

**A STUDY FOR DIABETES PREDICTION USING
HYPERPARAMETER TUNING AND MACHINE
LEARNING TECHNIQUES**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE AWARD OF DEGREE
OF
MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by:

TARUN KHARKWAL
2K20/SWE/24

Under the supervision of
MS SHWETA MEENA
(Assistant Professor)



DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

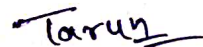
Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, Tarun Kharkwal, Roll No. 2K20/SWE/24 student of M. Tech (Software Engineering) hereby declare that the project Dissertation titled "A Study for Diabetes Prediction using Hyperparameter Tuning and Machine Learning Techniques" which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 27/05/2022



TARUN KHARKWAL

2K20/SWE/24

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “A Study for Diabetes Prediction using Hyperparameter Tuning and Machine Learning Techniques” which is submitted by Tarun Kharkwal, 2K20/SWE/24 Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 27/5/2022

Shweta
27/5/2022

MS. SHWETA MEENA

SUPERVISOR

Assistant Professor,

Department of Software Engineering,

Delhi Technological University

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

ACKNOWLEDGMENT

The success of this project requires the assistance and input of numerous people and the organisation. I am grateful to everyone who helped in shaping the result of the project. I express my sincere thanks to Ms Shweta Meena, my project guide, for providing me with the opportunity to undertake this project under her guidance. Her constant support and encouragement have made me realise that it is the process of learning which weighs more than the end result. I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout with new ideas, provided information necessary and pushed me to complete the work.

I also thank all my fellow students and my family for their continued support.

Tarun

TARUN KHARKWAL

2K20/SWE/24

ABSTRACT

Diabetes is the most frequent metabolic condition that causes excess sugar levels in our blood. Patients with diabetes have a body that can't metabolise insulin adequately or can't make enough insulin. Diabetes is a disease which occurs when the glucose level increases in blood. It is a persistent disease that occurs mainly in two ways: First one is, if the adequate insulin is not produced by the pancreas and the second one, if insulin is not used by the body effectively. Insulin Hormone is responsible for regulating Blood sugar. Diabetes can harm our body parts too like eyes, kidneys, nerve, heart and blood vessels. Therefore Predicting diabetes in the earlier phase is very essential to control the diabetes and to save lives. In this study first we did a survey on the previous studies on this topic and after that we implemented our model on the basis of the survey.

Presenting a method of detection by symptoms that the person might observe may motivate the person to seek medical treatment more immediately, concluding in a more precise diagnosis and treatment. In this study, We have taken 35 papers (studies in the time span of 2014-2021) out of which we chose 23 papers for further study based upon our requirements. And then we analysed current research in order to conclude the risk factors for diabetes. This research investigates the accuracy of diabetes prediction. It is concentrated on current advancements that have a significant influence on diabetes diagnosis and detection. and we also see by using which medical information and Machine Learning techniques we can predict better. On the basis of our small survey we concluded that RF is the popular technique used among all techniques and PIMA Indians Dataset is used more frequently. And among all used techniques we found out that boosting and SVM has the highest accuracy.

Furthermore, We have used in this research logistic regression, KNN, DT, naive bayes, RF and SVM classifiers. After these techniques we tried to optimise our model. We found that all techniques with optimised models after smote performed very well. and for KNN and RF we achieved the highest accuracy.

Then we used hyper parameter tuning to optimise our KNN and RF classifier, and we also tuned XGB. Firstly For RF model using GridsearchCV optimization we achieved accuracy nearly 89.21%. And for XGB after hyper parameter tuning we got an

accuracy of 90.13 percent. Which is the best accuracy among all the ML techniques that we have used. As a result we can say that among all the ML models XGB boost after hyper parameter tuning is performed very well.

CONTENTS

Candidate's Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Contents	vi
List of Figures	viii
List of Tables	ix
List of Abbreviations	x
CHAPTER 1 INTRODUCTION	1
1.1 Artificial Intelligence and Machine Learning	1
1.1.1 Supervised Learning	2
1.1.2 Unsupervised Learning	3
1.1.3 Reinforcement Learning	3
1.2 Motivation	3
1.3 Objective	4
1.4 Diabetes and Its Types	4
1.4.1 Type-1 Diabetes (IDDM)	5
1.4.2 Type-2 Diabetes (NIDDM)	6
1.4.3 Gestational Diabetes	6
1.5 Project Overview	6
1.6 Thesis Outline	7
CHAPTER 2 RELATED WORK	8
2.1 Literature Review	8
2.2 Insights from Literature Review	14
CHAPTER 3 DATASET AND TECHNIQUES USED	18
3.1 Dataset Used	18

3.1.1	Data Preprocessing	19
3.1.2	Data Normalisation	19
3.1.3	Data Balancing using Smote	20
3.1.4	Hyperparameter Tuning	21
3.2	Insights From The Heatmap	21
3.3	Techniques Used	23
3.3.1	Logistic Regression	23
3.3.2	K-Nearest Neighbours Classifier	23
3.3.3	Decision Tree	24
3.3.4	Random Forest	24
3.3.5	Naive Bayes	24
3.3.6	Support Vector Machine	25
3.3.7	Gradient Boosting	25
	CHAPTER 4 EXPERIMENTAL SETUP	26
4.1	Hardware Used	26
4.2	Software Used	26
4.3	Framework of Our Work	27
4.4	Methodology Used	28
	CHAPTER 5 RESULTS AND DISCUSSION	29
	CHAPTER 6 CONCLUSION AND FUTURE WORK	36
6.1	Summary	36
6.2	Threats to Validity	36
6.3	Validation of Prediction Techniques	37
6.4	Conclusion	37
6.5	Future Scope	38
	REFERENCES	39
	LIST OF PUBLICATIONS	43

LIST OF FIGURES

1.1	Artificial Intelligence and ML	2
2.1	Dataset vs Frequency	17
3.1	Head of our dataset	18
3.2	Comparison of “Outcome” attribute in dataset	19
3.3	Outcome class in dataset after using SMOTE	20
3.4	Smote applied on the dataset	20
3.5	Heatmap of dataset	22
3.6	Blood pressure vs diabetes in function of BP and Age	22
4.1	Framework	27
5.1	Comparison of accuracy before smote for normal and optimised ML algorithm	30
5.2	Comparison of accuracy after smote for normal and optimised ML algorithm	31
5.3	Comparison of accuracy among optimised ML algorithm before smote and after smote	33
5.4	Comparison of training and testing score for KNN	34
5.5	Comparison of RF and XGB after parameter tuning	35

LIST OF TABLES

2.1	Datasets and Techniques	15
2.2	Country Vs Number of Studies taken	15
2.3	Frequencies of Classifiers	16
2.4	Maximum Accuracy vs Model Name	17
5.1	Comparison table for accuracy before smote without/with optimization	29
5.2	Comparison table of accuracy after smote without/with optimization	31
5.3	Comparison table for datasets before and after optimization and smote	32

LIST OF ABBREVIATIONS

1. ML : Machine learning
2. IDDM : Insulin Dependent Diabetes Mellitus
3. NIDDM : Non Insulin Dependent Diabetes Mellitus
4. KNN : K- Nearest Neighbour
5. DT : Decision Tree
6. RF : Random Forest
7. XGB : Gradient Boosting
8. SMOTE : Synthetic Minority Over-sampling Technique
9. AI : Artificial Intelligence
10. SVM : Support Vector Machine

CHAPTER 1

INTRODUCTION

Diabetes mellitus (alias diabetes) is one of the most perilous diseases. Many people in the world are afflicted from diabetes. It is responsible for a major health care problem that is exploding in the world. This disease does not spare even the youngsters. It has been researched that people with unhealthy diets and obesity are at higher risk of generating this disease. If a doctor diagnoses these symptoms in the patients earlier then he will recommend a good diet, exercise and medicines for the treatment and the patient can be saved. Predicting diabetes in the starting phase is very essential to control the diabetes and to save lives. So as we know for prediction problems ML is widely used in the world. For the same purpose we worked on diabetes prediction using ML. Our ultimate aim is: “We have given the details of patients features using that we have to check whether a patient has diabetes or not.”

1.1 Artificial Intelligence and Machine Learning

AI refers to a machine's capacity to mimic human actions. AI is developed by studying and analysing how a human brain learns, determines, and functions when attempting to solve a problem. The study of intelligent agents is characterised as an AI research area. The agents observe their surroundings and execute activities in order to reach a goal. ML and Deep Learning are subcategories of AI, allowing it to be more versatile and efficient. ML is a branch of AI that has a well stated purpose. It's a computer science subfield. In essence, It is the process through which a machine learns to do actions on its own, such as making decisions based on previously collected data, in a human-like fashion. In order to achieve this function in action, Its algorithms recognise patterns in a dataset and build a model that can subsequently be improved in order to anticipate actions based on fresh data. Machines may be programmed to do tasks in a variety of fields. One

method that technology is being used to assist diabetes predictions is through its algorithms, which include data interpretation.

