A PROJECT REPORT

on

Diabetes Prediction Using Different Machine Learning Techniques

Submitted in partial fulfillment of the requirements

for the award of the degree of

Bachelor of Technology

in

Information Technology

by

Rahul Kumar Sachdeva (Roll. No. 2000970130088)

Rahul Lodhi (Roll. No. 2000970130089)

Prakhar Kumar Singh (Roll. No. 2000970130078)

Group No.: 23IT709

Under the Supervision of

Ms. Anam Khan

Galgotias College of Engineering & Technology

Greater Noida 201306

Uttar Pradesh, INDIA

Affiliated to

Dr. A.P.J Abdul Kalam Technical University

Lucknow

May 2024

GALGOTIAS COLLEGE OF ENGINEERING & TECHNOLOGY GREATER NOIDA - 201306 , UTTAR PRADESH, INDIA.

DECLARATION

We hereby declare that the project work presented in this project report entitled **"Diabetes Prediction Using Different Machine Learning Techniques"** in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Information Technology, submitted to A.P.J. Abdul Kalam Technical University, Lucknow, is based on my work carried out at Department of Information Technology, Galgotias college of engineering and technology, Greater Noida. The work contained in the report is original and project work reported in this report has not been submitted by me/us for the award of any other degree or diploma.

Signature:

Name: Rahul Kumar Sachdeva

Roll No: 2000970130088

Signature:

Name: Rahul Lodhi

Roll No: 2000970130089

Signature:

Name: Prakhar Kumar Singh

Roll No: 2000970130078

Date: 08/05/2024

Place: Greater Noida

ACKNOWLEDGEMENT

We want to give special thanks to our Project coordinator, **Dr. Javed Miya** and our project guide, **Ms. Anam Khan** for the timely advice and valuable guidance during designing and implementation of this project work.

We also want to express our sincere thanks and gratitude to **Prof**. **Dr. Sanjeev Kumar Singh**, Head of Department (HOD), and Information Technology Department for providing us with the facilities and for all the encouragement and support.

Finally, we express our sincere thanks to all staff members in the department of Information Technology branch for all the support and cooperation.

Rahul Kumar Sachdeva
(2000970130088)

Rahul Lodhi
(2000970130089)

Prakhar Kumar Singh
(2000970130078)

GALGOTIAS COLLEGE OF ENGINEERING & TECHNOLOGY GREATER NOIDA - 201306 , UTTAR PRADESH, INDIA.

CERTIFICATE

This is to certify that the project report entitled “DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES” submitted by Rahul Kumar Sachdeva (2000970130088), Rahul Lodhi (2000970130089), Prakhar Kumar Singh (2000970130078) to the Galgotias College of Engineering and Technology, Uttar Pradesh in partial fulfilment for the award of Degree of Bachelor of Technology in Information Technology is a bonafide record of the project work carried out by them under my supervision during the year 2023-2024.

Ms. Anam Khan
(Assistant Professor)
Deptt. of IT

Dr. Sanjeev Kumar Singh
(HOD IT)

ABSTRACT

Diabetes mellitus, particularly type-2 diabetes, represents a substantial portion of global diabetes cases, exerting significant pressure on healthcare systems worldwide[1]. This metabolic disorder, marked by inadequate insulin production or response leading to heightened blood sugar levels, is linked with numerous health complications, including heart and kidney diseases. Conventional diagnosis involves frequent visits to diagnostic centers, consuming both time and financial resources. However, the advent of machine learning technologies offers a promising solution to this challenge. By leveraging advanced data processing techniques, machine learning models can predict the onset of diabetes, enabling early intervention and improved patient outcomes. This research aims to support physicians in the timely identification and effective diagnosis of type 2 diabetes. Supervised machine learning techniques were executed to “Pima dataset”, utilizing six predictors to develop predictive models. The study employs classification algorithms such as SVM, KNN, Naive Bayes, Gradient Boosting Classifier, Logistic Regression, and Random Forest. Results indicate promising accuracy levels across the models, with Support Vector Machine achieving 76%, KNN 80%, Naive Bayes 76%, Gradient Boosting Classifier 85%, Logistic Regression 80%, and Random Forest 96%. These outcomes underscore the efficacy of machine learning approaches in diabetes prediction, offering a valuable tool for healthcare professionals to enhance diagnosis and patient care. This study advances the creation of accurate and effective type 2 diabetes diagnosis tools by utilizing machine learning's predictive capabilities. The findings highlight the potential of machine learning algorithms to analyze large volumes of diabetes-related data, enabling proactive healthcare interventions and ultimately improving patient outcomes. Moreover, the study underscores the importance of ongoing research and confirmation efforts to guarantee the dependability and effectiveness of machine learning in clinical settings.

Keywords— ***Machine Learning, SVM, KNN, Naive Bayes, Gradient Boosting Classifier, Random Forest Algorithm***

CONTENTS

LIST OF FIGURES

ABBREVIATIONS

NCD	Non-Communicable Disease
IDF	International Diabetes Federation
CBT	Cognitive-behavioral therapy
CGM	Continuous glucose monitoring
ICSDF	International Conference on Smart Data Intelligence
SVM	Support Vector Machines
CART	Classification and Regression Tree
KNN	k-Nearest Neighbor
IARJSET	International Advanced Research Journal in Science, Engineering and Technology.
ICCCI	International Conference on Computer Communication and Informatics
GLP-1	Glucagon-Like Peptide-1
SGLT 2	Sodium-Glucose Cotransporter 2
ICSMDI	International Conference on Sustainable Materials, Design and Innovations
UCI	University Of California
ANN	Artificial Neural Network
ROC	Receiver Operating Characteristic
LGBM	Light Gradient Boosting Classifier
RF	Random Forest
GB	Gradient Boosting
AUC	Area Under the Curve

CHAPTER 1

INTRODUCTION

Diabetes, a chronic metabolic NCD, poses a significant global health challenge, with an estimated 415 million cases worldwide, projected to rise to 642 million by 2040. It is characterized by abnormally high blood glucose levels, primarily caused by insulin dysfunction. While the human body requires glucose for energy, inefficient insulin production or utilization leads to hyperglycemia, the hallmark of diabetes. Type 2 diabetes, the most prevalent form, often stems from a combination of unhealthy lifestyle habits and insufficient physical activity. Consequently, glucose remains in the bloodstream, contributing to various systemic complications affecting the kidneys, eyes, neurological system, and arteries. Hyperglycemia, a key feature of diabetes, can result from insulin deficiency, as observed in type 1 diabetes, where pancreatic beta cells fail to produce adequate insulin. Type 2 diabetes, on the other hand, involves insulin resistance, where the body cannot efficiently utilize the insulin it produces. It is essential to comprehend the complex nature of diabetes and the underlying mechanisms in order to create preventative and management plans that work. This abstract provides a concise overview of diabetes, highlighting its global prevalence, etiology, and the distinction between type 1 and type 2 diabetes.

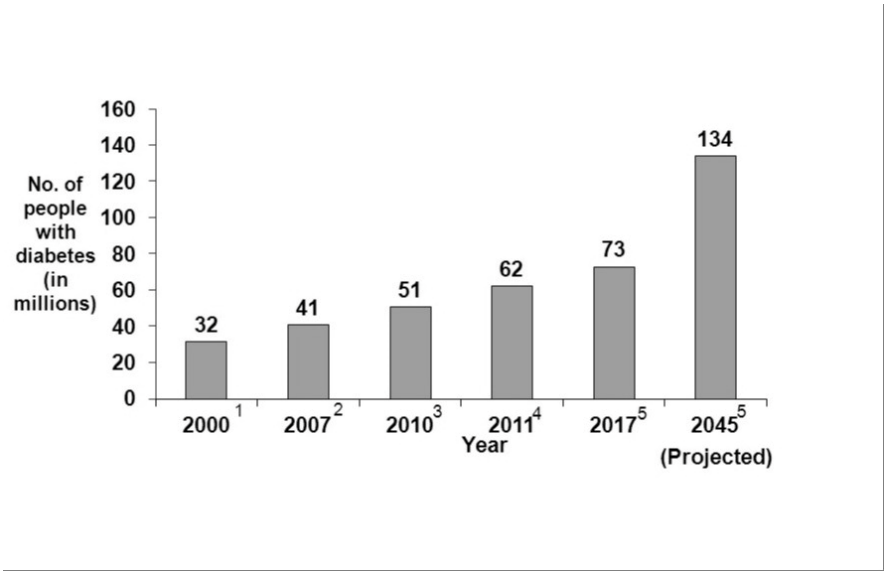


Fig: 1.1: Exploring Diabetes Epidemic in India

1.1 Global Prevalence and Impact

The global prevalence of diabetes has reached alarming proportions, affecting individuals across diverse demographics and geographic regions. The IDF reports that diabetes accounts for a considerable burden on healthcare systems, with substantial economic costs associated with its management and complications. The disease is a leading cause of morbidity and mortality, significantly impacting quality of life. Regions with rapid urbanization and lifestyle changes, particularly in low- and middle-income countries, have seen the most substantial increases in diabetes prevalence. This rising trend necessitates urgent action from healthcare providers, policymakers, and the public to address both the prevention and management of diabetes.

1.2 Understanding Diabetes: A Metabolic Disorder

Diabetes is characterized by abnormally high blood glucose levels, primarily due to insulin dysfunction. Insulin, a hormone produced by the pancreatic beta cells, is crucial for regulating blood glucose levels by facilitating the uptake of glucose into cells for energy production. In the absence or inefficiency of insulin, glucose remains in the bloodstream, leading to hyperglycemia, the hallmark of diabetes. The condition is chronic and requires ongoing medical attention and self-management to prevent acute complications and reduce the risk of long-term complications.

1.3 Types of Diabetes

There are several types of diabetes, with Type 1 and Type 2 being the most common.

Type 1 Diabetes: Type 1 diabetes, also known as juvenile-onset or insulin-dependent diabetes, is an autoimmune condition where the body's immune system attacks and destroys the insulin-producing beta cells in the pancreas. This leads to a deficiency of insulin, requiring individuals to depend on external insulin administration for survival. Although it can develop at any age, it predominantly affects children and young adults. The exact cause of the autoimmune response is not well understood, but it is believed to involve a combination of genetic and environmental factors.

Type 2 Diabetes: Type 2 diabetes is the most prevalent form, accounting for about 90-95% of all diabetes cases. It usually develops in adults over the age of 45 but is increasingly seen in younger populations, including children, adolescents, and young adults. Type 2 diabetes is characterized by insulin resistance, where the body's cells do not respond effectively to insulin. Over time, the pancreas cannot produce enough insulin to maintain normal blood glucose levels. Several factors contribute to the development of type 2 diabetes, including genetic predisposition, obesity, physical inactivity, and unhealthy dietary habits.

1.4 Systemic Complications of Diabetes

Diabetes affects multiple organ systems, leading to a range of complications if not adequately managed. The chronic hyperglycemia associated with diabetes damages blood vessels and nerves, contributing to various microvascular and macrovascular complications.

Microvascular Complications

These include diabetic retinopathy, nephropathy, and neuropathy. Diabetic retinopathy is a leading cause of blindness, resulting from damage to the blood vessels in the retina. Diabetic nephropathy, characterized by kidney damage, can lead to end-stage renal disease. Diabetic neuropathy affects peripheral nerves, causing symptoms such as pain, numbness, and increased risk of foot ulcers and amputations.

Macrovascular Complications

These include cardiovascular diseases such as coronary artery disease, stroke, and peripheral artery disease. Diabetes accelerates atherosclerosis, the buildup of fatty deposits in the arteries, increasing the risk of heart attacks and strokes. Managing blood glucose levels, blood pressure, and lipid profiles is crucial in reducing the risk of these complications.

1.5 Psychological Impact of Diabetes

Living with diabetes imposes a significant psychological burden on individuals, deeply affecting their mental and emotional well-being. The relentless need to monitor blood glucose levels, adhere to strict dietary restrictions, and manage a complex regimen of medications can be overwhelming. This constant vigilance can lead to chronic stress and anxiety, which, over time, may erode an individual's resilience and capacity to cope effectively with the disease. The psychological toll of diabetes is multifaceted, impacting various aspects of a person’s life and well-being. A prominent aspect of the psychological impact is diabetes distress, a condition specifically associated with the emotional burdens of managing diabetes. Diabetes distress is characterized by feelings of frustration, defeat, and burnout due to the continuous demands of diabetes management. Individuals may feel overwhelmed by the need to constantly balance their blood sugar levels, which can fluctuate unpredictably despite their best efforts. This distress is not merely transient but can persist, leading to a sense of helplessness and loss of control over one's health.

Furthermore, the psychological impact of diabetes is not limited to the individual but also affects their relationships with family and friends. Family members may experience stress and anxiety about their loved one's health, and the dynamics within the family can shift as everyone adjusts to the demands of diabetes management. This can create tension and strain in relationships, particularly if family members feel burdened by the responsibility of providing support. Addressing the psychological aspects of diabetes is essential for comprehensive diabetes care. This involves integrating psychological support into diabetes management plans. Mental health professionals, such as psychologists and counselors, play a crucial role in providing therapy and support to help individuals cope with the emotional burdens of diabetes. CBT has been shown to be particularly effective in helping individuals develop healthier thought patterns and coping strategies. Additionally, peer support groups offer a platform for individuals to share their experiences and gain support from others who understand the unique challenges of living with diabetes.

