcare and diabetes management, demonstrating the power of machine learning in predicting and diagnosing medical conditions. The optimized models and feature importance analysis open avenues for future research in healthcare decision support systems and personalized medicine, advancing the application of machine learning in improving public health and disease prevention.

6.1 FUTURE RESEARCH

Looking ahead, the domain of healthcare analytics holds promising opportunities for future research. The integration of advanced classifiers and automation, along with the application of

machine learning techniques, stands as a potential avenue to further enhance patient outcomes and propel the field forward. Enlarging the existing dataset and delving into additional potent machine learning algorithms could pave the way for improved prediction accuracy, ultimately leading to more precise and tailored patient care. A critical area of investigation is the realm of personalized medicine. Exploring the customization of prediction models based on individual patient characteristics and medical histories has the potential to revolutionize diabetes management and treatment plans. By tailoring interventions to unique patient profiles, healthcare practitioners can optimize care strategies for better results.

Additionally, the real-world impact of machine learning models can be magnified through clinical integration. Collaborative efforts between researchers and medical professionals to seamlessly incorporate these models into clinical practice could streamline diagnostics and interventions, ensuring timely and effective healthcare delivery.

Expanding the scope of research by incorporating larger and more diverse datasets holds immense potential. Including data from various demographic backgrounds, ethnicities, and geographical regions could bolster the models' reliability and applicability, contributing to more robust prediction models. Diversifying the toolbox of machine learning algorithms is another avenue to explore. Beyond the algorithms examined in this study, delving into advanced methods such as neural networks, support vector machines, or deep learning architectures could unlock new dimensions of predictive accuracy and offer fresh insights into diabetes prediction.

In essence, the future research landscape in healthcare analytics offers a spectrum of possibilities. By embracing advanced technologies, tailoring interventions, collaborating closely with healthcare practitioners, and expanding the datasets and algorithmic approaches, researchers can collectively drive innovation in diabetes management, advance patient care, and make meaningful strides towards a healthier future.

References

Tama, B. & Hermansyah, H. (2011). An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 9*, 287. doi:10.12928/telkomnika.v9i2.699

Anderson, J., Parikh, J., Shenfeld, D., Ivanov, V., Marks, C., Church, B., Rublee, D. (2015). Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. *J Diabetes Sci Technol, 10*(1), 6-18. doi:10.1177/1932296815620200

Barik, S., Mohanty, S., Mohanty, S., & Singh, D. (2021). Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques. In (pp. 399-409).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res., 13*(null), 281–305.

Bergstra, J., Yoshua, B. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*.

Bilous, R. D. R. I. I. (2021). *HANDBOOK OF DIABETES*. [S.l.]: JOHN WILEY & SONS [S.l.].

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Bruen, D., Delaney, C., Florea, L., & Diamond, D. (2017). Glucose sensing for diabetes monitoring: recent developments. *Sensors, 17*(8), 1866.

Debjit, K., Islam, M. S., Rahman, M. A., Pinki, F. T., Nath, R. D., Al-Ahmadi, S., . . . Awal, M. A. (2022). An Improved Machine-Learning Approach for COVID-19 Prediction Using Harris Hawks Optimization and Feature Analysis Using SHAP. *Diagnostics, 12*(5), 1023. Retrieved from https://www.mdpi.com/2075-4418/12/5/1023

Durairaj, M. (2015). *PREDICTION OF DIABETES USING BACK PROPAGATION ALGORITHM*.

Erdebilli, B., & Devrim-İçtenbaş, B. (2022). Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. *Mathematics, 10*(14), 2466. Retrieved from https://www.mdpi.com/2227-7390/10/14/2466

Healthcare Engineering, J. o. (2023). Retracted: A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering, 2023*, 9872970. doi:10.1155/2023/9872970

Jayalakshmi, T., & Santhakumaran, A. (2010, 9-10 Feb. 2010). *A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks.* Paper presented at the 2010 International Conference on Data Storage and Data Engineering.

Kharwar, A., & Thakor, D. (2022). An Ensemble Approach for Feature Selection and Classification in Intrusion Detection Using Extra-Tree Algorithm. *International Journal of Information Security and Privacy, 16*, 21. doi:10.4018/IJISP.2022010113

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering, 2022*, 1684017. doi:10.1155/2022/1684017

Lesmana, I. P. D., Purnama, I. K. E., & Purnomo, M. H. (2011). Abnormal condition detection of pancreatic Beta-cells as the cause of Diabetes Mellitus based on iris image. *2011 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*, 150-155.

Alam, T., Iqbal, M., Ali, Y., Wahab, A., Ijaz, S., Baig, T., . . . Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked, 16*, 100204. doi:10.1016/j.imu.2019.100204

Molnar, C. (2020). *Interpretable machine learning*: Lulu. com.

Muller, P. S., Sundaram, S., Nirmala, M., & Nagarajan, E. (2015). Application of computational technique in design of classifier for early detection of gestational diabetes mellitus. *Applied Mathematical Sciences, 9*, 3327-3336. doi:10.12988/ams.2015.54319

Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems, 2022*, 6521532. doi:10.1155/2022/6521532

Phyo, P.-P., Byun, Y.-C., & Park, N. (2022). Short-Term Energy Forecasting Using Machine-Learning-Based Ensemble Voting Regression. *Symmetry, 14*(1), 160. Retrieved from https://www.mdpi.com/2073-8994/14/1/160

Pradhan, M., & Bamnote, G. (2014). Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. *Advances in Intelligent Systems and Computing, 327*, 763-770. doi:10.1007/978-3-319-11933-5_86

Richard, I. G., Holt, C. C., Flyvbjerg, J. (2011).

Textbook of Diabetes.

Roglic, G. (2016). WHO Global report on diabetes: A summary. *Int J Non-Commun Dis, 1*. doi:10.4103/2468-8827.184853

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, 6-7 Sept. 2018). *Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.* Paper presented at the 2018 24th International Conference on Automation and Computing (ICAC).

Sarwar, N., Gao, P., Seshasai, S. R., Gobin, R., Kaptoge, S., Angelantonio, E., Danesh, J. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet, 375*(9733), 2215-2222. doi:10.1016/s0140-6736(10)60484-9

Lundberg, S.-I. L. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved from https://arxiv.org/abs/1705.07874

Shan, R., Sarkar, S., & Martin, S. S. (2019). Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. *Diabetologia, 62*(6), 877-887. doi:10.1007/s00125-019-4864-7

Shapley, L. S. (1953). 17. A Value for n-Person Games. In K. Harold William & T. Albert William (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307-318). Princeton: Princeton University Press.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science, 132*, 1578-1585. doi:https://doi.org/10.1016/j.procs.2018.05.122

Taser, P. Y. (2021). Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction. *Proceedings, 74*(1), 6. Retrieved from https://www.mdpi.com/2504-3900/74/1/6

Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017, 29-31 Oct. 2017). *Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset.* Paper presented at the 2017 International Conference on Computing Networking and Informatics (ICCNI).

Vijayan, V., & Anjali, C. (2015). *Prediction and diagnosis of diabetes mellitus — A machine learning approach*.

Wibawa, A., & Hery Purnomo, M. (2006). *Early Detection on the Condition of Pancreas Organ as the Cause of Diabetes Mellitus by Real Time Iris Image Processing*.

Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., . . . Zhang, P. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice, 162*, 108072.

Zhang, B., Vijaya kumar, B. V., & Zhang, D. (2014). Noninvasive diabetes mellitus detection using facial block color with a sparse representation classifier. *IEEE Trans Biomed Eng, 61*(4), 1027-1033. doi:10.1109/tbme.2013.2292936