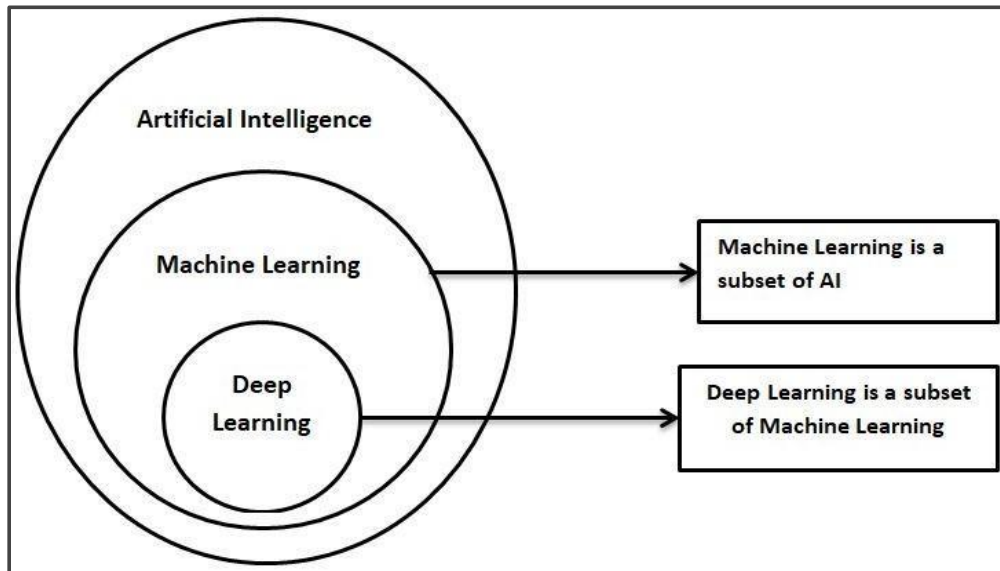


Figure 1.1 Artificial Intelligence and Machine Learning

The capacity to forecast whether or not a person will develop diseases like diabetes is governed by the ability to make predictions. Predictive analysis is a machine learning-based approach for forecasting data. This methodology includes ML algorithms, which are data extraction procedures and statistical methodologies for predicting future events. In the medical field, these predictions can be used to forecast a diagnosis as precisely as possible. ML may be divided into three categories. Let's take a look at each of the three strategies one by one.

1.1.1 Supervised Learning

We give an input (x) and its matching output variable (y) in supervised learning, and we run an algorithm that learns to execute mapping functions from input to output.

$$Y = f(x) \quad (1.1)$$

The basic goal of this is to effectively estimate the mapping function as output variables can be anticipated when fresh input data is added. The machine is trained using labelled data in this process of an algorithm learning using the training dataset.

1.1.2 Unsupervised Learning

Unsupervised learning is the use of AI systems to detect patterns in data sets that include no classified or labelled data items. Only the input data (X) is available in unsupervised learning, whereas the associated output variables are absent. Its main goal is to reveal hidden patterns in data. Unsupervised learning teaches the agent to recognise essential elements in incoming data and then group them together based on their comparable behaviour or traits.

1.1.3 Reinforcement Learning

In reinforcement learning strategy we reward those actions which are desirable while we penalise undesirable ones. Rather of being explicitly taught, an agent in reinforcement learning learns through interacting with an environment and watching the effects. Reinforcement learning is based on the premise that an agent will learn from its actions and will be rewarded for doing the correct thing.

1.2 Motivation

Diabetes affects 537 million people between the ages of 20 and 79. The entire population of diabetes is anticipated to reach 643 million by 2030, and 783 million by 2045. Diabetes claims the lives of over 6.7 million people each year. Over time, a high glucose level in the blood could lead to a variety of complications, including heart problems, renal failure, vision problems, dental problems, nerve damage, and foot problems. increased thirst and urine, increased appetite, tiredness, fuzzy vision, significant impairment in the hands or feet, unhealed wounds, and weight loss are all indicators that something is wrong with you if you have diabetes. One method that technology is being used to assist diabetes predictions is through ML algorithms, which include data interpretation. The capacity to forecast whether or not a person will develop diseases like diabetes is governed by the ability to make predictions. Predictive analysis

is a machine learning-based approach for forecasting data. This methodology includes ML algorithms, which are data extraction procedures and statistical methodologies for predicting future events. In the medical field, these predictions are used to forecast a diagnosis as precisely as possible. ML is an area of AI that has a well-defined purpose. It's a computer science subfield. In essence, ML is the process through which a machine learns to do activities on its own, such as making decisions based on previously collected data, in a human-like fashion.

1.3 Objective

The following objectives will be considered while the research study is carried out:

1. To give a brief introduction about ML in the field of diabetes prediction like what are the studies that have been done in this field.
2. To implement various ML algorithms to predict diabetes. The different algorithms that have been used by us in this research are logistic regression, KNN, DT, naive bayes, RF and SVM classifiers.
3. To balance our dataset using SMOTE (Synthetic Minority Oversampling Technique) and to compare accuracy of each model using SMOTE and without using SMOTE. And also tried tuned hyper parameters.
4. To compare the accuracy and efficiency of ML algorithms with PIMA Indian Dataset and SMOTE dataset.

1.4 Diabetes and Its Types

Diabetes mellitus (alias diabetes) is the most perilous disease. It is responsible for a major health care problem that is exploding in the world. This disease does not spare even the youngsters. Diabetes occurs when the pancreas, which is a human body organ in the body, is not able to produce adequate insulin or if the tissues and cells in the body are unable to use the insulin generated. More than 72 million people in India have diabetes apart from this lots of people are under risk of having diabetes. Our major source of

bioenergy is blood glucose, which we get from the food that we have. Pancreas is responsible for producing insulin hormone, and insulin aids glucose absorption into cells as energy. Often it happens that our body is not able to produce adequate insulin and also it does not manage it properly. Glucose always remains in our blood and as a result does not reach our cells. Having lots of glucose in our blood might lead to health issues in the long run. Although there is no medical breakthrough, we can monitor and manage it and be healthy. Now the question arises what are the factors for causing it? These are the factors responsible for diabetes: genes, environment, food and lifestyle. Type-1, Type-2, and Gestational diabetes are kinds of it.

1.4.1 Type-1 Diabetes (IDDM)

Diabetes type-1 which is indicated by the deficiencies of insulin also known as Insulin Dependent Diabetes Mellitus. Our body is not able to produce insulin if we have type-1 diabetes, and the resistant system picks up and damages the cells that are responsible for insulin generation in our pancreas. These cells, known as beta cells, are found on tiny islands of endocrine cells called pancreatic islets, together with other types of cells. These cells generate insulin, a hormone that helps in the movement of glucose from meals into cells throughout the body, where it is used for energy. However, if these cells are killed, as a result no insulin is created, and the glucose remains in the bloodstream, causing significant damage to all of the body's organ systems. It is most commonly observed in children and adolescents, although it can affect anybody at any age. People must take insulin on a regular basis to stay alive.

Experts are unsure how this causes type-1 diabetes, but they feel it's a combination of autoimmune, genetic, and environmental factors. Type-1 diabetes has a greater probability of developing than almost all other serious chronic illnesses in children. Girls are around 10 to 12 years old, while guys are around 12 to 14 years old. According to the Juvenile Diabetes Research Foundation, It affects as many as 3 million Americans. As per the above discussion, it is a serious medical concern in the world. The medical profession has done a lot of study into how to cure this condition.

1.4.2 Type-2 Diabetes (NIDDM)

Type 2 which is the body's insulin resistance alias Non-Insulin Dependent Diabetes Mellitus. If we talk about this, our body does not produce or utilise insulin well in this type. Age doesn't matter for this type of diabetes anyone can be affected, including youth. This kind of diabetes is more common among middle-aged and elderly people. This is frequently seen in patients.

1.4.3 Gestational Diabetes

Some women get this kind of diabetes while pregnant. After birth, this kind of diabetes normally goes away. However, if you've had gestational diabetes, you're more likely to get type 2 diabetes later in life. It's conceivable that type-2 diabetes develops during pregnancy.

1.5 Project Overview

On diabetes data, there has been a lot of ML research. especially using the Pima Indian Diabetes dataset from the UCI repository. ML algorithms for post-diagnosis treatment have aroused a lot of interest, including forecasting blood glucose levels to adjust insulin dose and using to forecast by association rules. the prevalence of specific illnesses in diabetes patients. However, unlike both of these approaches, we use a dataset that is not restricted to a specific religion but to a specific type of diabetes, and our objective is to train a model to forecast the emergence of this form of diabetes using the patient's prior medical records, not to monitor diabetic individuals. Using subject test results, forecast diabetes diagnoses from non-diabetic instances, we principally used these supervised ML classifiers : Logistic Regression, DT, RF, Naive Bayes, SVM, KNN, XGB. Developing a strong prediction model was not straightforward, despite the fact that this is not a well-known difficult topic for physicians. Next, To forecast diabetes, we utilised the same base classifiers, However, the features were alterations in test findings across subsequent same types of testing for an individual. This method holds the potential

of predicting whether or not someone is at risk for diabetes before any test results indicate they do. SVM, RF, KNN and XGB classifiers gave interesting results. Surprisingly, To do this, it was required to deliberately embed absent properties to get over 90% accuracy in diabetes prediction. And we achieved it for the XGB classifier after hyper parameter tuning. The fact that variations in test results over time might predict diabetes with roughly 80% accuracy without utilising data from the previous time period preceding diagnosis as diabetic was of particular interest. Another component that was explored was whether oversampling of minority class instances would increase the classifiers' prediction accuracy. The oversampling technique employed was the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic cases using a closest neighbour approach. and the oversampling strategy also enhanced overall prediction accuracy (using XGB and SVM).

1.6 Thesis Outline

There are six sections in this dissertation. First chapter provides an introduction, while the subsequent chapters thoroughly discuss the recommended techniques.

Chapter 1 This Chapter discusses background, motivation, problem statement, objectives, and importance. Aside from this chapter, there are six others in the thesis:

Chapter 2 This chapter looks at the various studies and related work carried out in order to get the required information about the diabetes detection using ML algorithms.

Chapter 3 This chapter tells about the dataset and ML techniques that we have used to achieve the proposed work. In the first section of this chapter we will discuss the dataset thoroughly and in the next section we will discuss all the ML techniques and optimization techniques that we have used in this research.