In summary, the psychological impact of diabetes is profound and multifaceted, affecting mental health, emotional well-being, self-image, and social relationships. Recognizing and addressing these psychological aspects is crucial for holistic diabetes care, as mental health

is inextricably linked to physical health. Through comprehensive support that includes psychological care, education, and community resources, individuals with diabetes can achieve better health outcomes and improved quality of life.

1.6 Preventative and Management Strategies

Effective management of diabetes requires a comprehensive approach that integrates lifestyle modifications, medication, and regular monitoring. For Type 1 diabetes, insulin therapy is indispensable. Patients must administer insulin, often through injections or insulin pumps, to regulate their blood glucose levels. CGM systems provide real-time feedback on glucose levels, enabling more precise adjustments to insulin dosage. Carbohydrate counting, which involves calculating the carbohydrate content of meals to match insulin doses, is also crucial for maintaining optimal blood glucose levels. Technological advances, such as closed-loop systems (also known as artificial pancreas systems), have significantly improved the quality of life for individuals with Type 1 diabetes by automating insulin delivery and minimizing blood glucose fluctuations.

In managing Type 2 diabetes, lifestyle interventions are paramount. Weight loss is a primary goal, as even modest reductions in body weight can improve insulin sensitivity and glucose control. Increased physical activity, such as regular aerobic exercise and resistance training, enhances insulin sensitivity and helps lower blood glucose levels. Dietary changes, including a balanced diet rich in whole grains, fruits, vegetables, lean proteins, and healthy fats, are fundamental to managing blood sugar levels. Reducing the intake of refined sugars and processed foods is particularly beneficial.

Pharmacotherapy is often necessary for Type 2 diabetes management. Metformin is commonly the first-line medication due to its effectiveness in lowering blood glucose levels and its favorable safety profile. Other medications, such as sulfonylureas, GLP-1 receptor agonists, and SGLT2 inhibitors, can be added based on individual patient needs. GLP-1 receptor agonists and SGLT2 inhibitors not only help control blood glucose but also offer cardiovascular benefits and promote weight loss, which are advantageous for many patients with Type 2 diabetes.

Regular monitoring is essential for both types of diabetes. Patients should frequently check their blood glucose levels to ensure they remain within target ranges. HbA1c testing, which reflects average blood glucose levels over the past two to three months, is a critical measure for assessing long-term glucose control. Additionally, routine screening for complications, such as retinopathy, nephropathy, and neuropathy, is vital to identify and manage potential issues early.

In conclusion, the management of diabetes, whether Type 1 or Type 2, necessitates a multifaceted approach. By combining lifestyle modifications, appropriate medication, and vigilant monitoring, individuals with diabetes can achieve better health outcomes and improve their quality of life.

1.7 Education and Support

Education and support are essential pillars in the comprehensive management of diabetes. Effective diabetes education programs are designed to empower individuals with the knowledge and skills necessary to manage their condition proactively. These programs encompass a wide range of topics, including understanding the pathophysiology of diabetes, recognizing the importance of blood glucose monitoring, mastering the principles of carbohydrate counting, and learning the correct administration of medications, including insulin. By fostering a deeper understanding of the disease, these educational initiatives enable individuals to make informed decisions about their daily care and long-term health.

Self-management education is particularly crucial, as diabetes requires daily attention and careful planning. Individuals are taught how to monitor their blood glucose levels accurately and interpret the results to make necessary adjustments in their diet, physical activity, and medication. They also learn how to recognize and respond to the signs of hypo- and hyperglycemia, which is vital for preventing acute complications. Furthermore, education on the importance of regular medical check-ups and screenings for diabetes-related complications, such as retinopathy, nephropathy, and neuropathy, helps in the early detection and management of potential issues.

Support groups play a vital role in the emotional and psychological well-being of individuals with diabetes. These groups provide a platform for individuals to share their experiences, challenges, and successes, fostering a sense of community and belonging. The shared experiences within these groups can offer practical advice and strategies for overcoming common obstacles, as well as emotional support that helps individuals feel less isolated in their journey. Support groups can be found in various settings, including community centers, hospitals, and online forums, making them accessible to a wide range of people.

Counselling services are another critical resource, offering professional guidance to help individuals manage the emotional and psychological aspects of living with diabetes. Diabetes can be overwhelming, and the constant vigilance required to manage the disease can lead to significant stress and anxiety. Professional counsellors can help individuals develop coping strategies to manage these emotions, reduce stress, and improve their overall mental health. This support is essential for maintaining a balanced and positive outlook, which is crucial for effective diabetes management.

Engaging family members and caregivers in the education process is also important. They play a crucial role in providing day-to-day support and encouragement. Family members who understand the complexities of diabetes management can better assist with meal planning, remind individuals to monitor their blood glucose levels, and offer emotional support during challenging times. Educating family members helps create a supportive home environment, which can significantly enhance the individual's ability to manage their diabetes effectively.

Furthermore, diabetes education should be an ongoing process. As new research emerges and treatment options evolve, continuous education ensures that individuals stay informed about the latest developments in diabetes care. Healthcare providers should encourage regular participation in refresher courses and updates to help individuals stay current with their knowledge and skills.

In addition to structured education programs, informal education through reliable sources such as reputable websites, books, and online courses can also be beneficial. Patients

should be encouraged to seek information from credible sources to avoid misinformation, which can be detrimental to their health.

1.8 Public Health Initiatives

Public health initiatives are crucial in combating the diabetes epidemic by addressing its root causes and promoting healthier lifestyles within communities. These initiatives encompass a range of programs and policies designed to prevent the onset of type 2 diabetes, detect the disease early, and manage its complications effectively. By targeting the population at large, public health efforts aim to reduce the overall burden of diabetes and improve health outcomes on a broad scale.

One of the primary strategies involves promoting healthy lifestyles through community-based programs that encourage regular physical activity and healthy eating habits. These programs often include public campaigns that highlight the importance of exercise in maintaining a healthy weight and improving insulin sensitivity. Community fitness events, accessible recreational facilities, and partnerships with local gyms can help make physical activity more attractive and attainable for everyone. In tandem, nutrition education initiatives teach individuals about the benefits of a balanced diet rich in fruits, vegetables, whole grains, and lean proteins while reducing the intake of processed foods, sugary drinks, and unhealthy fats. School-based programs can instil these healthy habits from a young age, creating a foundation for lifelong health.

Another critical component of public health initiatives is the promotion of smoking cessation. Smoking is a significant risk factor for diabetes and its complications, and quitting smoking can significantly improve health outcomes for individuals at risk for or living with diabetes. Public health programs often provide resources such as counselling, support groups, and nicotine replacement therapies to assist individuals in quitting smoking. These efforts are frequently supported by public awareness campaigns that emphasize the health risks associated with smoking and the benefits of cessation.

Screening programs for early detection of diabetes and pre-diabetes are essential in identifying individuals at risk before they develop serious complications. By offering

regular screenings in community centers, workplaces, and healthcare facilities, public health initiatives can help catch diabetes in its early stages. Early detection allows for timely interventions, such as lifestyle modifications and medications, which can delay or even prevent the progression to full-blown diabetes. These programs are especially important in high-risk populations, including those with a family history of diabetes, overweight individuals, and certain ethnic groups.

Public health initiatives also involve public awareness campaigns that educate communities about the risk factors and complications of diabetes. These campaigns use various media, including television, radio, social media, and print materials, to reach a wide audience. They provide valuable information on recognizing the symptoms of diabetes, understanding the importance of regular check-ups, and knowing how to manage the disease effectively. By raising awareness, these campaigns motivate individuals to take proactive steps towards prevention and management, such as adopting healthier lifestyles and seeking medical advice when necessary.

CHAPTER 2

LITERATURE SURVEY

Early detection of diabetes is crucial for timely intervention and effective management to prevent adverse health outcomes. This literature review examines the application of machine learning (ML) classification methods in developing models for the early identification of diabetes. Several studies have explored various ML algorithms to predict diabetes risk accurately, demonstrating the potential of these technologies to enhance healthcare outcomes.

2.1 Shafi S, Ansari GA. Early prediction of diabetes disease &classification of algorithms using machine learning approach. In Proceedings of the ICSMDI 2021A function of an Algorithm and Classification ensemble Machine Learning approach is discussed. Depending on diabetes risk factors the J48 decision tree was used to identify hypertension whether the Patient is diabetic or non-diabetic. The research's finding shows that Adboost Machine Learning ensemble technique outclasses bagging and a J48 decision tree in terms of efficiency. It developed a glucose predictive model with the primary goal of predicting whether or not a patient will develop diabetes at a given age.

The conceptual model is widely used in machine learning and employs decision trees to solve problems to implement it. The observed results are accurate because the developed method performs well in detecting diabetic events at a specific age with greater accuracy by using a Prediction Model. For testing and evaluating the dataset for estimation support vector was used. High blood pressure collection of data sourced from UCI database was used to examine disease. Through Program is designed and we were able to obtain the best data. By reducing the time needed to generate a dataset reliability can be extremely high. It tends to be expense for diabetes prediction a hypertension data sampling model with two sub-modules.

In the first partition an Artificial Neural Network is used and in the second partition. Fasting Blood Sugar is used for decision tree and is used to identify an effects of hypertension on

the health of patients. It can also be applied to the algorithm which can classify the risk of diabetic mellitus. The author used four well-known Machine Learning classification methods to achieve the goal DT, ANN and Naïve Bayes.

Methods like classifiers are used to increase the reliability to build models. The Random Forest algorithm yields the best outcomes of all the algorithms and tested according to the results of the research. The Machine learning approach is critical for predicting a number of medical databases, including hypertension data. The goal of this research is to develop a structure in a medicine device that provides machine learning to predict hypertension using significant features that are mainly associated to the illness. Although data is one of the most essential issues of classification process, machine learning approaches are entirely reliant on it. When data is obtained from various sources in a raw format, there is a risk of several variations, which the model might not be able to manage. As a result, pre-processing is proposed to reduce all variations and create a pure set of data.

Early phase detection in diabetes is a serious research problem and a potential challenge for researchers in the field of human healthcare development. Diabetes is an actual harmful disease and early detection is always a challenge. This research employs machine learning classification algorithms to create a model is capable of overcoming all problems and is useful in the early prediction of diabetes disease. In this research structured efforts are made to design a Framework that can predict diabetes. The 3 Machine Learning approach classification algorithms are analyzed, tested on different measures as part of this research.

The PIDD is used in the experiments. Using the NB classification algorithm, the experimental results show that a built method is adequate with an accuracy of 74.28 percent. The developed framework, along with the Machine Learning different classifiers used could be used to identify or diagnose other diseases in the future. For diabetes research, the study could be extended and improved as well as several other Machine Learning approaches and also authors have planned to do classification for other algorithms with missing data.

2.2 Nai-Arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci.(2015): This paper applied a use of algorithms to classify the risk of diabetes mellitus. Four well known classification models that are Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes were first examined. Then, Bagging and Boosting techniques were investigated for improving the robustness of such models. Additionally, Random Forest was not ignored to evaluate in the study. Findings suggest that the best performance of disease risk classification is Random Forest algorithm. Therefore, its model was used to create a web application for predicting a class of the diabetes risk.

In this work, we proposed a web application by using a use of disease classifiers and a real data set. The data used in this creation are general information of 30,122 people who were collected from 26 Primary Care Units in Sawanpracharak Regional Hospital during 2012 – 2013. Before creating the web application, thirteen classification models were evaluated for seeking a predicting model. These models consist Decision Tree, Neural Network, Logistic Regression, Naïve Bayes and Random Forest algorithms including combination of Bagging and Boosting techniques except Random Forest algorithm. To investigate the robustness of each model, accuracy and ROC Curve were calculated and compared with others.

The results reveal that Random Forest was ranked first in both accuracy and ROC Curve. This might be because of variable selection. In the process of Random Forest, data were not only chosen randomly but also input variables were random selected by considering important variables. Hence, this causes accuracy values increase. Therefore this algorithm was selected to model the diabetes risk prediction and used for creating the application

2.3 Saravananathan K, Velmurugan T. Analyzing diabetic data using classification algorithms in data mining. Indian J SciTechnol. (2016): In this research work, the frequently used classification techniques J48, CART, SVMs, and kNN are analyzed, on the medical dataset to find the optimal solution for Diabetes. The performance indicators accuracy, specificity, sensitivity, precision, error rate are calculated for the given dataset. Accusation beside with a proper data pre-processing technique can get better the accuracy

of the classifier. The function of data normalization had noticeable impact on categorization performance and considerably enhanced the performance of J48. The performance of kNN algorithm has minimum accuracy. Based on the parameters taken for analysis, the performances of the four algorithms are analyzed.

The results show that the performance of J48 technique is significantly superior to the other three techniques for the classification of diabetes data. To improve the overall accuracy, it is necessary to use more data set with large number of attributes and use the best feature selection method in future.The impact of categorization is very important in authentic earth applications in all fields. To categorize the rudiments allowing to the applications of the elements during the predefined set of modules are used by classification methods. Very popular classification algorithms J48, SVM, CART and kNN for diabetic data are used for this research work.

2.4 S. V. K. R. Rajeswari and P. Vijayakumar, "Prediction of diabetes mellitus using machine learning algorithm," Annals of the Romanian Society for Cell Biology, (2021): From the study, it can be concluded that the LGBM algorithm provides the highest accuracy when compared with RF and GB classifiers. Therefore, the LGBM algorithm is well suited for the Pima dataset and the DMS dataset used in the study. The LGBM algorithm differs from RF and GB in the following ways: The parameters used in LGBM are different from those in GB and RF. The parameter tuning varies with each algorithm, and the model is built based on the classifier used.

Therefore, in this paper, a predictive model is built using LGBM algorithm, and the accuracy is obtained. The diabetes mellitus disease prediction can further be improved by enhancing the dataset using other advanced methodologies like transformer based learning. The attributes used can also be employed in different combinations for identification. The classifiers used can be fine-tuned more to predict the disease with higher accuracy, and the probability of occurrence of the disease can be calculated. This will further improve the accuracy percentage and deliver a more profound model to predict diabetes mellitus disease among affected people.

2.5 Sadhu A, Jadli A. Early-stage diabetes risk prediction: A comparative analysis of classification algorithms IARJSET (2021): In this paper we experimented with a diabetes dataset with different classification algorithms. Seven classification algorithms have been implemented on the validation sat of the used dataset. The results drawn from training several machine learning models clearly indicate that Random Forest Classifier proved to be the best model among the models used in the paper for the concerned dataset with an accuracy score of 98.0778%, ROC score of 0.9979 and F-score of0.9790. Top three classifiers for the dataset are Random Forest classifier, Multi-layer perceptron and Support Vector Machine. Although, rest of the algorithm showed an accuracy of more than 90% and a F-score and ROC value of more than 0.9, the random forest classifier stands out with the maximum score in all the three evaluation metrices.

Hence, as per the results obtained we can firmly believe that Random Forest Classifier is one of the most effective algorithms against binary-based classification datasets. For Multi-layer Perceptron to work with highest accuracy it needs to be fed more training data points. This is one of the most valid reason for its underperforming in the concerned dataset. In future more data must be collected from across the world for a more precise and accurate classification of the disease. Future study will concentrate on finding more factors that have the potential to cause diabetes and to include those potential factors in the dataset for a better classification. This can help in the enhancement and automation of diagnosis of the disease. Future studies on the disease and application of various data mining and ML algorithm can help in better early prediction of diabetes.

2.6 Classification of Diabetes Patients Using Kernel-Based Support Vector Machines, authored by G.A.Pethunachiyar and presented at the 2020 (ICCCI): This research paper shows that lot of results on diabetes all over the world. Diabetic data set from CPCSSN database taken for analysis. The researcher used the bagging ensemble techniques using J48 for classifying different age groups patients with diabetes mellitus. Regression based data mining technique used in for predictive analytic in diabetic treatment. Here, Oracle Data Miner tool was used for prediction analysis. The study was

conducted in 2012 for predicting diabetes by using the common risk factors. For the performance analysis, different classification techniques such as decision tree, Neural Networks and logistic regression were considered. The logistic model outperformed the other two in accuracy rate.

Common attributes taken for study are family history, characteristics and lifestyle risk. In, Investigation performed among participants who registered in an adapted Diabetes Prevention Program (DPP) for weigh reduction. The findings in the above support the participants for dietary monitoring by themselves and motivate the participants to increase the levels in exercise. The investigation also performed to analyze the treatment for hypertension based on regression technique.

Study conducted in shows the importance of machine learning methods in medical field and it gradually increasing over the years. The survey proves that the large amount of data was generated from the wide research carried out in all aspects of diabetes and machine learning algorithms had an important role in most of the applications. It also proves that Support vector machines (SVM) is the mostly used algorithm. Even though, DM can be categorized into many types, the major types are type 1 diabetes and type 2 diabetes. Type-2 diabetes is mainly caused by the resistance of insulin in our body and it is the general type in ninety percent of the diabetes patients. Ten percentages of the diabetic patients are affected by Type-1 diabetes. From the study, it is observed that Diagnosis of diabetes patients depends on blood glucose levels. Support Vector Machine is the widely used techniques in prediction of diabetes patients. It evaluates the performance based on accuracy level for various kernel functions of SVM. A result shows that SVM with Linear Kernel outperforms the two kernel functions. The SVM with Linear Kernel function produced 100%, SVM with Radial Kernel Produced 99% and SVM with Polynomial kernel produced 90% for the chosen data set.

2.7 L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. Garcia-Garcia, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," Diabetology & Metabolic Syndrome, vol. 13, no. 1, p. 148, 2021:

Diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. In particular, type 2 diabetes is associated with insulin resistance (insulin action defect), i.e., where cells respond poorly to insulin, affecting their glucose intake [2]. The diagnostic criteria established by the American Diabetes Association are: (1) a level of glycated hemoglobin (HbA1c) greater or equal to 6.5%; (2) basal fasting blood glucose level greater than 126 mg/dL, and; (3) blood glucose level greater or equal to 200 mg/dL 2 h after an oral glucose tolerance test with 75 g of glucose.

Diabetes mellitus is a global public health issue. In 2019, the International Diabetes Federation estimated the number of people living with diabetes worldwide at 463 million and the expected growth at 51% by the year 2045. The review finds that the structure of the dataset is relevant to the accuracy of the models, regardless of the selected features that are heterogeneous between studies. Concerning the models, the optimal performance is for tree-type models. However, even though they have the best accuracy, they require complementary techniques to balance data and reduce dimensionality by selecting the optimal features. Therefore, K Nearest Neighbor, and Support vector machines are frequently preferred for prediction. On the other hand, Deep Neural Networks have the advantage of dealing well with big data. However, they must be applied to datasets with more than 70,000 observations. At least three metrics and the AUC (ROC) should be reported in the results to allow estimation of the others to reduce heterogeneity in the performance comparison. Therefore, the areas of opportunity are listed below.

CHAPTER 3

DESIGNING AND METHODOLOGY

Detecting diabetes early is essential for swiftly intervening and managing the condition to prevent unforeseen outcomes. This study explores the use of machine learning (ML) classification methods to develop models for early identification of diabetes development. The following sections review various research efforts that have utilized ML algorithms for diabetes prediction.

3.1 Description of the Dataset

The dataset used in many studies for diabetes prediction is the "Pima Indian Diabetes Dataset," which is acquired from the UCI Machine Learning Repository.

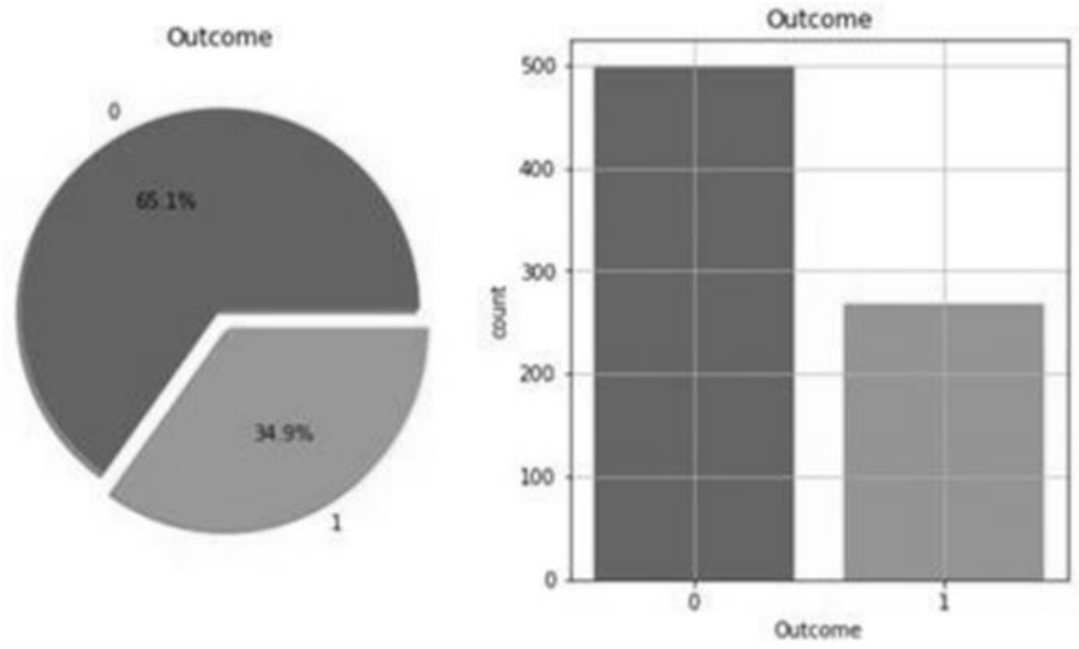


Fig. 3.1 Percentage of people having diabetes in the Pima Indian dataset

This dataset includes several attributes related to diabetes, as shown in Table 1:

S.No.	Attributes
I	Pregnancy
II	Glucose
III	Blood Pressure
IV	Skin Thickness
V	Insulin
VI	BMI(Body Mass Index)
VII	Diabetes Pedigree Function
VIII	Age

Table 1

3.2 Data Preprocessing

In healthcare datasets, the class variable indicates diabetic outcomes (0 for negative, 1 for positive). Data preprocessing is vital to address missing values and impurities, ensuring the accuracy and effectiveness of machine learning techniques. This process enhances data quality, leading to successful predictions. By preprocessing the data, researchers optimize the dataset for machine learning analysis, thereby improving the accuracy and reliability of predictive models for diabetic outcomes.

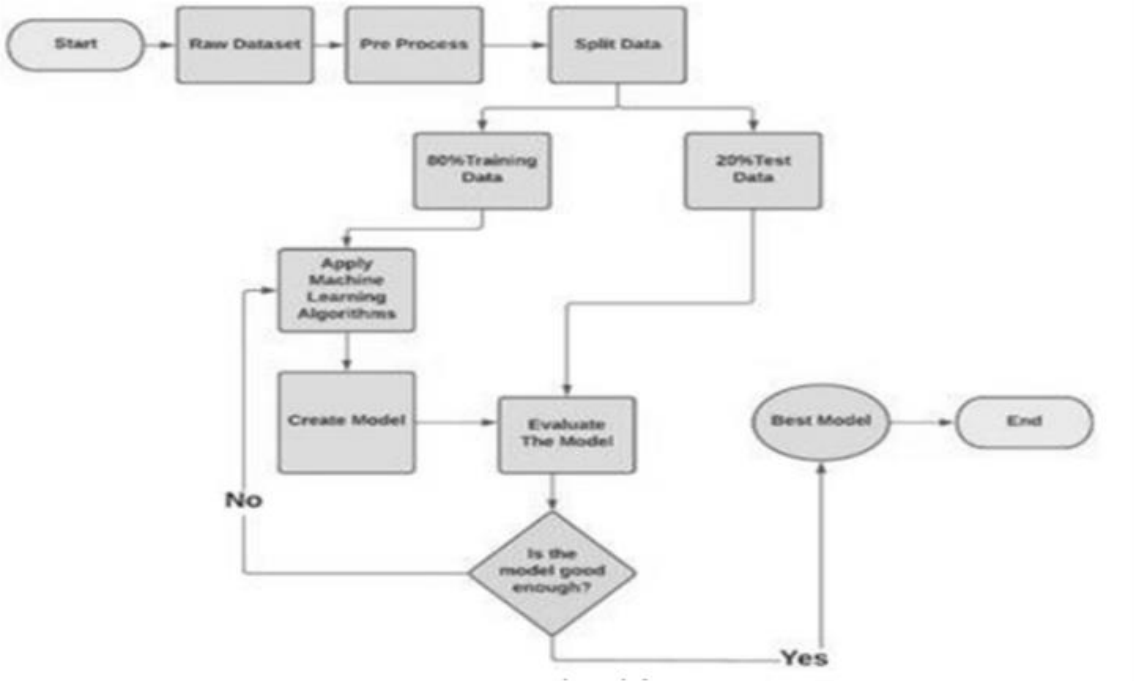


Fig. 3.2 Proposed Model

3.3 Data Selection and Preparation

The study begins with data selection from the UCI repository, addressing missing values, inconsistencies, and erroneous information. Subsequently, the data is prepared by splitting it into training and testing datasets, typically with a 70% and 30% allocation, respectively. This ensures that the models developed are both trained effectively and tested on unseen data to validate their performance.

3.4 Machine Learning: The techniques of Machine learning such as Naïve Bayes, SVM, and KNN are then employed for prediction, constituting a crucial stage in achieving research objectives. This methodical approach ensures data integrity and enables accurate predictions through diverse machine learning techniques:-

3.4.1 Support Vector Machine- In machine learning, support vector machines (SVMs, also support vector networks) are supervised max-margin models with associated

learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues. SVMs are one of the most studied models, being based on statistical learning frameworks of VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974).