Chapter 4 This section provides a brief summary of the hardware and software tools utilised in the proposed work. In this section, we will also discuss our performance measures, framework of this study and methodology.

Chapter 5 This chapter deals with results and discussions of the proposed study. It also elaborates on model classification performance using a variety of evaluation measures.

Chapter 6 This chapter summarises the works and provides conclusions. The future scope is also discussed so that further improvement can be done.

CHAPTER 2

RELATED WORK

Data science has recently acquired popularity in predicting the probability of diabetes disease. Various toolkits and algorithms have been investigated and developed by researchers. These toolkits and algorithms demonstrate the immense potential of this study subject. A significant number of researches has been conducted in diabetes prediction. Many algorithms have been proposed to use ML to automate diabetes prediction. The suggested model relies mostly on the most well-known diabetes prediction algorithms discovered via our survey. This chapter presents a survey of studies dealing with diabetes prediction in humans. And also we will discuss the insights that we have found out from the literature review.

2.1 Literature Review

This literature evaluation serves as a solid foundation for future work on this project. And also It will tell us about all the previous studies that have been done on this topic.

A. S. Alanazi et al. (2020) [1] presented a model in which the author combined two ML algorithms which were SVM and RF to predict diabetes. They collected diabetic patients data from the Security Force Primary Health Care. The data was carefully checked for missing or empty values. The data was deemed to be clean enough to move forward. The author concluded the article by saying, "The capacity of data science is to expedite early diabetes diagnosis." Diabetes may be diagnosed using ML techniques such as SVM and RF. According to the data, the RF algorithm has a predicted rate of 98% and a precision of 100%."

N. E. Costea et al. (2021) [2] offered a comparison of three ML algorithms: SVM, Naive Bayes, and RF. By putting their diabetes prediction abilities to the test on two publicly available databases: Pima Indians and Diabetes Dataset 2019. The goal of his

study is to examine different ways to acquire improved accuracy by analysing the performance of the algorithms using different measures. They determined that the SVM and RF achieved an accuracy of 80%.

D. Dutta et al. (2021) [3] investigated the key aspects of diabetes using three algorithms: Logistic Regression, SVM, and RF. Also discovered what are the key elements in the development of diabetes, In fields of application where data comprising dozens or even hundreds of factors are accessible, factor and feature identification has become the subject of major study. Similarly, the author focused on the most important characteristics to determine if a person may get diabetes in the future. The article concluded by stating that "RF is the best model for estimating diabetes," with an accuracy of roughly around 84 percent."

P. Sonar and K. Jaya Malini [4] presented model construction using classification approaches like DT, ANN, Naive Bayes, and SVM. The purpose of this study is to develop a system that can better predict a person's mellitus vulnerability. The study concluded with the observation that DT models provide precisions of 85 percent, Naive Bayes models provide precisions of 77 percent, and SVM models provide precisions of 77.3 percent. "SVM are really good when we have no notion about the data," the author stated in the research. Even with unstructured and semi-structured content kinds like text and images. The disadvantage of the SVM approach is that various essential parameters must be accurately specified in order to produce the best classification results for each particular scenario. The DT is simple to grasp and govern. There seems to be volatility in DT, which means that little modifications to the data model of the ideal DT can produce enormous changes. They're usually wrong. Naive Bayes is a robust algorithm that manages missing values by disregarding the probability estimate process. Concerned about how inputs are prepared. When the quantity of training datasets is increased, there is a risk of bias. Artificial Neural Network makes accurate predictions and is simple to apply. Dealing with massive data and complicated models is difficult. It takes a long time to process."

S. K. Reddy et al. (2021) [5] utilised the PIMA Indians Dataset to show two models: RF and KNN. The developers of the suggested system employed supervised learning of ML techniques and parameter adjustment utilising optimization approaches to get accurate diabetic outcomes. The accuracy of both models is 78.4 percent and 80.8 percent, respectively.

N. Mohan et al. (2020) [6] introduced the SVM technique and used it to predict diabetes. In this paper, the actual output of SVM kernels on diabetes dataset for diabetes detection is novel. This article compares the results of SVM kernels. The classifier is constructed using four different SVM kernels, and the accuracy of its predictions is determined using a testing set. Four kernels are used to test the SVM: linear, polynomial, sigmoid, and RBF. For various kernels, the efficiency of the SVM algorithm is investigated. For prediction, the strongest kernel is selected and employed. And the author claimed that the RBF kernel was the finest of the four.

A. M. Posonia et al. (2020) [7] suggested a ML knowledge for diabetes forecasting, such as DT J48 computation. DT is amongst the most effective classifiers. There were 768 patient records in the dataset, each with eight key features and a target category with the outcome "Positive" or "Negative." Weka was used for the experiment, and the results reveal that DT J48 computation provides higher efficiency with less processing time. For the study, the author used the Pima Indians Diabetes Dataset. The planned research task is 91.2 percent efficient.

G. A. Pethunachiyar (2020) [8] developed SVM models with different kernels and stated that SVM is the most often used approach in diabetes patient prediction. It assesses the performance of SVM kernel functions based on their accuracy level. The SVM with Linear Kernel outperforms the other two kernel functions, according to the results. The SVM with Linear Kernel yielded 100%, the SVM with Radial Kernel function achieved 95%, and the SVM with Polynomial Kernel gave 90% for the supplied sample group.

V. Mounika et al. (2021) [9] proposed ML methods such as diabetes were predicted using logistic regression and RF classifications. The same algorithms were used to revamped the PIMA Indians dataset. For the PIMA dataset, RF often delivers the maximum accuracy. While employing the two independent learning algorithms, all of the algorithms achieved good results for specific metrics such as accuracy, sensitivity to recall, and so on. They compared all of the classifier predictions, and the study concluded that RF and Logistic Regression were 95 percent and 97 percent accurate, respectively.

M. A. R. Refat et al. (2021) [10] presented various ML models to predict the diabetes and found out that XGB has 100% accuracy and apart from that also concluded that RF and LSTM model has second highest accuracy 92.3%.

H. Abbas et al. (2019) [11] used ML to predict the continued prospects of NIDDM using data from the San Antonio Heart Study. To create the forecasting models, they used

SVM and ten factors that are universally acknowledged in the research as significant important predictors of diabetes. They used 10-fold cross-validation to fit the classifier and a retain dataset to verify it because the data was unbalanced in terms of class labelling. The validation accuracy in this study was 84.1 percent, with a mean recall rate of 81.1 percent across 100 iterations. The latest results might help identify groups at high risk of developing NIDDM in the upcoming times. The investigation concluded with the observation that employing SVM attained an accuracy of 84 percent.

K. Drisset al. (2020) [12] proposed a method based on three major steps: cleaning, modelling, and narrative. The first step is to do an imputation technique to eliminate missing data. The KNN method is then used to categorise patients. Two metrics, the F1 score and the ROC curve, were employed to predict the performance of the suggested strategy. The F1 score and ROC curve clearly distinguish diabetes and non-diabetic patients. Author chose k's value 11 in KNN technique for better results, and the study concluded with the statement that employing KNN produced the greatest F1 score of 83 percent.

J. Ma et al. [13] suggested several ML methods. To create a diabetes prediction model, the author employed six traditional ML methods, including logistic regression, SVM, DT, RF, boosting, and NN. The author used data from the UCI ML Repository, which was acquired by direct questionnaires from Sylhet Diabetes Hospital patients in Sylhet, and endorsed by a physician. The author tuned the parameters of each model to achieve a balance of accuracy and complexity. According to the testing error, RF, boosting, and NN outperformed logistic regression, SVM, and DT. The author ended by stating that the NN of the test dataset reaches 96 percent accuracy, making it the best model among these models for predicting diabetes.

K. Vijiya Kumar et al. (2019) [14] developed a model that provides the RF approach to reliably diagnose diabetes in a person. RF technique is a type of ensemble method that is often used for regression and classification issues. The exactness is higher when compared to other techniques. The proposed model offers the best diabetic predictive performance, and the data showed that the prediction system is capable of forecasting the diabetes condition effectively, efficiently, and, most significantly, accurately and instantaneously. The author used PIMA Indians dataset and ended by saying that RF algorithms were investigated and assessed on several measures throughout this paper.

L. V. R. Kumari et al. (2021) [15] provided many ML models, and the goal of this study is to develop a much more accurate early sign of diabetes by employing a range of ML algorithms. By constructing models using patient information, ML approaches increase diabetes diagnosis. This paper uses methods such as Naive Bayes, KNN, Logistic Regression, and RF. Both have various levels of precision. According to the author's results, the KNN model has 78.57 percent accuracy, 87 percent precision, and 72 percent specificity. The logistic regression model has a 72 percent accuracy rate, an 84 percent precision rate, and a 63 percent specificity rate. The Naive Bayes model has a 71 percent accuracy rate, an 81 percent precision rate, and a 60 percent specificity rate. The RF model has a 76 percent accuracy rate, an 84 percent precision rate, and a 67 percent specificity rate. Among them, the study concluded that the KNN algorithm delivers an accuracy of 78.57 percent.

S. Ghane et al. (2021) [16] discussed numerous ML models and recommended ML algorithms that assist predict diabetes, including KNN, SVM, DT, RF, Light XGB Machine (LGBM) and Adaboost. All of these algorithms are constructed utilising the Pima Indian Diabetes dataset and numerous characteristics such as glucose, skin thickness, insulin, age, and so on that aid in the detection of diabetes. We trained our models and found that LGBM outperformed all others, with an accuracy of 89.85 percent and an AUC of 0.95. As a result, LGBM is a more effective algorithm for identifying diabetics from non-diabetics. and concluded with the remark that the Neural Network has the highest accuracy of 82.54% for practice fusion dataset and for pima indians dataset DT has the highest accuracy of 72.6%.