The popularity of SVMs is likely due to their amenability to theoretical analysis, their flexibility in being applied to a wide variety of tasks, including structured prediction problems. It is not clear that SVMs have better predictive performance than other linear models, such as logistic regression and linear regression.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

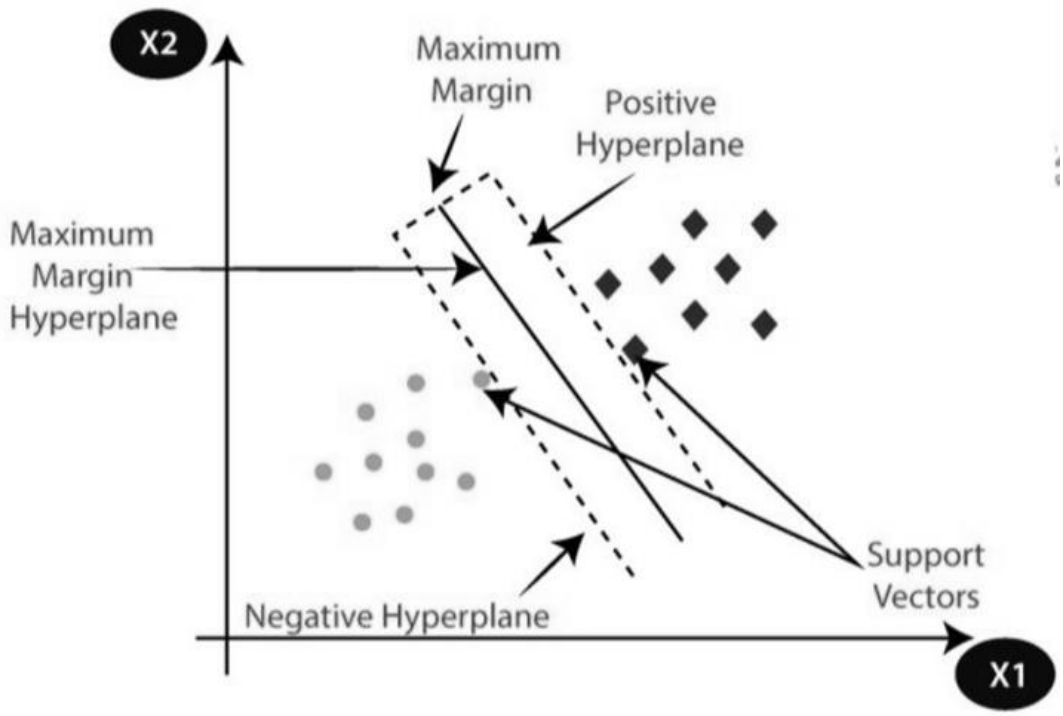


Fig. 3.4.1 Support Vector Machine

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

SVM is a "Supervised Machine Learning" algorithm, commonly abbreviated as SVM. [8] The most used categorization method is SVM. Two classes are divided by a hyperplane made by SVM. In high-dimensional space, it can produce a single hyperplane or a set of hyperplanes. Regression and classification are further uses for this hyperplane. In addition to classifying entities that lack data support, SVM distinguishes between instances within particular classes. The closest training point for each class is reached through the use of a hyperplane for separation.

Algorithm-

• Identify the hyperplane that optimally separates the classes.

• Calculate the margin, which the distance measured between each data point and the hyperplanes.

• A low distance between classes increases the likelihood of misclassification, and vice versa.

• Opt for the class with the highest margin, where the margin is computed as the sum of distances to positive and negative points.

3.4.2 K-Nearest Neighbor- K-NN algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution. It can also handle both numerical and categorical data, making it a flexible choice for various types of datasets in classification and regression tasks. It is a non-parametric method that makes predictions based on the similarity of data points in a given dataset. K-NN is less sensitive to outliers compared to other algorithms.

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data.

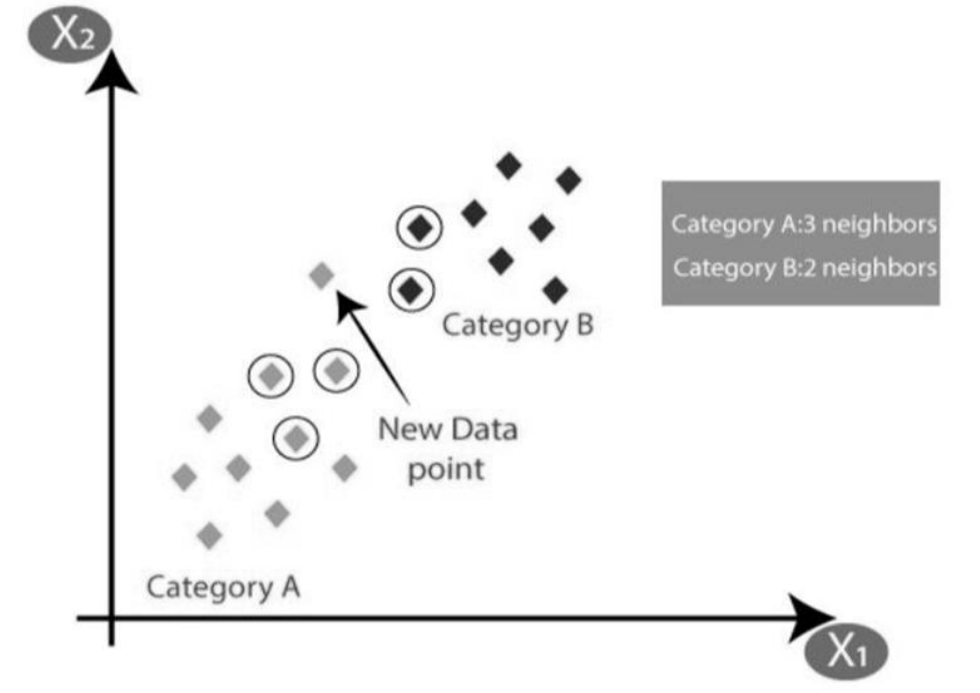


Fig. 3.4.2 K-Nearest Neighbor

'KNN' is another type of "Supervised Machine Learning" algorithm, which assists in addressing both regression as well as classification tasks. KNN presumes that similar objects are situated nearby each other. Data points which are alike are located adjacent to themselves. KNN aids in categorizing novel work in light of similarity metrics.

All of the data is documented by the "KNN Algorithm", which then categorizes them on how alike they are accordingly. Uses a tree structure to compute distance between points. This algorithm identifies the closest neighbors of a new data point in the training data set for generating a prediction for it. The value of K, which stands for "number of nearby neighbors," is always positive.

The neighbor's value is picked from a set of classes. Euclidean distance is the primary measure used to define "closeness". The "Euclidean Distance" between two points 'P' and 'Q' i.e. P (p1, p2, …. pn) and Q (q1, q2,..qn) is determined by this subsequent formula:

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

• Utilize a sample dataset, such as "Pima dataset", consisting of rows and columns.

• Prepare a test dataset containing attributes and rows.

• Calculate the Euclidean distance using the appropriate formula.

$$EuclideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

• Determine a "random value K", representing number of closest neighbors.

• Utilize the "minimum distance" and "Euclidean distance" to determine nth column for each.

Obtain corresponding output values for the determined columns. If values are identical, then the patient has diabetes, otherwise not.

3.4.3 Naive Bayes Classifier- Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. To start with, let us consider a dataset. One of the most simple and effective classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities.

In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers.[4] Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification Naïve Bayes algorithm is used for

classification problems. It is highly used in text classification. In text classification tasks, data contains high dimension (as each word represent one feature in the data). It is used in spam filtering, sentiment detection, rating classification etc. The advantage of using naïve Bayes is its speed. It is fast and making prediction is easy with high dimension of data.The likelihood that an event will occur depends on past knowledge of potential event related circumstances, as determined by Naive Bayes.[10] The most straightforward and quick classification algorithm, Naive Bayes, works well with large data blocks. The NB classifier is used in many different applications, including recommender systems, text categorization, sentiment analysis, and spam filtering. The probability of the unknown classes is predicted using the Bayes theorem.

The Naive Bayes algorithm is simple to understand and apply. This is why sparse data sets have the potential to outperform more complex models.

P(h|e) = (P(h|e) * P(h)) / P(e) where,

• "(P(h|e))" signifies 'posterior probability', representing the probability of 'h'(hypothesis) given 'e'11(event).

• "(P(e|h))" signifies likelihood, indicating the probability 'e'(event) given that 'h'(hypothesis) is true. • "(P(h))" represents 'prior probability', denoting the probability of 'h'(hypothesis) being true.

• "(P(e))" signifies probability of 'e'(event) occurring.

3.4.4 Gradient Boosting- Gradient boosting is a machine learning technique based on boosting in a functional space, where the target is *pseudo-residuals* rather than the typical residuals used in traditional boosting. It gives a prediction model in the form of an ensemble of weak prediction models, i.e., models that make very few assumptions about the data, which are typically simple decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion

as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

Gradient Boosting is mainly of two types depending on the target columns:
Gradient Boosting Regressor: It is used when the columns are continuous.
Gradient Boosting Classifier: It is used when the target columns are classification problems.

Gradient Boosting is a classification technique and the most potent ensemble method for prediction. To create powerful learner models for prediction, it combines week learners collectively. The Decision Tree model is employed. It is a very popular and effective method for classifying complex data sets. The performance of the gradient boosting model gets better with each iteration.
Gradient boosting can be used in the field of learning to rank. The commercial web search engines Yahoo and Yandex use variants of gradient boosting in their machine-learned ranking engines. Gradient boosting is also utilized in High Energy Physics in data analysis. At the Large Hadron Collider (LHC), variants of gradient boosting Deep Neural Networks (DNN) were successful in reproducing the results of non-machine learning methods of analysis on datasets used to discover the Higgs boson.[16] Gradient boosting decision tree was also applied in earth and geological studies – for example quality evaluation of sandstone reservoir.

Algorithm-

• Begin with sample desired values, denoted as ‘P’.

• Calculate the error present in the desired values.

• Adjust and update the weights to minimize the error, denoted as ‘M’.

• Update the target values using the formula P[x] = p[x] + alpha * M[x].

• Analyze and compute the performance of model learners using a loss function F.

• Perform the aforementioned measures until the P(desired result) is achieved.

3.4.5 Logistic Regression- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd *et al.* using logistic regression.

Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).

Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or Any Other Party, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc.

In economics, it can be used to predict the likelihood of a person ending up in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks, such as identifying whether an email is spam or not and diagnosing diseases by assessing the presence or absence of specific conditions based on patient test results. This approach utilizes the logistic (or sigmoid) function to transform a linear combination of input features into a probability value ranging between 0 and 1. This probability indicates the likelihood that a given input corresponds to one of two predefined categories. The essential mechanism of logistic regression is grounded in the logistic function's ability to model the probability of binary outcomes accurately. With its distinctive S-shaped curve, the logistic function effectively maps any real-valued number to a value within the 0 to 1 interval. This feature renders it particularly suitable for binary classification tasks, such as sorting emails into "spam" or "not spam". By calculating the probability that the dependent variable will be categorized into a specific group, logistic regression provides a probabilistic framework that supports informed decision-making.

Disaster planners and engineers rely on these models to predict decision take by householders or building occupants in small-scale and large-scales evacuations ,such as building fires, wildfires, hurricanes among others. These models help in the development of reliable disaster managing plans and safer design for the built environment.

Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

It is much similar to the Linear Regression except that how they are used. It is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

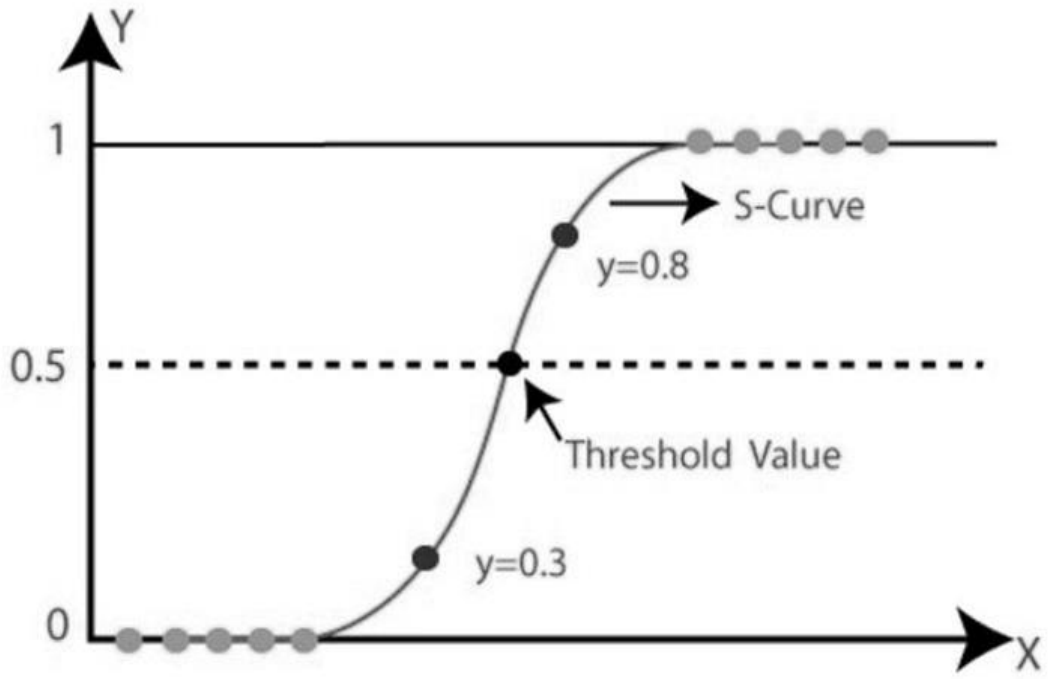


Fig. 3.4.5 Logistic Regression

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

The probability in logistic regression establishes if a particular data entry belongs to the class indicated by the number [4] (‘Brownlee’, 2016c). The data is modelled using ‘sigmoid function’ in logistic regression in the following ways:

$$P(X) = \frac{1}{1+e^{-y}}$$

In this case, ‘y’ represents the real numerical value, ‘e’ is base of the natural logarithms, and ‘P(X)’ is probability that X lies between 0 and 1.

3.4.6 Random Forest– Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition.

This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks).

This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.

For problems requiring regression and classification, this kind of collective learning approach is used. On comparing with other models, the accuracy it offers is higher. Huge datasets can be managed with ease through this method. “Leo Bremen” is the one who created Random Forest. It seems to be highly liked method for group learning.

On decreasing variance, Random Forest Enhances Decision Tree Performance. To operate, this algorithm constructs numerous decision trees in the training phase. It subsequently produces a classification based on either the average prediction (for regression) from the individual trees or the consensus classification of all the trees combined.