P. S. Kohli et al. (2018) [17] described and applied a variety of ML models and classification methods, each with its own advantage, on three distinct sickness databases (diabetes, heart disease, and cancer) available in the UCI repository for prediction and diagnosis. To select attributes for each dataset, reverse modelling using the p-value test was utilised. The study's results support the hypothesis of utilising ML to detect diseases early on. The paper concluded that for the PIMA Indians dataset, Logistic Regression achieved the highest accuracy of 82.46 percent.

R. Akula et al. (2019) [18] showed several ML models and discovered that all techniques except Naive Bayes had very poor accuracy. As a result, they took it one step beyond and merged all of the methods into a weighted mean or softer polling ensemble classifier, in which each model contributes to a simple majority in assessing whether or not a patient has diabetes. The Ensemble learning on Practice Fusion is 85 percent

accurate. By far the most new method in this area is our ensemble technique. We are confident that the weighted mean ensemble approach not only performed well in overall metrics, but also aided in the recuperation of faulty forecasts and precise NIDDM prediction. Our precise innovative model can serve as a warning to people to seek medical attention as soon as possible. The study concluded by stating that the RF model, with an accuracy of 98 percent, is the best model across all models on the Early stage diabetes risk prediction dataset.

B. S. MURTHY et al. (2021) [19] showed various ML models and they conducted a large-scale investigation on diabetes datasets using ML methods such as DT Classifier and Logistic Regression. To choose the best forecasting algorithm, both methods are investigated and contrasted using a variety of criteria, including recall, accuracy, precision, specificity, and sensitivity. Following the execution results produced in this study, it is clear that Logistic Regression provides more accuracy than DT Classifier. As a result, it is possible to infer that Logistic Regression is the best diabetes prediction algorithm and that it aids in the provision of appropriate medicine. Finally, the article concluded that the SVM is the best model among all models, with an accuracy of 78.2 percent on PIMA Indians datasets.

M. Rady et al. (2021) [20] offered a machine learning-based solution to the problem. On a data set of 521 individuals, they used eight methods. The outcomes are compared in order to determine the optimal algorithm for this assignment. The techniques utilised were from a variety of families, including logistic regression, SVM-linear and nonlinear kernel, RF, DT, XGB, KNN, and naive bayes. The results demonstrate that RF has a distinct edge, with an accuracy of 98 percent after utilising 20 percent for testing and 80 percent of the dataset for training, and determined that the RF model obtained 98 percent accuracy.

A. Mir et al. (2018) [21] presented many models with the purpose of constructing a classifier method to forecast diabetes utilising Naive Bayes, SVM, RF, and the Simple CART method using the WEKA tool. The study sought to offer the best model for diabetic illness prediction based on effective process findings. Each method's experimental findings on the data were explored. With highest accuracy, the SVM was demonstrated to be the most efficient in forecasting malignancy. The article has demonstrated that SVM had the highest accuracy of any classifier, at 78.2 percent.

M. A. Sarwar et al. (2018) [22] shown and analysed many healthcare models and research. For the experiment, a dataset of a patient's medical record is acquired, and 6

major ML algorithms are applied to the data. The efficiency and correctness of the algorithms utilised are compared. A comparison of the various machine learning algorithms used in this study shows which model is most suited for diabetes forecasting. This study use machine learning approaches to aid physicians and practitioners in the early identification of diabetes. Finally, the author concluded that SVM with KNN achieved a highest accuracy of 77 percent.

R. Priyadarshini et al. (2014) [23] created multiple models and used the concept of customised advanced learning machines to categorise individuals as diabetic or non-diabetic based on past supplied data, enabling medical practitioners in deciding if someone is diabetic or not. It also discusses and analyses the use of two standard ML techniques, back propagation neural network and modified Extreme learning machine, as binary classifiers to the diabetes prediction issue. These two methods are applied on the same type of multi - class classification classification datasets, and the study attempts to draw some conclusions based on training and testing results. The data were taken from the University of California, Irvine's learning repository.

From the literature review we have found out that most used models in diabetes prediction are RF, SVM and DT. In all studies we found out that these models are performing very well in predicting diabetes. And most of the authors use PIMA Indians dataset for their study.

2.2 Insights From Literature Review

In this survey, Various factors have been listed and each factor has used a dataset. We have taken 23 studies for this survey. We found that different authors used different datasets which are Security Force Primary Health Care, PIMA Indians, Sylhet Diabetes Hospital San Antonio Heart Study (SAHS), Practice Fusion, Early Stage Diabetes Risk Prediction. And authors also used different models which are RF, SVM, Naive Bayes, Logistic Regression, DT, KNN, Boosting, MLP, LSTM, Neural Network and ANN. and all the research that we have taken is in the range of 2014-2021. All the details about datasets and techniques used are given in the Table 2.1 below.

Table 2.1. Datasets and Techniques

SN	Dataset	Techniques Used
1	Security Force Primary Health Care	RF and SVM
2	PIMA Indians	RF, NB, LR, SVM, DT, KNN and ANN
3	Sylhet Diabetes Hospital	RF, DT, KNN, SVM, LR, NN
4	San Antonio Heart Study (SAHS)	SVM
5	Practice Fusion, PIMA Indians	KNN, SVM, DT, RF, Boosting, NN, NB
6	Early stage diabetes risk prediction	NB, SVM, KNN, DT, RF, LR, Ada Boost

Below Table 2.2 country vs number of studies that are taken to conduct this survey is shown. Here we took most of the studies from India which are in majority after that the United Kingdom is in second place and from other countries we took just one research paper.

Table 2.2. Country Vs Number of Studies taken

SN	Country	Number of Studies
1	Saudi Arabia	1
2	Romania	1
3	Canada	1
4	India	14
5	Spain	1
6	UK	2
7	China	1
8	USA	1
9	Egypt	1

We may simply deduce from the below table 2.3 which tells us about the frequency of the classifiers. RF is the most favoured approach among researchers for predicting diabetes.

Table 2.3. Frequencies of Classifiers

SN	Techniques Names	Used in Papers	Frequency
1	RF	[1], [2], [3], [5], [6], [9], [10], [13], [14], [15], [16], [17], [18], [20], [21], [22]	16
2	SVM	[1], [2], [3], [4], [6], [8], [10], [11], [13], [16], [17], [18], [20], [21], [22]	15
3	Naive Bayes	[2], [4], [15], [18], [20], [21], [22]	7
4	Logistic Regression	[3], [9], [10], [13], [15], [17], [19], [22]	8
5	DT	[4], [7], [10], [13], [16], [17], [18], [19], [20], [21], [22]	11
6	KNN	[5], [10], [12], [15], [16], [18], [20], [22], [23]	9
7	Boosting	[10], [13], [16], [17], [18], [20]	6
8	MLP	[10]	1
9	LSTM	[10], [16]	2
10	Neural Network	[13], [18], [23]	3
11	ANN	[4], [10]	2

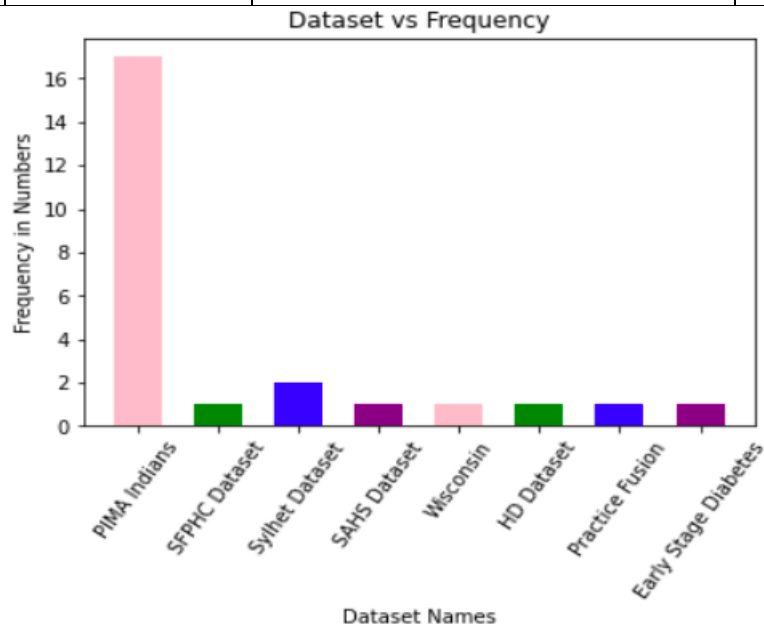


Figure 2.1 Dataset Vs Frequency

From the above image we can see that PIMA Indians Dataset is a widely used dataset for diabetes prediction.

Table 2.4. Maximum Accuracy vs Model Name

SN	Model Name	Maximum Accuracy
1	RF	98% [1]
2	SVM	100% [8]
3	Naive Bayes	82% [4]
4	Logistic Regression	97% [9]
5	DT	96.6% [10]
6	KNN	84.6% [10]
7	Boosting	100% [10]
8	MLP	88.5% [10]
9	LSTM	92.3% [10]
10	Neural Network	96.2% [13]
11	ANN	88.5% [10]

In above Table 2.4 the maximum accuracy achieved by each algorithm is given. Here we can see that SVM and Boosting technique is performing well for diabetes dataset.

This chapter provides an overview of ongoing research and focuses on recent advances in ML that have had a substantial influence on the detection and diagnosis of diabetes. This evaluation provides a thorough assessment of the techniques, approaches, characteristics, and shortcomings. Finally, this chapter focuses on the advantages and disadvantages seen in present research work within each category as a source of inspiration for future advancements in diabetes illness prediction.

CHAPTER 3

DATASET AND TECHNIQUES USED

This chapter tells about the dataset and ML techniques that we have used to achieve the proposed work. In the first section of this chapter we will discuss the dataset thoroughly and in the next section we will discuss all the ML techniques and optimization techniques that we have used in this research.

3.1 Dataset Used

This section deals with the brief description of the dataset that we have used in the proposed work.

In this study we have used ML classification models for predicting whether a patient has diabetes or not. We have taken the “UCI ML Repository’s Pima Indians Diabetes Dataset” [34] which consists of 768 rows and 9 attributes. Attributes are given below in the image. Where the first eight attributes are independent variables and the last one named ‘Outcome’ is Dependent variable. In Figure 3.1 we have shown the head of our data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 3.1 Head of our dataset

The diabetic patients classification is given in the figure below. We developed ML models for diabetes prediction. But the dataset was not balanced. Nearly 500 classes

are marked as 0 which stands for not diabetic and 268 marked as 1 stands for positive (diabetic).

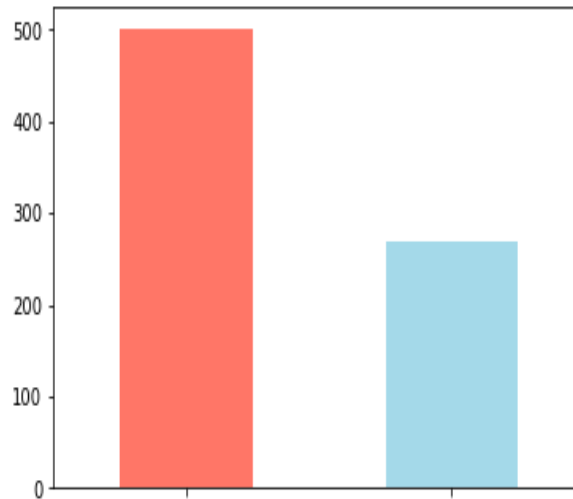


Figure 3.2 Comparison of “Outcome” attribute in dataset

3.1.1 Data Pre-processing

The most important step is data Pre-processing. Data Pre-processing is performed to increase the quality and efficacy of the mining results. When utilising ML Techniques on a dataset, this strategy is required for accurate findings and effective prediction. It entails eliminating values that are missing or inconsistent.

3.1.2 Data Normalisation

Following data pre-processing, we do max min normalisation. Min Max is a method of data normalisation that is comparable to Z score, decimal scaling and standard deviation normalisation. It helps with data normalisation. The data will be scaled from 0 to 1. This standardisation makes the data easier to understand. It's used to scale numbers. Following that, we divided the data into training and testing sets.

3.1.3 Data Balancing Using SMOTE

To manage the unbalanced classification in PIMA Indians data, we combine oversampling and under sampling utilising SMOTE. Following that, we fit our model to the balanced sampled data and analyse its performance on the testing set to compare the various models. After balancing the dataset with SMOTE, the dataset distribution is as shown below.

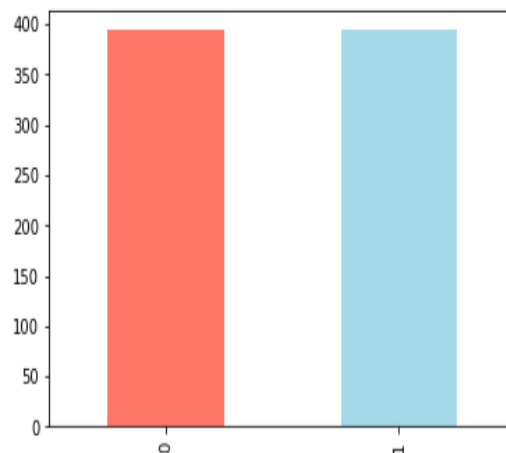


Figure 3.3 “Outcome” class in dataset after using SMOTE

```
print("Positive Values before Smote :",y_train.value_counts(normalize=True)[1]/(y_train.value_counts(normalize=True)[0]+
print("Negative Values before Smote :",y_train.value_counts(normalize=True)[0]/(y_train.value_counts(normalize=True)[0]+

print()
print('Shape of x before applying SMOTE :', X_train.shape)

smote = SMOTE()
X_train,y_train = smote.fit_resample(X_train,y_train)

print('Shape of x after applying SMOTE : ', X_train.shape)
print()

print("Positive Values after Smote :",y_train.value_counts(normalize=True)[1]/(y_train.value_counts(normalize=True)[0]+
print("Negative Values after Smote :",y_train.value_counts(normalize=True)[0]/(y_train.value_counts(normalize=True)[0]+

Positive Values before Smote : 35.83061889250814 %
Negative Values before Smote : 64.16938110749186 %

Shape of x before applying SMOTE : (614, 8)
Shape of x after applying SMOTE : (788, 8)

Positive Values after Smote : 50.0 %
Negative Values after Smote : 50.0 %
```

Figure 3.4 Smote applied on the dataset

Following is the snapshot of the SMOTE for balancing the data. Earlier 'Outcome' attribute was in ratio 35:65. Then we used SMOTE and changed it to a ratio of 50:50.

3.1.4 Hyperparameter Tuning

We can see hyper parameter tuning as a machine learning tool in which each model includes a set of dials that we may adjust to control how it performs. Changing these settings has the potential to improve or degrade model performance. This is known as hyperparameter tuning.

3.2 Insights From the Heatmap

We found there is the highest correlation between the outcome and glucose. This is obviously an overwhelmingly contributing factor in patients with diabetes. The BMI and skin thickness also have a strong correlation indicating that the two are closely connected in terms of body fat content. We know from basic biology that insulin regulates the usage of glucose in your body and the correlation between the 2 reflects that here as well.

In the below figure, the scatter plot is drawn for diabetes in function of BP and Age and Blood pressure. Where red dots determine the patient has diabetes and blue dots tell the patient has no diabetes. From the plot we can get insights that there is less probability of diabetes in age less than 32.



Figure 3.5 Heatmap of dataset

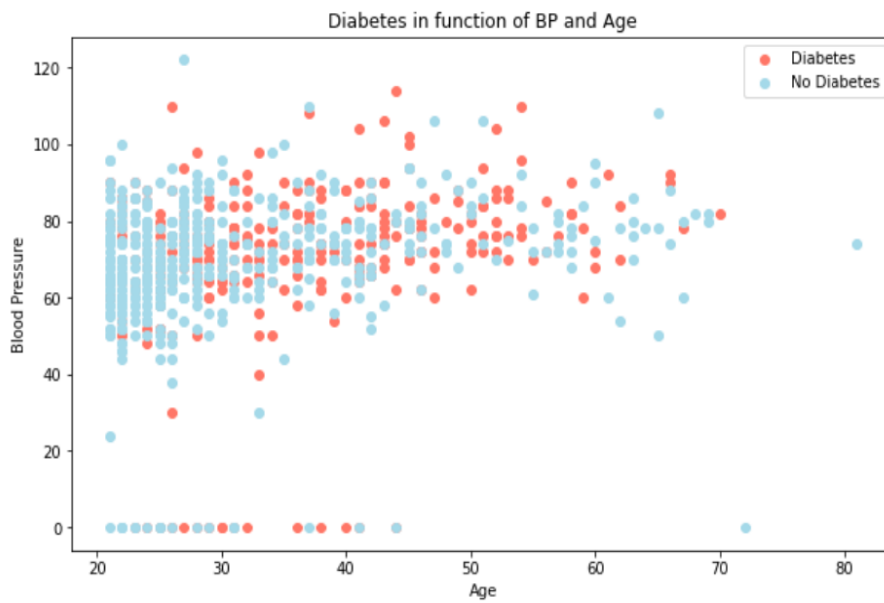


Figure 3.6 Blood pressure vs diabetes in function of BP and Age

3.3 Techniques Used

In this study, various ML classification techniques are used by us. All the algorithms used in this study are described below.

3.3.1 Logistic Regression

It is an algorithm that comes under supervised ML. It is used when we have dependent variable as categorical and we predict the probability of the target variable. Here the dependable variable “Outcome” is categorical, Therefore we used logistic regression. For example there are only two cases in which either the patient has diabetes (1) or patient has no diabetes (0). Here we use sigmoid function for model building. We can define the sigmoid function using the following expression [31].

$$Y = 1/(1 + e^{-x}) \quad (3.1)$$

Here Y is the outcome of the addition of weighted variables. It depends on output value, if it is greater than 0.5 then the answer is 1 otherwise 0.

3.3.2 K-nearest Neighbours

It is an algorithm that comes under supervised ML used to solve classification problems easily. Main ideology behind KNN is, It thinks that similar things exist in adjacent areas. That means they are near to each other. Basically we deal with distance in KNN, there are lots of ways to calculate distance like Manhattan, Euclidean, Hamming distance, Minkowski distance, Kullback-Leiber (KL) divergence, BM25 etc.

Algorithm

1. First we have to load the data, after the data loading We have to initialise k according to the chosen number of neighbours.
2. Then in the dataset for every sample we have to do the following steps :

- a. First we have to find the euclidean distance among the given sample and current sample from the data.
 - b. Then we have to add these distances, index of the sample to a collection in ordered form.
3. Then we have to sort the distances in ascending order for knn.
4. After that we chose the first K entries from the pool of distances.
5. Then we are left with the labels of the selected K entries.
6. At the end we have to return the mode of the K labels.

3.3.3 Decision Tree

It is a classification technique that is used for supervised ML. It is mainly used when we have categorical variables. DTs have a tree-like shape in which each inner node is for testing on the attribute, and each branch is for the result of the test, and each node which is in leaf holds a class name.