Algorithm-

• Begin by selecting "K" features from the total "M" features, where K is significantly smaller than M. • Identify the best split point within the selected "R" features for each node.

• Based on the optimal split, break the node into subnodes.

• Repeat steps 1 to 3 until the desired number of nodes, denoted as "l," has been achieved.

• Construct a forest by iteratively repeating steps 1 to 4 for a certain amount of times, creating "z" total trees.

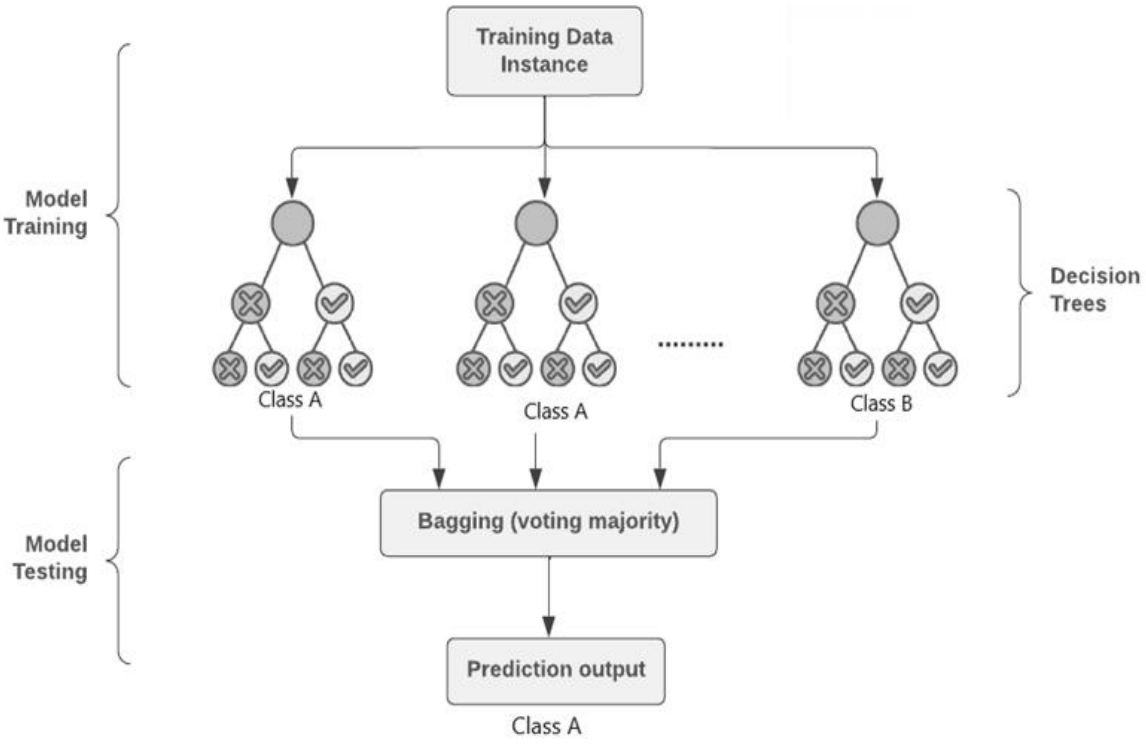


Fig. 3.4.6 Random Forest

CHAPTER 4

IMPLEMENTATION

Introduction

The process of building a machine learning model for diabetes prediction encompasses several critical stages:

- Data Preprocessing: Cleaning and preparing the data for analysis.
- Exploratory Data Analysis (EDA): Understanding the dataset's characteristics and relationships through statistical summaries and visualizations.
- Model Selection: Choosing the appropriate machine learning algorithms to use.
- Training: Feeding the data to the chosen models to learn patterns.
- Evaluation: Assessing the model's performance using various metrics.
- Visualization of Results: Presenting the findings in an interpretable manner.

This guide provides an in-depth look at each step, highlighting the techniques and tools essential for building an effective diabetes prediction model.

4.1 Tech Stack

To implement this model, we will use the following technologies:

1. Programming Language: Python: A versatile language popular for data science and machine learning due to its readability and extensive libraries.

2. Libraries:

a. Data manipulation and analysis:

- **pandas:** Provides data structures and functions needed to manipulate structured data seamlessly.
- **numpy:** Offers support for large multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

b. Data visualization:

- **seaborn:** Based on matplotlib, it provides a high-level interface for drawing attractive and informative statistical graphics.
- **matplotlib:** A comprehensive library for creating static, animated, and interactive visualizations in Python.

c. Machine learning: scikit-learn: A robust library offering simple and efficient tools for data mining and data analysis, built on numpy, scipy, and matplotlib.

3. IDE/Environment: Jupyter Notebook: An open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.

4.2 Planning and Implementation

1. Data Loading and Inspection

a. Loading the Dataset: The first step involves loading the dataset into the environment. This is typically done using pandas, which can read data from various formats (CSV, Excel, SQL, etc.).

b. Inspecting the Structure: After loading the data, it's crucial to inspect its structure. This includes:

- Viewing the first few rows to get a sense of the data.
- Checking the data types to ensure they are appropriate for the analysis.
- Summarizing the data to identify any missing values, outliers, or anomalies

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
df=pd.read_csv('diabetes.csv')
df.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

The head() method gives a preview of the dataset, showing the first five rows. This is followed by descriptive statistics and dataset information:

```
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	3.703500	121.182500	69.145500	20.935000	80.254000	32.193000	0.470930	33.090500	0.342000
std	3.306063	32.068636	19.188315	16.103243	111.180534	8.149901	0.323553	11.786423	0.474498
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	63.500000	0.000000	0.000000	27.375000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	40.000000	32.300000	0.376000	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	130.000000	36.800000	0.624000	40.000000	1.000000
max	17.000000	199.000000	122.000000	110.000000	744.000000	80.600000	2.420000	81.000000	1.000000

```
df.info()
```

```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
```

Checking for any missing values:

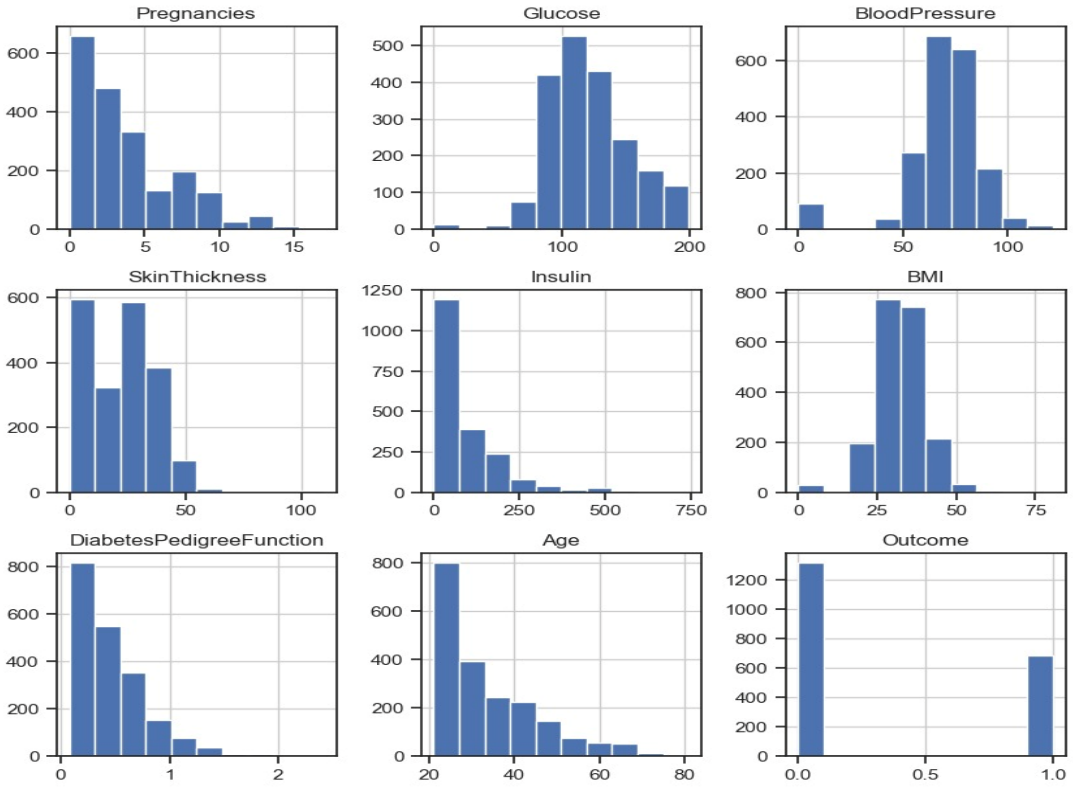
```
df.isnull().values.any()
```

```
False
```

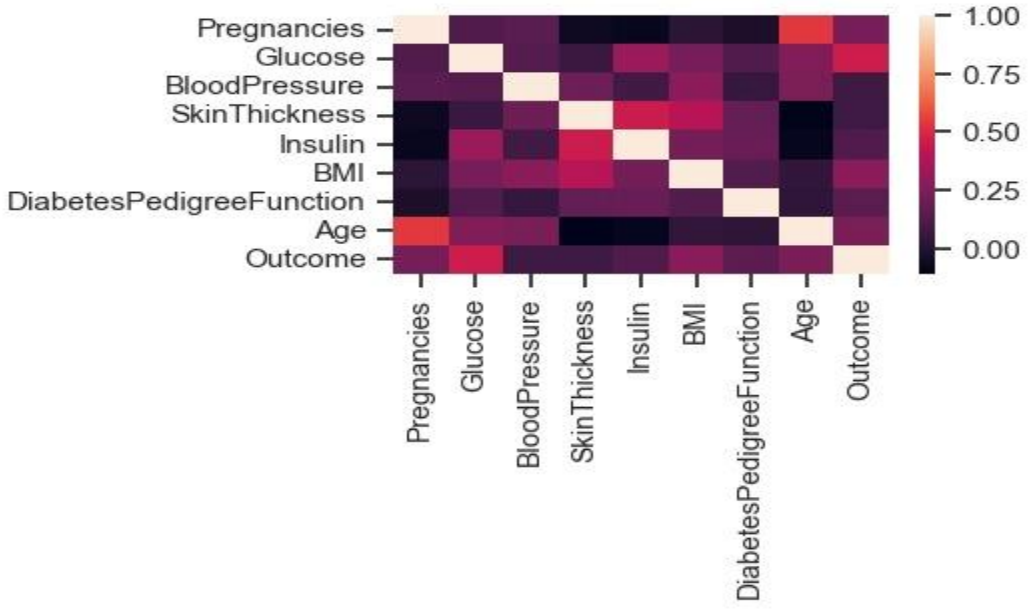
2. Exploratory Data Analysis (EDA)

Visualizing the distribution of data helps in understanding the underlying patterns:

```
df.hist(bins=10,figsize=(10,10))
plt.show()
```

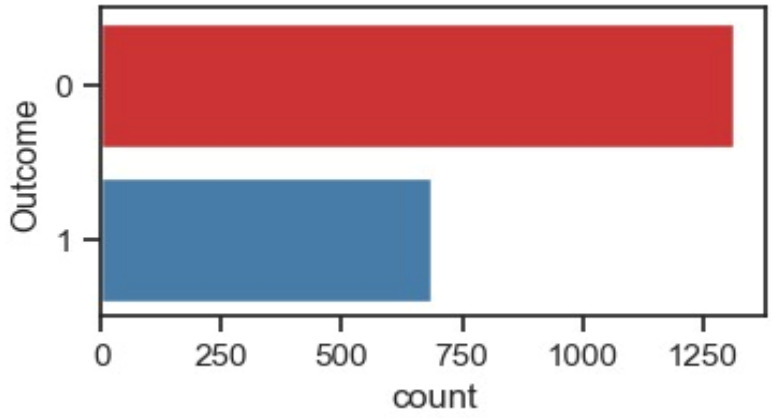