3.3.4 Random Forest

It is a ML classification algorithm which is based on many DT. Using it we can predict for which observation belongs to which class. In this technique we have to take a big number of DT that work as an ensemble. And each DT has its own decision, using all these DT we will take the highest frequency of decision as an answer of the RF [33]. It is the most powerful technique of ML because we used multiple decision trees here to train our model.

3.3.5 Naive Bayes

It is based on a probabilistic ML model which is used for classification tasks. The base of the classifier is based on the Bayes theorem.

$$P(X/Y) = P(Y/X) * P(X)/P(Y) \quad (3.2)$$

Here $P(X/Y)$ is the probability of A's happening if B has already happened. X is the hypothesis and Y is evidence. The assumption is that all the features are independent [36].

3.3.6 Support Vector Machine

It is also known as SVM, it is a system that separates the guarded machine. In this process we create a hyperplane in the N space that separates all the categories and clearly separates the data points. Basically hyperplanes are decision-making limits that help to separate data points. Data points on different sides of the hyperplane are divided into different classes. The points near the hyperplane separate the classes known as supporting vectors and affect the position and position of the hyperplane [32].

Algorithm

1. First select the hyper plane by calculating the length among the planes and the data.
2. If length among the classes is high then the probability of wrong assumption is low and vice-versa.
3. At the end select the class which has the most margin.

$$\text{Margin} = \text{positive point distance} + \text{negative point distance} \quad (3.3)$$

3.3.7 Gradient Boosting

In general it is Adaptive Boosting which is a famous ensemble method. It is an iterative process which generates a powerful classifier which has lots of weighted classifiers which are complemented for each other. All base learners trained on various subsets which are taken from the original dataset. The crux over this technique is that at each loop more importance is given on examples which were classified wrongly in the previous loop. The value of importance is countable by a weighted average that is given to all occasions in the training reproduction.

CHAPTER 4

EXPERIMENTAL SETUP

This section provides a brief summary of the hardware and software tools utilised in the proposed work. In this section, we will also discuss our performance measures, framework of this study and methodology.

4.1 Hardware Used

The suggested work is based on ML techniques, which revolves on classification, only a computer system is needed as a hardware tool for implementation. The model is implemented and executed on an LAPTOP with the minimum hardware parameters listed be

- System Type Windows 10, Macintosh
- Processor Core i3 processor
- RAM 4GB
- Hard disk 500GB

4.2 Software Used

All ML models were implemented using Anaconda Navigator and Jupyter Notebook.

Python 3.6 Python is a general-purpose programming language that is interpreted and object-oriented. It is an open source language that has become one of the most popular due to its brief, simple, and large library support. It offers excellent code readability and a simple syntax. Python has many packages for diverse purposes; some of the packages we utilise in our work are numpy, pandas, scikit learn, seaborn, matplotlib etc.

4.3 Framework of Our Work

Above is the Framework that we have used in our project. First of all we collected our data online then preprocessed it and normalised it and then balanced our dataset. After that we did a training and testing split, then we performed ML techniques on it, and at the end we compared the result and each technique on the basis of accuracy.

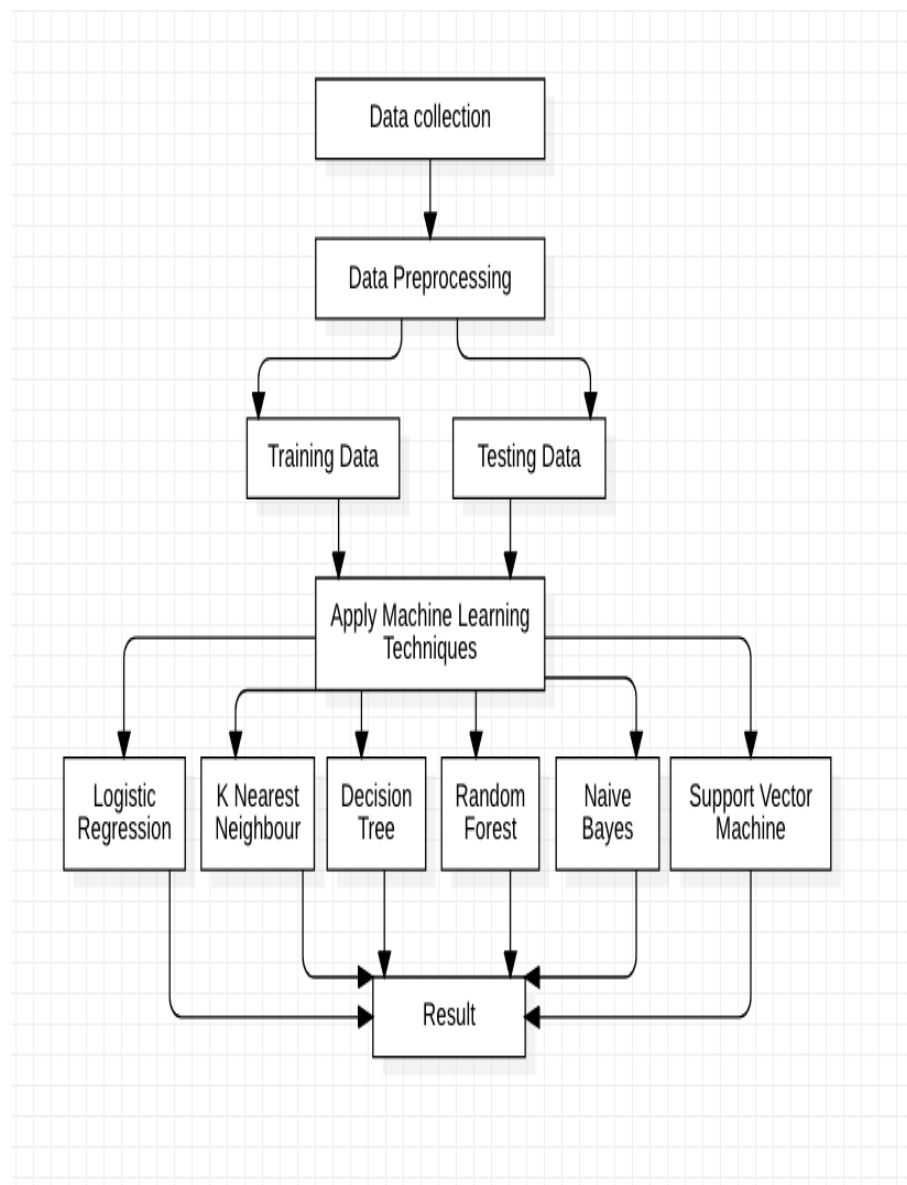


Figure 4.1. Framework

4.4 Methodology Used

Here, we will discuss the methodology used to implement the proposed work. We have a dataset in the form of Comma Separated Value of PIMA Indians datasets which is taken from UCI repository. These are the steps that we have taken to implement the diabetes prediction using ML.

1. First of all we did a survey on the previous paper and found out which are the best ML techniques that have been used for diabetes prediction in previous studies and what are the popular datasets.
2. After that we listed down all popular techniques and datasets for our further study.
3. Then we imported all the necessary libraries such as numpy pandas and other ML libraries that are used in this study.
4. After that we uploaded the PIMA Indians dataset..
5. Then we did pre-processing on the dataset to handle missing data and to remove anomalies.
6. After that we tried to understand the data and patterns and did data analysis.
7. Then the dataset is not balanced so to balance it we used Synthetic Minority Over-sampling Technique (SMOTE). In which we create an artificial minority class for the unbalanced dataset to make it balanced.
8. After that we splitted data in a ratio of 80:20 respective Training set and Test set data.
9. Then we tried to train our model using training dataset with various ML techniques.
10. After that first we checked accuracy in the original dataset by using all the ML techniques and then we checked for the modified data the same.
11. After that we did hyperparameter tuning in our model and tried to increase the accuracy. Only for KNN, RF and XGB, because we obtained high accuracy in it.
12. After that we compared the accuracy of all the algorithms.
13. At last we compared the result for RF and XGB with hyper parameter tuning.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter defines the arrangement applied during the experiments. The major and unique elements of the arrangement were discussed as part of the dissertation work progress. Here we evaluated the performance of the suggested ML techniques. The results are contrasted, as is the examination of the without and after smote models. In the full explanation of the results, the features, benefits, and shortcomings of the proposed model and diabetes prediction are examined. The following are the findings from our qualitative and quantitative method analyses:

In table 5.1, compares maximum accuracy before smote without hyper parameter tuning to maximum accuracy before optimization. We can plainly observe that after performing hyper parameter tuning to all ML algorithms, their accuracy rose.