```
sns.heatmap(df.corr())
```

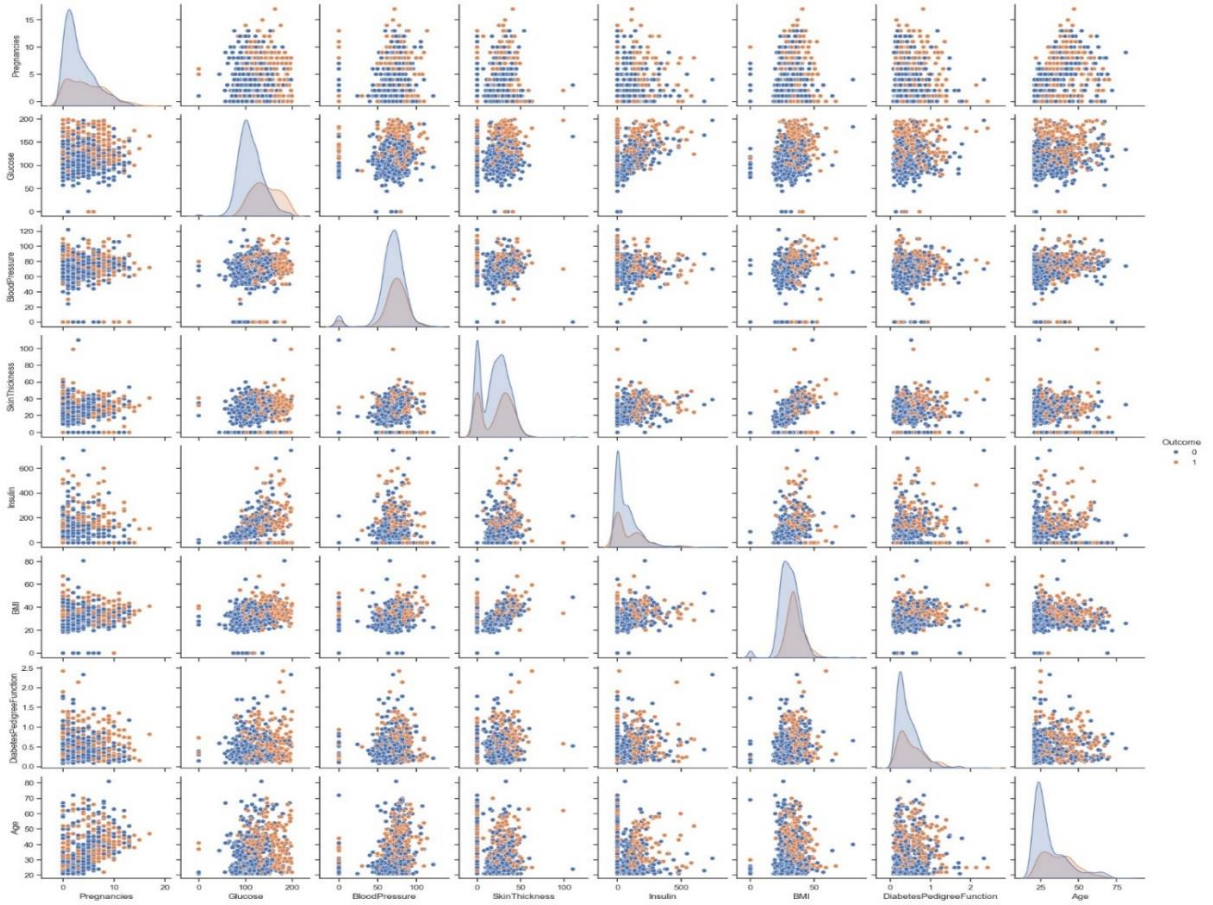


The heatmap shows the correlation between different features. We also visualize the count of diabetic vs non-diabetic cases:

```
sns.countplot(y=df['Outcome'],palette='Set1')
```

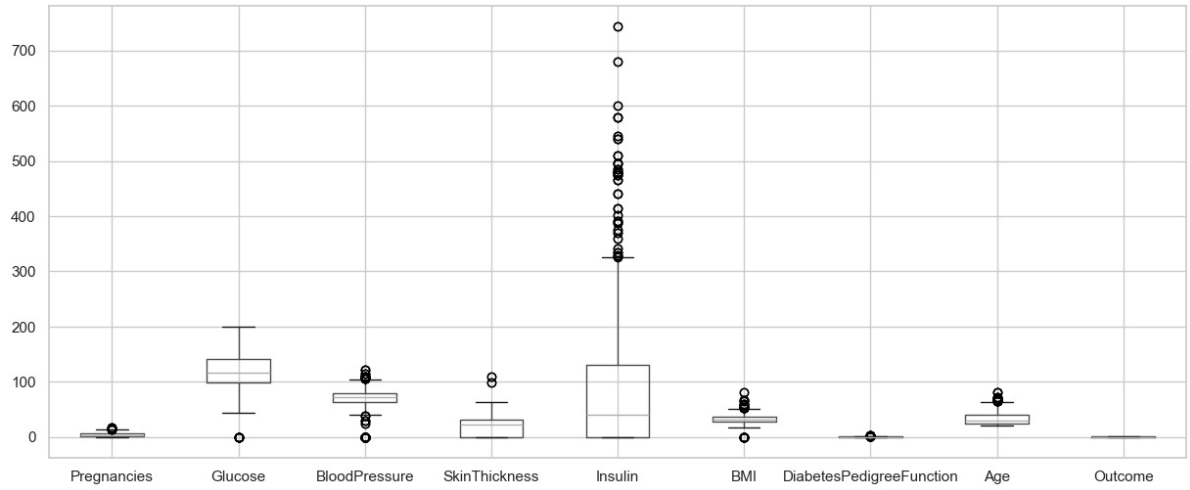



```
sns.set(style="ticks")
sns.pairplot(df, hue="Outcome")
```

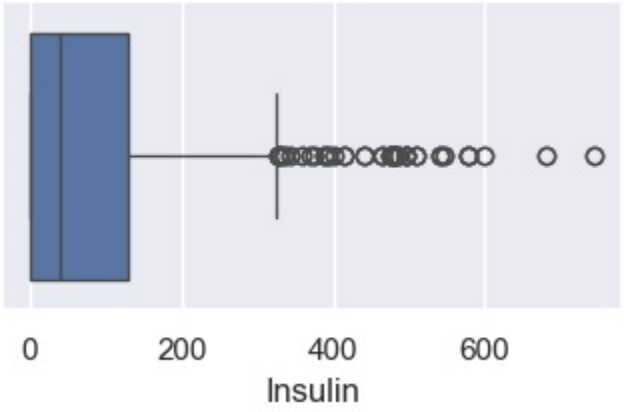


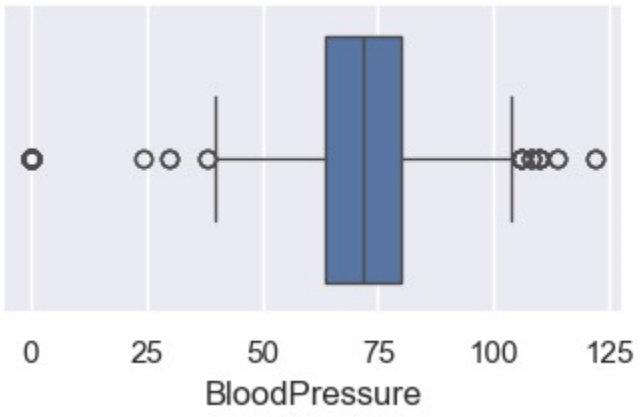
Box plots are useful for identifying outliers:

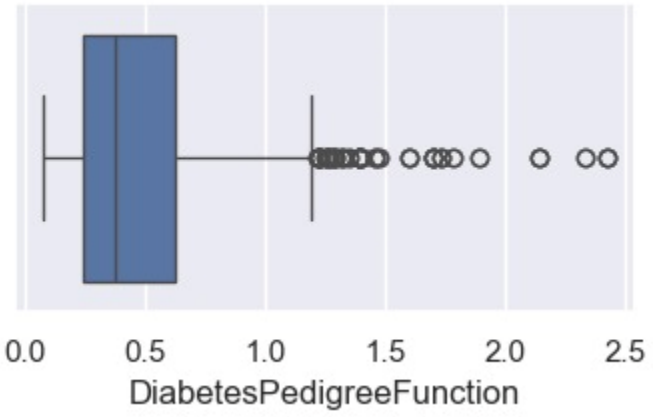
```
sns.set(style="whitegrid")
df.boxplot(figsize=(15,6))
```




```
sns.set(style="whitegrid")

sns.set(rc={'figure.figsize':(4,2)})
sns.boxplot(x=df['Insulin'])
plt.show()
sns.boxplot(x=df['BloodPressure'])
plt.show()
sns.boxplot(x=df['DiabetesPedigreeFunction'])
plt.show()
```







3. Data Preprocessing

To improve model performance, we handle outliers by using the Interquartile Range (IQR) method:

```
Q1=df.quantile(0.25)
Q3=df.quantile(0.75)
IQR=Q3-Q1

print("---Q1--- \n",Q1)
print("\n---Q3--- \n",Q3)
print("\n---IQR---\n",IQR)

#print((df < (Q1 - 1.5 * IQR))|(df > (Q3 + 1.5 * IQR)))
```

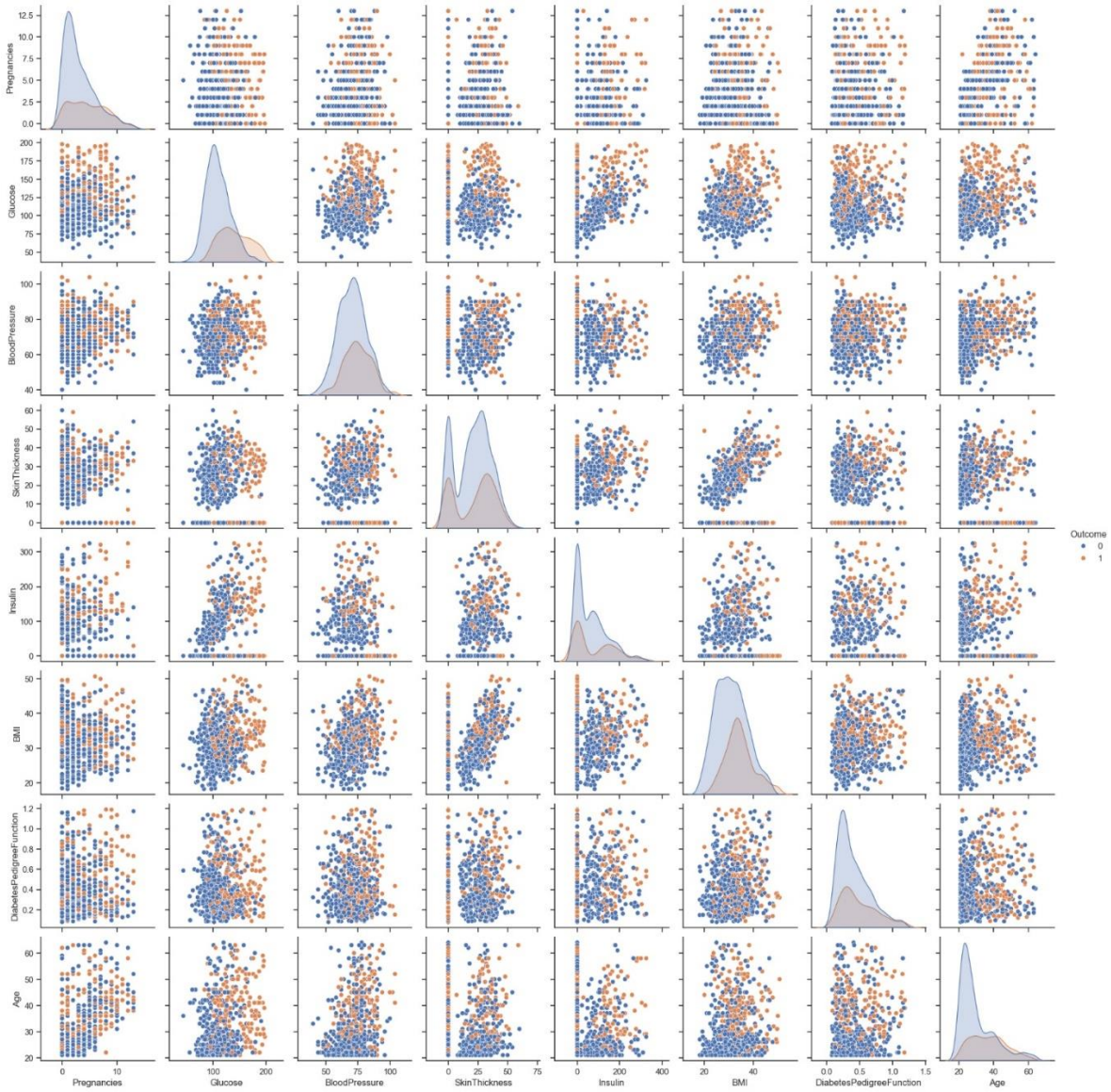
```
---Q1---
 Pregnancies                 1.000
Glucose                     99.000
BloodPressure               63.500
SkinThickness                0.000
Insulin                      0.000
BMI                         27.375
DiabetesPedigreeFunction     0.244
Age                         24.000
Outcome                      0.000
Name: 0.25, dtype: float64


---Q3---
 Pregnancies                 6.000
Glucose                    141.000
BloodPressure               80.000
SkinThickness               32.000
Insulin                    130.000
BMI                         36.800
DiabetesPedigreeFunction     0.624
Age                         40.000
Outcome                      1.000
Name: 0.75, dtype: float64

---IQR---
...
DiabetesPedigreeFunction     0.380
Age                         16.000
Outcome                      1.000
dtype: float64
```

```
df_out = df[~((df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR))).any(axis=1)]
df.shape,df_out.shape
```

```
((2000, 9), (1652, 9))
```

After removing outliers, we re-visualize the data to ensure cleanliness:

```
sns.set(style="ticks")
sns.pairplot(df_out, hue="Outcome")
plt.show()
```



4. Feature and Target Extraction

We separate features (X) and target (y):

```
X=df_out.drop(columns=['Outcome'])
y=df_out['Outcome']
```

Splitting the data into training and testing sets:

```
from sklearn.model_selection import train_test_split
train_X,test_X,train_y,test_y=train_test_split(X,y,test_size=0.2)
```

```
train_X.shape,test_X.shape,train_y.shape,test_y.shape
```

```
((1321, 8), (331, 8), (1321,), (331,))
```

5. Model Building and Evaluation

We employ several machine learning models to find the best-performing one. We use accuracy and ROC AUC as evaluation metrics. Functions to calculate confusion matrix components and display results are defined:

```
from sklearn.metrics import confusion_matrix,accuracy_score,make_scorer
from sklearn.model_selection import cross_validate

def tn(y_true, y_pred): return confusion_matrix(y_true, y_pred)[0, 0]
def fp(y_true, y_pred): return confusion_matrix(y_true, y_pred)[0, 1]
def fn(y_true, y_pred): return confusion_matrix(y_true, y_pred)[1, 0]
def tp(y_true, y_pred): return confusion_matrix(y_true, y_pred)[1, 1]

scoring = {'accuracy': make_scorer(accuracy_score),'prec': 'precision'}
scoring = {'tp': make_scorer(tp), 'tn': make_scorer(tn),
           'fp': make_scorer(fp), 'fn': make_scorer(fn)}

def display_result(result):
    print("TP: ",result['test_tp'])
    print("TN: ",result['test_tn'])
    print("FN: ",result['test_fn'])
    print("FP: ",result['test_fp'])
```

Logistic Regression:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

acc=[]
roc=[]

clf=LogisticRegression()
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.7583081570996979 ROC 0.7084831056793672
```

```
TP:  [17 22 21 21 23 20 18 24 23 22]
TN:  [87 85 81 86 84 89 80 75 82 89]
FN:  [23 17 19 19 17 20 22 16 17 18]
FP:  [ 6  8 11  6  8  3 12 17 10  3]
```

Support Vector Machine:

```
from sklearn.svm import SVC

clf=SVC(kernel='linear')
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.7673716012084593 ROC 0.7154924514737598
TP:  [17 22 21 23 21 20 16 23 23 22]
TN:  [87 84 83 87 84 87 82 79 83 89]
FN:  [23 17 19 17 19 20 24 17 17 18]
FP:  [ 6  9  9  5  8  5 10 13  9  3]
```

K-Nearest Neighbors (KNN):

```
from sklearn.neighbors import KNeighborsClassifier

clf=KNeighborsClassifier(n_neighbors=3)
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.8429003021148036 ROC 0.81845594696062
TP:  [27 32 26 31 32 28 31 27 33 31]
TN:  [86 85 84 90 86 80 78 84 84 85]
FN:  [13  7 14  9  8 12  9 13  7  9]
FP:  [ 7  8  8  2  6 12 14  8  8  7]
```

Random forest:

```
from sklearn.ensemble
import RandomForestClassifier

clf=RandomForestClassifier()
clf.fit(train_X,train_y)

y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.9788519637462235 ROC 0.9739595814362171
TP:  [37 38 38 39 39 38 36 36 34 37]
TN:  [88 92 90 91 90 89 87 91 89 90]
FN:  [3 1 2 1 1 2 4 4 6 3]
FP:  [5 1 2 1 2 3 5 1 3 2]
```

Naive Bayes Theorem:

```
from sklearn.naive_bayes import GaussianNB

clf=GaussianNB()
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.770392749244713 ROC 0.7410735681763718
TP:  [18 27 23 25 26 24 24 24 24 25]
TN:  [81 82 78 82 75 82 71 74 78 78]
FN:  [22 12 17 15 14 16 16 16 16 15]
FP:  [12 11 14 10 17 10 21 18 14 14]
```

Gradient Boosting Classifier:

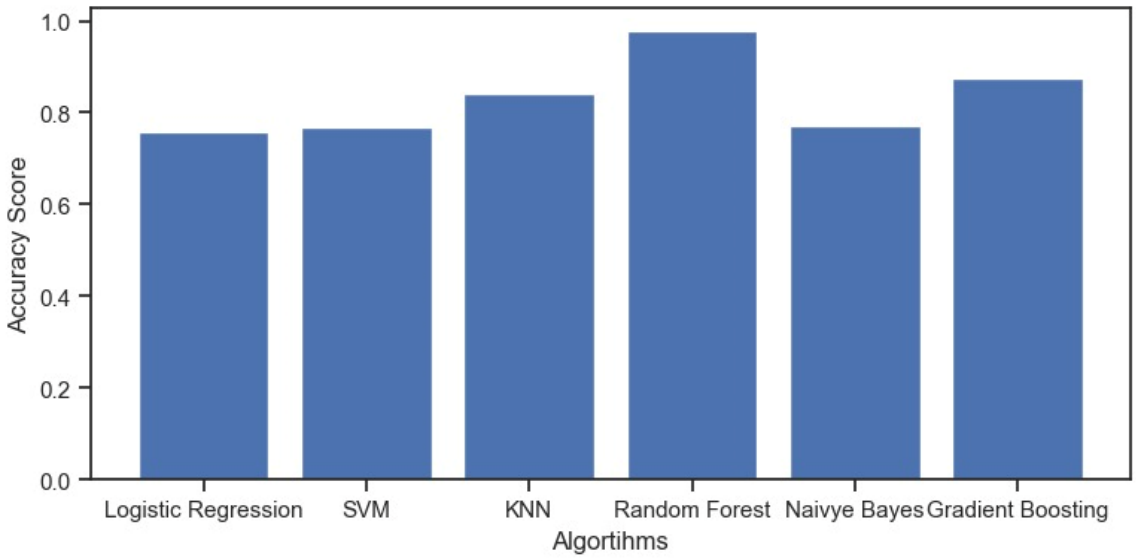
```
from sklearn.ensemble import GradientBoostingClassifier
clf=GradientBoostingClassifier(n_estimators=50,learning_rate=0.2)
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
ac=accuracy_score(test_y,y_pred)
acc.append(ac)

rc=roc_auc_score(test_y,y_pred)
roc.append(rc)
print("\nAccuracy {0} ROC {1}".format(ac,rc))

result=cross_validate(clf,train_X,train_y,scoring=scoring,cv=10)
display_result(result)
```

```
Accuracy 0.8761329305135952 ROC 0.8499680485661794
TP:  [29 31 24 33 30 31 29 29 29 30]
TN:  [87 91 90 89 86 87 86 88 81 87]
FN:  [11  8 16  7 10  9 11 11 11 10]
FP:  [ 6  2  2  3  6  5  6  4 11  5]
```

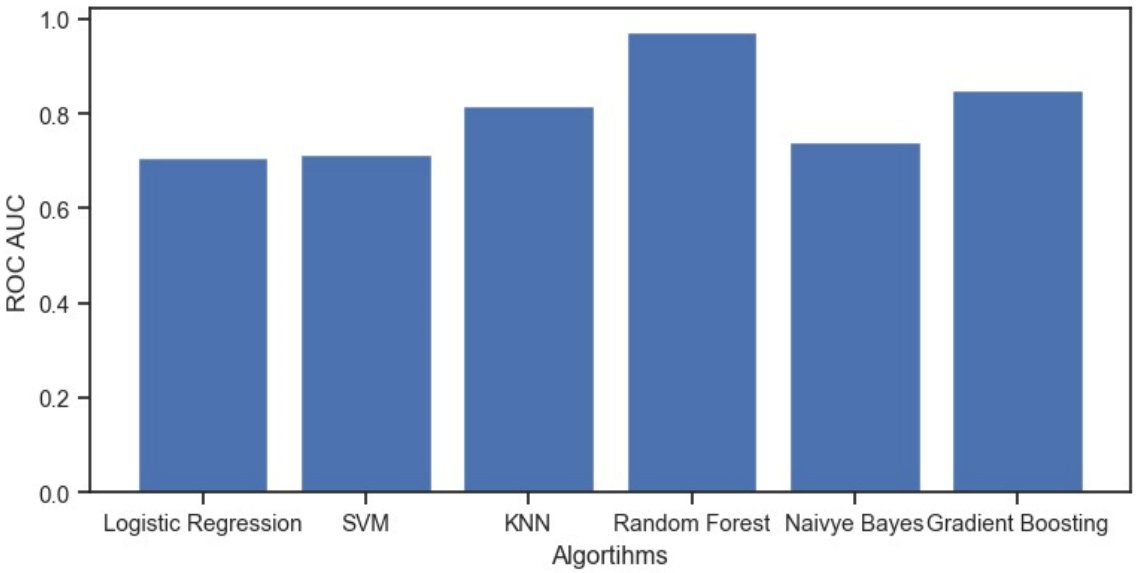
6. Result Visualization

We visualize the accuracy and ROC AUC scores of different models:

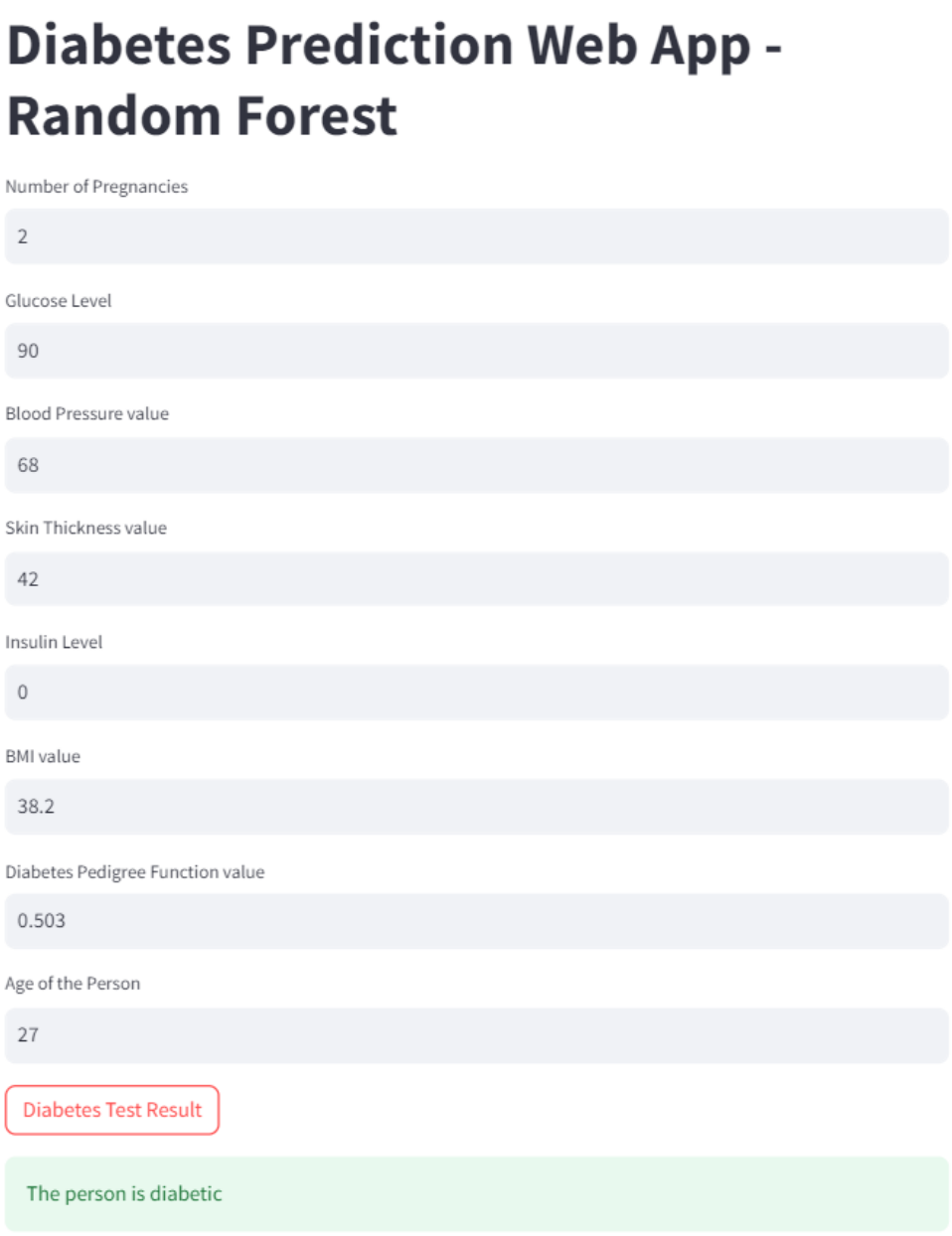
```
ax=plt.figure(figsize=(9,4))
plt.bar(['Logistic Regression','SVM','KNN','Random Forest','Naivye
Bayes','Gradient Boosting'],acc,label='Accuracy')
plt.ylabel('Accuracy Score')
plt.xlabel('Algortihms')
plt.show()
```




```
ax=plt.figure(figsize=(9,4))
plt.bar(['Logistic Regression','SVM','KNN','Random Forest','Naivye
Bayes','Gradient Boosting'],roc,label='ROC AUC')
plt.ylabel('ROC AUC')
plt.xlabel('Algortihms')
plt.show()
```



7. Web Interface



CHAPTER 5

RESULT AND CONCLUSION

5.1. Results:

In this work, various ensemble and classification techniques were implemented using Python to maximize the accuracy of predicting diabetic outcomes. The models evaluated include Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, Naive Bayes, and Gradient Boosting classifiers. The goal was to identify the model that performs best in terms of accuracy and ability to distinguish between diabetic and non-diabetic cases.

Accuracy Score

The Accuracy Score measures the proportion of correctly classified instances among the total instances. It is a straightforward metric that gives an overall sense of the model's performance.