Table 5.1 Comparison table for accuracy before smote without/with optimization

SN	ML Technique Used	Accuracy before smote optimization	Accuracy before with optimization
1	KNN	72.07	77.94
2	Logistic Regression	76.62	84.05
3	DT	66.88	84.05
4	RF	74.02	88.40
5	Naive Bayes	74.02	85.50
6	SVM	77.92	84.05

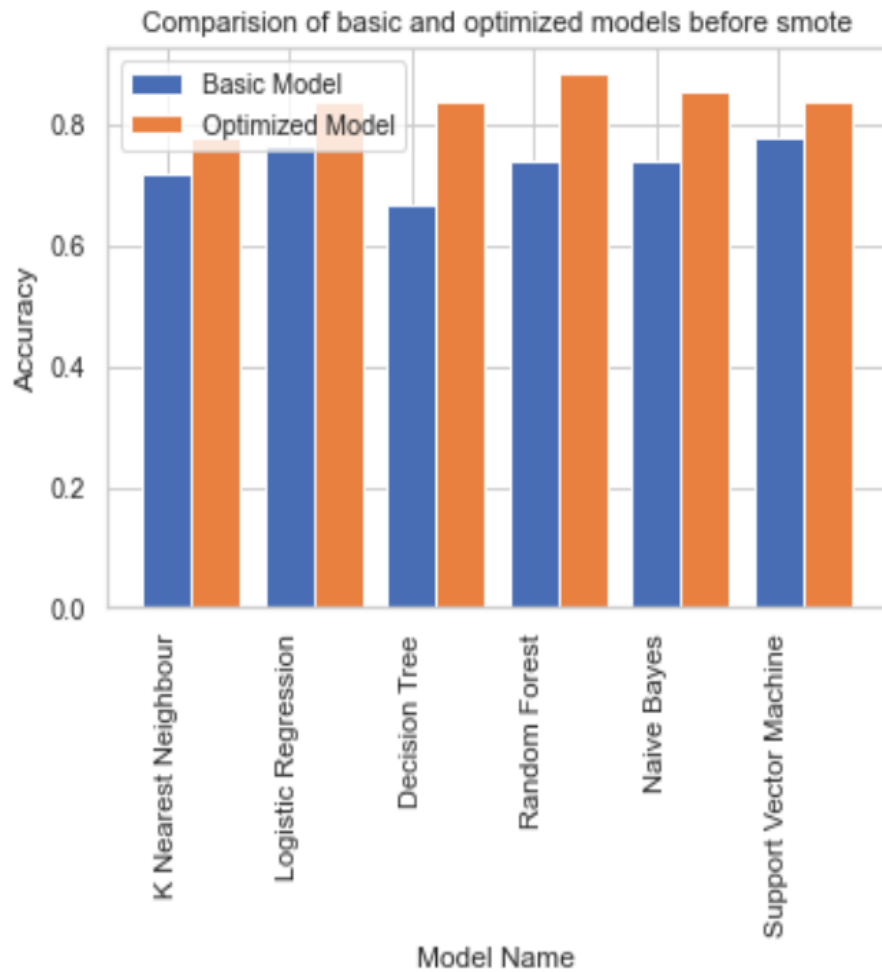


Figure 5.1 Comparison of accuracy before smote for normal and optimised ML algorithm

In figure 5.1, we created grouped bar chart for clear visualisation of accuracy of all the techniques and it is clearly shown here that for RF we achieved the highest accuracy of 88.40 percent. Which we achieved after optimization of our model before smote.

Here in this method we encountered a challenge that our dataset was not properly distributed and also it is not properly balanced. So to make it a balanced dataset we used the smote technique and balanced our dataset on the basis of ‘Outcome’ attribute in a ratio of 50 is to 50, which was earlier nearly to 65:35.

Table 5.2 Comparison table of accuracy after smote without/with optimization

SN	ML Technique Used	Accuracy after smote without optimization	Accuracy after smote with optimization
1	KNN	77.94	89.65
2	Logistic Regression	84.05	81.81
3	DT	84.05	89.65
4	RF	88.40	88.50
5	Naive Bayes	85.50	80.68
6	SVM	84.05	85.22

After balancing our dataset we again compared the results of our models after smote with optimization and without optimization. Table 5.2 compares highest accuracy after smote without hyper parameter tuning to before and after optimization. Except for logistic regression and naïve bayes, we can plainly observe that accuracy rose following optimization for all ML techniques.

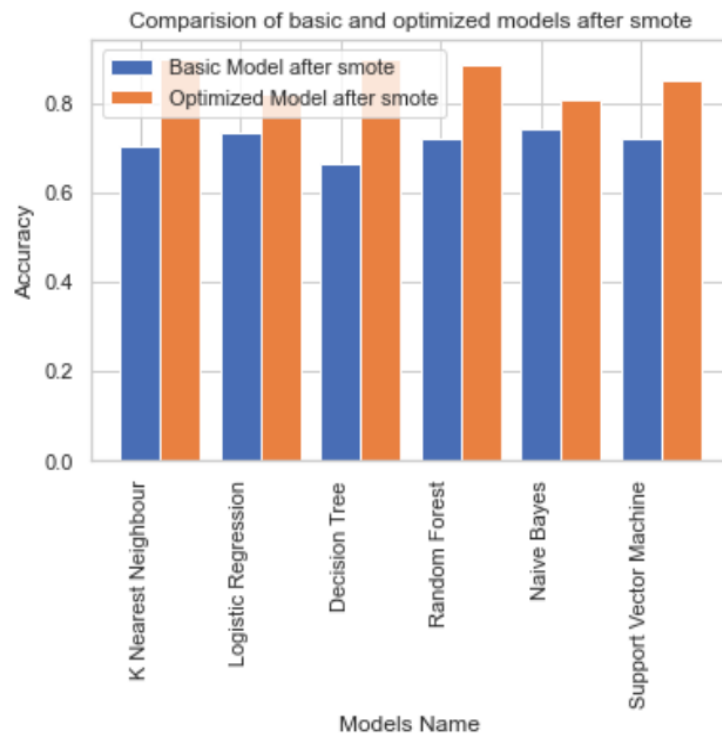


Figure 5.2 Comparison of accuracy after smote for normal and optimised ML algorithm

In figure 5.2, we constructed a grouped bar chart to easily visualise the accuracy of all approaches, and it is apparent that for KNN and DT, we reached the maximum accuracy of 89.65%. Which we accomplished after optimising our model and balancing our dataset with smote. We also noted that after employing smote, the accuracy of our models rose, and we got the maximum accuracy using the RF method of 88.40 percent, which was before 74.02 without optimization.

After performing optimization in both "dataset before smote" and "dataset after smote," we compared the results of only the optimised part, which was the crux of this research. Experimentally, we discovered that the algorithms on balanced dataset worked exceptionally well, outperforming previous studies in this field using ML, and we also achieved maximum accuracy using KNN after using smote and optimising our model.

Table 5.3 Comparison table for datasets before and after optimization and smote

SN	ML Technique Used	Performance before smote with optimization	Performance after optimization smote
1	KNN	77.94	89.65
2	Logistic Regression	84.05	81.81
3	DT	84.05	89.65
4	RF	88.40	88.50
5	Naive Bayes	85.50	80.68
6	SVM	84.05	85.22

Above in table 5.3 we can clearly see the comparison among all techniques with optimised models before and after smote. and for the same we have also given visualisation by grouped bar graph in figure 5.3.

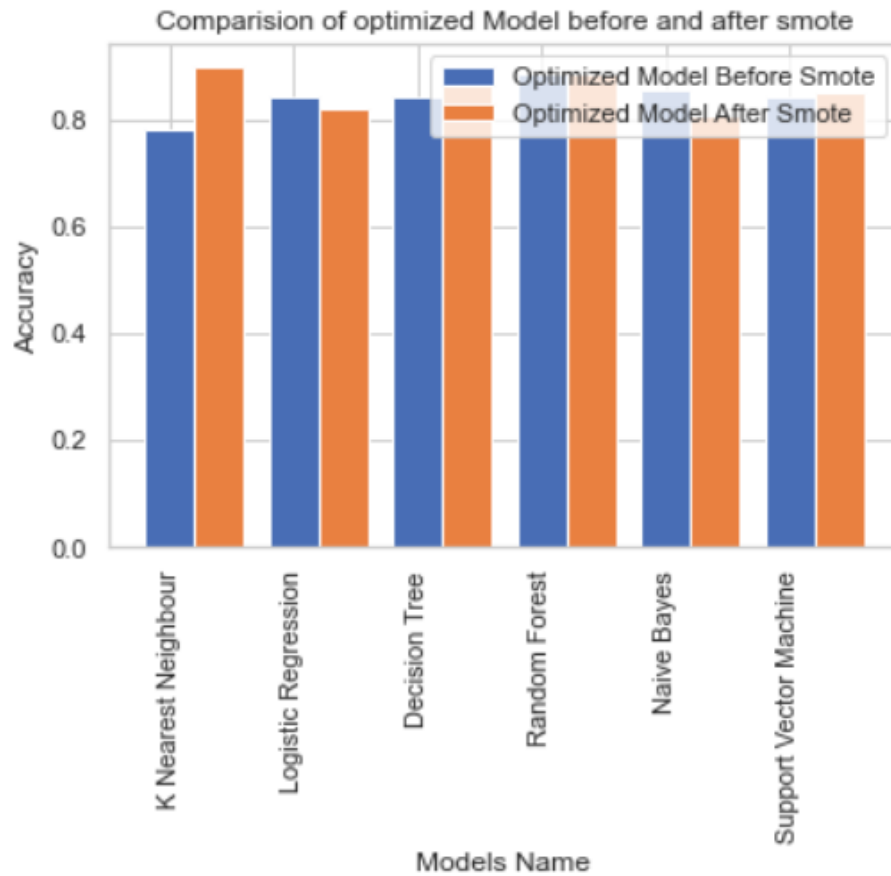


Figure 5.3 Comparison of accuracy among optimised ML algorithm before smote and after smote

After all these optimizations we also tried hyper parameter optimization in well performed techniques which are KNN and RF. We tried to tune our parameters using different techniques.

In KNN we change the value of k and try to check on which value of k we are achieving the higher value of accuracy. So as a result we found that at $k = 4$ we have found the maximum value of accuracy which is near to 90% for our test data. and $k=4$ comes out as a hyper parameter for our model, and the mean KNN score for our test data and training data is 85.93% and 87.08% respectively. Which is shown in below figure 5.4.