Accuracy Score Analysis,

In Figure 1, the accuracy scores for different models are plotted

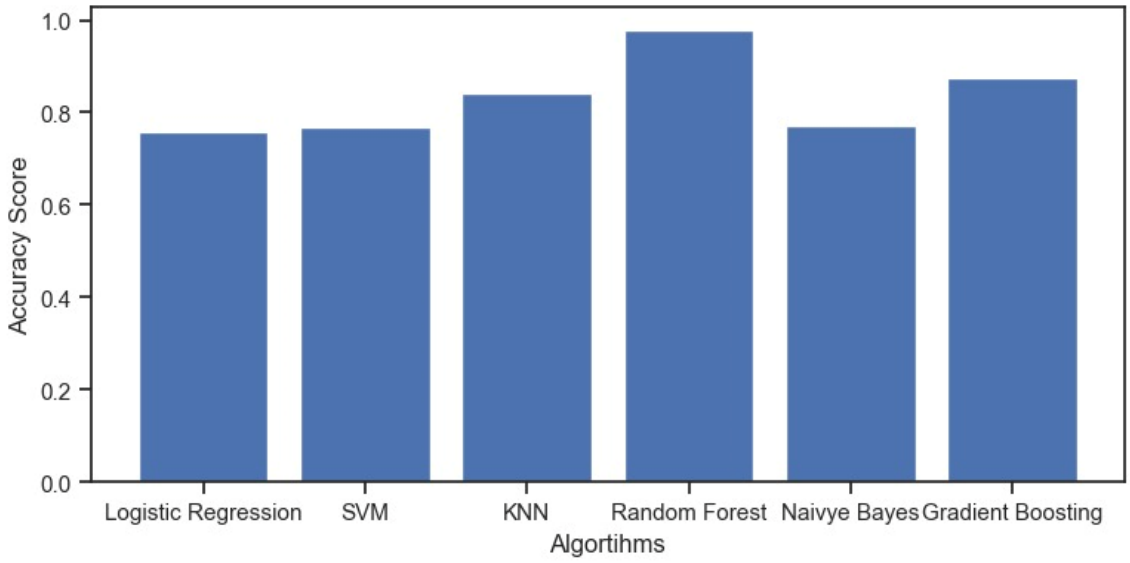


Figure 5.1.1: Performance Comparison of Machine LearningAlgorithms (Accuracy)

Random Forest:

This model achieves the highest accuracy, indicating it correctly classifies the most instances. This is consistent with its high ROC AUC score.

Logistic Regression:

This model also demonstrates high accuracy, reinforcing its robustness in classification tasks.

SVM and Gradient Boosting:

Both models show competitive accuracy, performing well in classifying instances.

KNN:

While its accuracy is reasonable, it does not match the performance of the top models.

Naive Bayes:

This model shows the lowest accuracy, reflecting its poor performance compared to other models.

The accuracy scores align with the ROC AUC results, highlighting that Random Forest and Logistic Regression are the most effective models for predicting diabetic outcomes.

Receiver Operating Characteristic Area Under the Curve (ROC AUC)

The ROC AUC is a performance measure used to evaluate the capability of a classification model to differentiate between classes. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The True Positive Rate, or sensitivity, is the ratio of correctly identified positive cases (diabetic patients). The False Positive Rate is the ratio of negative cases (non-diabetic patients) that are incorrectly classified as positive.

ROC AUC Analysis

In Figure 2, the ROC AUC curves for various models are displayed. The AUC value ranges from 0 to 1, where a value of 1 represents a perfect classifier and 0.5 indicates a model with no discriminative ability, equivalent to random guessing.

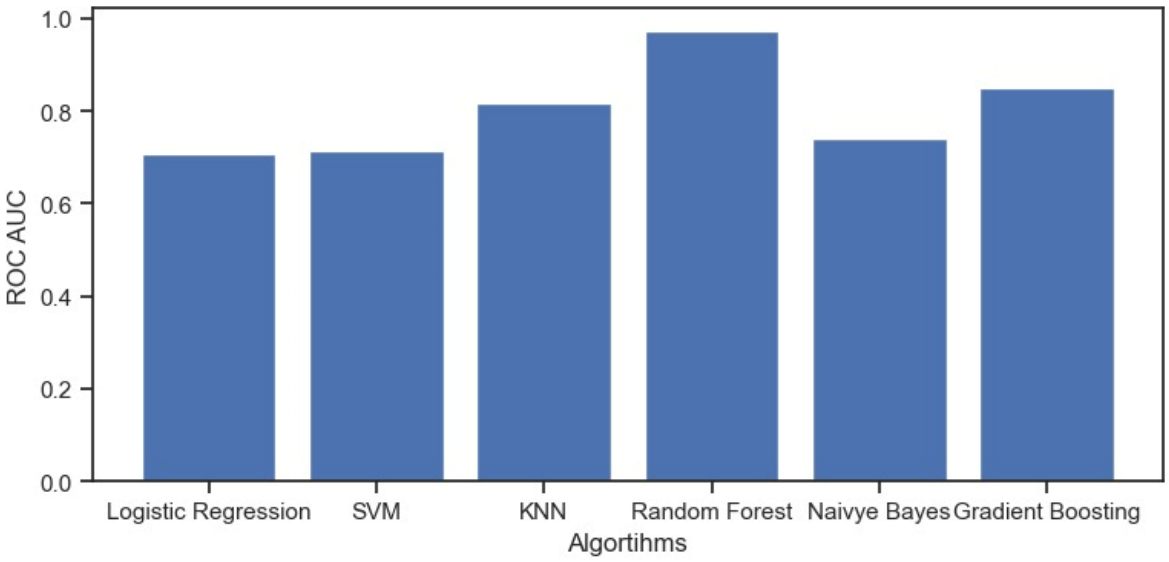


Figure 5.1.2: Performance Comparison of Machine LearningAlgorithms (ROC AUC)

5.2. Conclusion:

In conclusion, the project successfully achieved its primary objective of implementing machine learning techniques to predict diabetes as well as conducting analysis of the functionality. The suggested framework integrates ensemble learning and classification techniques, such as 'SVM', 'Decision Tree', 'Gradient Boosting classifiers', 'Random Forest', 'KNN' and 'Logistic Regression'. The experiment outcomes offer valuable insights for medical professionals, enabling them to make early predictions and informed decisions for diabetes treatment, ultimately contributing to saving lives.

Upon analyzing the performance metrics, it is evident that Logistic Regression and Random Forest classifiers stand out with the highest ROC AUC and accuracy scores. These models are highly effective in distinguishing between diabetic and non-diabetic cases. In contrast, Naive Bayes exhibits the poorest performance, indicating it is less suitable for this task. The evaluation underscores the importance of using robust machine learning techniques to achieve high prediction accuracy in medical diagnosis

CHAPTER 6

FUTURE SCOPE

1. Exploration of Deep Learning Techniques

Neural Network Architectures: Investigate the application of advanced neural network architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for the classification task. These architectures have shown promise in handling complex patterns in medical data.

Hyperparameter Tuning: Utilize automated hyperparameter tuning techniques like Grid Search and Random Search to optimize deep learning models for better accuracy and efficiency.

Transfer Learning: Leverage pre-trained models and fine-tune them on the diabetes dataset to improve performance and reduce training time.

2. Dataset Expansion and Enhancement

Larger Datasets: Incorporate larger and more diverse datasets to improve the generalizability of the models. This could involve merging multiple datasets from different sources or using publicly available medical databases.

Additional Features: Collect datasets with additional features such as genetic information, lifestyle factors, and more detailed medical history to enhance predictive accuracy.

Data Augmentation: Implement data augmentation techniques to artificially increase the size of the dataset and introduce variability, which can help in improving the robustness of the models.

3. Integration of Cloud Services

Amazon Web Services (AWS): Deploy the models on AWS to provide scalable and reliable access. Use AWS services like SageMaker for model training and deployment, Lambda for serverless computing, and S3 for data storage.

Other Cloud Platforms: Explore deploying the models on other cloud platforms like Google Cloud Platform (GCP) and Microsoft Azure to reach a wider audience and ensure redundancy.

API Development: Develop APIs using cloud functions to enable easy integration of the prediction models into various applications and services.

5. Model Interpretability and Explainability

Explainable AI (XAI): Implement techniques for model interpretability to ensure that the predictions made by the models can be understood by healthcare professionals and patients. Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be used.

Transparency in Decision-Making: Ensure transparency in the model's decision-making process to build trust among users and comply with regulatory standards.

6. Collaboration with Medical Experts

Interdisciplinary Research: Collaborate with medical professionals and researchers to validate the models and ensure they meet clinical standards. This will involve conducting rigorous clinical trials and studies.

User Feedback: Collect feedback from actual users and healthcare providers to continuously improve the models and applications based on real-world usage.

7. Compliance and Ethical Considerations

Data Privacy: Ensure strict adherence to data privacy laws and regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation).

Ethical AI: Develop ethical guidelines for the use of AI in healthcare to prevent bias and ensure fair treatment of all patients. Regular audits and evaluations of the models will be necessary to maintain ethical standards.

8. Customization for Diverse Populations

Personalized Medicine: Adapt models to cater to diverse populations by incorporating demographic-specific factors. This can involve creating specialized models for different age groups, ethnicities, and geographic locations.

Localization: Customize the applications for different languages and cultural contexts to make them accessible to a global audience.

9. Longitudinal Studies and Chronic Disease Management

Long-Term Monitoring: Implement longitudinal studies to track patients over time, enabling the models to provide more accurate predictions based on historical data.

Chronic Disease Management: Extend the models to assist in the management of other chronic diseases by integrating multi-disease prediction capabilities.

10. Educational and Training Programs

Training for Healthcare Providers: Develop training programs for healthcare providers to effectively use the prediction models and interpret the results.

Patient Education: Create educational resources for patients to understand their health data and the significance of the predictions made by the models.

11. Open Source and Community Involvement

Open Source Contributions: Release the code and models as open source to encourage community involvement and contributions. This will enable collaboration and innovation from a wider pool of researchers and developers.

Community Support: Establish forums and support channels for users and developers to share insights, troubleshoot issues, and collaborate on improvements.

By pursuing these future work initiatives, we aim to enhance the predictive accuracy, scalability, and accessibility of the diabetes prediction models, ultimately contributing to better healthcare outcomes and advancements in the field of medical AI.

REFERENCES

[1] **International Diabetes Federation, Diabetes Atlas, 3rd ed. Brussels, Belgium**: International Diabetes Federation.

[2] **Classification of Diabetes Patients Using Kernel-Based Support Vector Machines**, 'authored by **G.A. Pethunachiyar** and presented at the 2020 International Conference on Computer Communication and Informatics (ICCCI)

[3] **Brownlee, J. (2016c). Logistic regression for machine learning.**https://www.geeksforgeeks.org/understandinglogistic-regression/

[4] **Nai-Arun N, Moungmai R**., "Comparison of classifiers for the risk of diabetes prediction," Procedia Comput Sci., 2015:69:132–42.

[5] **Saravananathan K, Velmurugan T**., "Analyzing diabetic data using classification algorithms in data mining," Indian J Sci Technol., 2016:9:1–6.

[6] **K. Anandha Kumar**, "A survey on diabetes mellitus prediction using machine learning techniques," International Journal of Applied Engineering Research, vol. 11, 2022.

[7] **S. V. K. R. Rajeswari and P. Vijayakumar**, "Prediction of diabetes mellitus using machine learning algorithm," Annals of the Romanian Society for Cell Biology, vol. 25, pp. 5655–5662, 2021.

[8] **Shafi S, Ansari GA**. "Early prediction of diabetes disease & classification of algorithms using machine learning approach." In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). Available from SSRN 3852590 (2021).

[9] **Sadhu A, Jadli A.** "Early-stage diabetes risk prediction: A comparative analysis of classification algorithms." Int Adv Res J Sci Eng Technol (IARJSET) 2021;8:193–201.

[10] **R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki et al**., "A novel diabetes healthcare disease prediction framework using machine learning techniques," Journal of Healthcare Engineering, vol. 2022, Article ID 1684017, 10 pages, 2022.

[11] **K. J. Rani, "Diabetes prediction using machine learning**," International Journal of Scientific Research in Computer Science Engineering and Information Technology, vol. 6, pp. 294–305, 2020

[12] **S. Kodama, K. Fujihara, C. Horikawa et al**., "Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: a meta-analysis," Journal of Diabetes Investigation, vol. 13, no. 5, pp. 900–908, 2022.

[13] **I. Tasin, T. U. Nabil, S. Islam, and R. Khan**, "Diabetes prediction using machine learning and explainable AI techniques," Healthcare Technology Letters, vol. 10, pp. 1–10, 2023.

[14] **M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan**, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020.

[15] **Mounika, V. , Neeli, D.S. , Sree, G.S. , Mourya, P. , Babu, M.A**. : Prediction of type-2 diabetes using machine learning algorithms. In: International Conference on Artificial Intelligence and Smart Systems, pp. 127–131 (2021)

[16] **Olisah, C.C. , Smith, L. , Smith, M**. : Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Comput. Methods Programs Biomed. 220, 1–12 (2022).

[17] **Raja Krishnamoorthi, Shubham Joshi, and Hatim Z. Almarzouki,** "A Novel Diabetes Healthcare Disease Prediction Framework using Machine Learning Techniques," Journal of Healthcare Engineering, pp. 1-10 2022.

[18] **Tigga, N. P., and Garg, S. (2020)**. Prediction of type 2 diabetes using machine learning classification methods. Proc. Comp. Sci. 167, 706–716. doi: 10.1016/j.procs.2020.03.336

[19] **Sehly, R., and Mezher, M. (2020).** "Comparative analysis of classification models for pima dataset," in International Conference on Computing and Information Technology (ICCIT-1441), 1–5. doi: 10.1109/ICCIT-144147971.2020.9213821

[20] **Patil, M. K., Sawarkar, S. D., and Narwane, M. S. (2019**). Designing a model to detect diabetes using machine learning. Int. J. Eng. Res. Technol. 8, 333–340. Available online at: https://www.ijert.org/designing-a-model-to-detect-diabetes-using-machine-learning

[21] **Aishwarya Mujumdar, V Vaidehi**, "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science, Volume 165, 2019.

[22] **Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar**," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017

[23] **B. Nithya and Dr. V. Ilango**," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017

[24] **M. Gollapalli, et al**. A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM Comput. Biol. Med., 147 (2022), Article 105757