Mean KNN score on the training data: 87.08%
Mean KNN score on the testing data: 85.93%

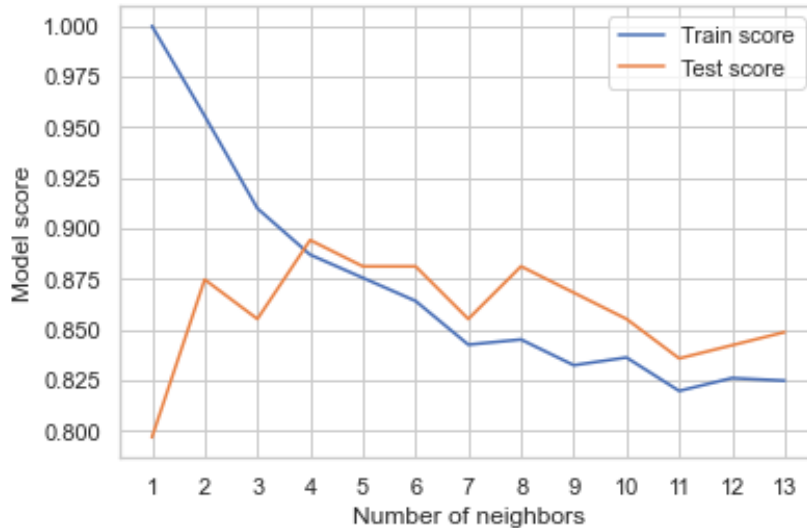


Figure 5.4 Comparison of training and testing score for KNN

After KNN we tuned our RF model using GridsearchCV optimization, and as a result we found the hyper parameter as 'max_depth': 8, 'max_features': 7, 'min_samples_split': 2, 'n_estimators': 500. Applying all these parameters we increased our accuracy for RF nearly 89.21% which was earlier 88.5 percent.

Then we used XGB Classifier and tuned it as well using GridsearchCV optimization and after Fitting 10 folds for each of 720 candidates, totalling 7200 fits we have found these hyper parameters {'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 0.1, 'n_estimators': 100, 'subsample': 1.0} and using these hyper parameters we achieved accuracy of 90.13 percent. Which is the best accuracy among all the techniques that we have used.

In figure 5.5 we can clearly see that the comparison between RF and XGB classifier after hyper parameter tuning as a result we can say that among all the ML models XGB boost after hyper parameter tuning is performed well

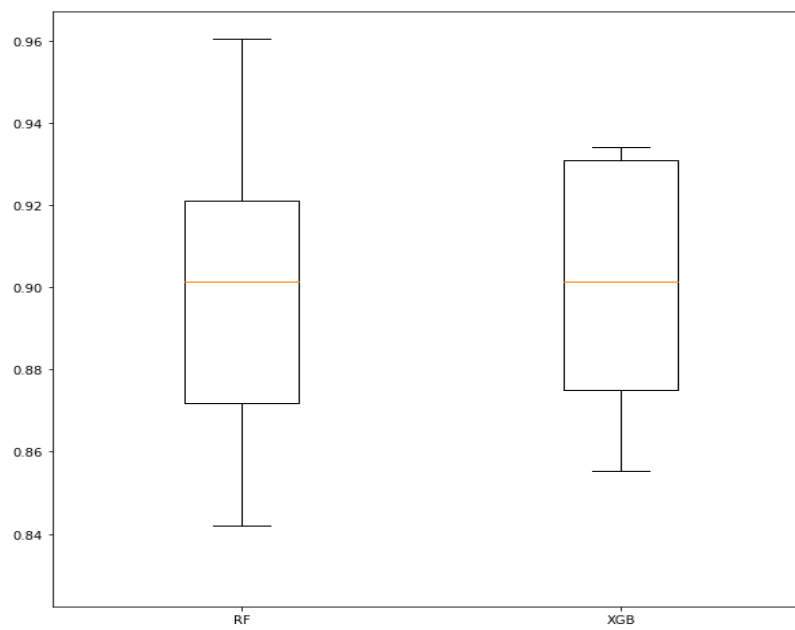


Figure 5.5 Comparison of RF and XGB after parameter tuning

In this section, we have clearly seen all the comparisons and our results that we have found from our study.

CHAPTER 6

CONCLUSION AND FUTURE WORK

After concluding results and discussion over the proposed approach this part manages summary, final conclusion and potential future work in this topic.

6.1 Summary

Diabetes identification early and actively is one of the most significant health concerns confronting the entire planet. The idea of this research is to develop an appropriate prediction model for diabetes prediction that is based on a ML method. As we discussed, diabetes is a chronic disease and we should consult a doctor in time whenever we see symptoms of diabetes in our body. In this paper we implemented six classification techniques which are logistic regression, DT, RF, SVM, KNN and Naive Bayes to predict diabetes on two datasets. Out of one is modified using SMOTE and another one is original. And then we compared the result of all classification techniques together, We also used bagging technique. also we found out that RF with ensemble bagging technique gives the highest accuracy of 81.74%. and there is no improvement in model after using SMOTE except for the bagging classifier in which achieved the highest accuracy. In future research I would like to say that we can also use ANN and other deep learning techniques to predict diabetes.

6.2 Threats to Validity

All models identified by step by step literature search, and we take models which are available in the verification research. Other powerful things include the size of the research sample, possible composition, confirmation of the occurrence of diabetes, and more data on individual property. However, some drawbacks of our research need to be

mentioned. Furthermore, our data had some limitations regarding the availability of variables, so we made an effort to provide all the variables and meanings used as closely as possible. To address the inconsistencies, we used SMOTE, which gave similar results. Next, we use glucose concentration that does not discard itself. We cannot say this strikes our conclusions. We know that glucose is an important predictor of diabetes. Using data from only verified cases we were able to avoid false positives in the remaining group as diabetes will be unchanged for months to years.

6.3 Validation of Prediction Techniques

Retrieval models returned vary widely in the type and number of predictions, age range, model type, time of follow-up, and result of outcome. Three recent step by step reviews represents a comprehensive overview of the research that improved these models or validated specific models.

All reviews also showed that most of the models have never been verified by outsiders. Our research has now examined the effectiveness of highly developed techniques for predicting future diabetes in outsiders and shows that many normal techniques do good to identify those at more risk for diabetes and that extended techniques perform slightly better. In general, the performance of a guessing technique decreases when it is used in a validated database.

6.4 Conclusion

Regardless of the absence of ample proof in the research literature, most of the dataset attributes may not have a clear correlation. It must be recognised early to enable effective cure, and ML and Deep Learning have changed research into prognosis for early phase diabetes. To ensure effective treatment, diabetes must be identified initially, and ML and deep learning have turned research into risk prediction for early phase mellitus. Researchers are enthusiastic about experimenting with various classification methods and introducing new models to improve accuracy of diabetes prediction. With the same vision the frequency of usage of ML techniques and accuracy of all ML and deep learning

classifiers used in the past years were examined in this research. The accuracy of these ML approaches was between 82 to 100 percent. The maximum accuracy obtained for the Deep Learning algorithms was 96.2%. In the future, the non-used classifiers might be used in other datasets in a combined model to improve the accuracy of diabetes mellitus prediction. The study's findings might aid health-care practitioners in diagnosing diabetes early and making better clinical decisions regarding diabetes treatment, perhaps saving lives. Despite the fact diabetes can be predicted with great accuracy.

6.5 Future Scope

The proposed approach, together with the recommended ML classification algorithms, may be useful in the prediction or diagnosis of various diseases in the near future. The study work may be changed and enhanced for diabetes prediction analysis, including a few more ML techniques. As future work, the missing data is totally learnt using meta-heuristic methods. For future work, the algorithms are programmed to be capable of learning missing data prediction. Furthermore, the study may be expanded for diabetes prediction by gathering data from various locations across the globe and producing a more precise and conventional, discriminating framework. The work can be changed and enhanced in order to automate the diabetes analysis.

Following that, we will focus on identifying other parameters in the dataset that might predict diabetes in an earlier stage, and we will also try the deep learning techniques in this field so that we can detect diabetes in the early phase with higher accuracy.

REFERENCES

- [1] A. S. Alanazi and M. A. Mezher, "Using machine learning algorithms for prediction of diabetes mellitus," 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020.
- [2] N. E. Costea, E. V. Moisi, and D. E. Popescu, "Comparison of machine learning algorithms for prediction of diabetes," 2021 16th International Conference on Engineering of Modern Electric Systems (EMES), 2021.
- [3] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018.
- [4] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019.
- [5] S. K. Reddy, T. Krishnaveni, G. Nikitha, and E. Vijaykanth, "Diabetes prediction using different machine learning algorithms," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021.
- [6] N. Mohan and V. Jain, "Performance analysis of Support Vector Machine in diabetes prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020.
- [7] A. M. Posonia, S. Vigneshwari, and D. J. Rani, "Machine learning based diabetes prediction using decision tree J48," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020.
- [8] G. A. Pethunachiyar, "Classification of diabetes patients using kernel based support vector machines," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020.
- [9] V. Mounika, D. S. Neeli, G. S. Sree, P. Mourya, and M. A. Babu, "Prediction of type-2 diabetes using machine learning algorithms," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021.
- [10] M. A. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A comparative analysis of early stage diabetes prediction using machine learning and Deep Learning Approach," 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021.

- [11] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani, and K. Qaraqe, "Predicting diabetes in healthy population through machine learning," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019.
- [12] K. Driss, W. Boulila, A. Batool, and J. Ahmad, "A novel approach for classifying diabetes' patients based on imputation and machine learning," 2020 International Conference on UK-China Emerging Technologies (UCET), 2020.
- [13] J. Ma, "Machine learning in predicting diabetes in the early stage," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2020.
- [14] K. VijayaKumar, B. Lavanya, I. Nirmala, and S. S. Caroline, "Random Forest algorithm for the prediction of diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019.
- [15] L. V. R. Kumari, P. Shreya, M. Begum, T. P. Krishna, and M. Prathibha, "Machine learning based diabetes detection," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021.
- [16] S. Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote, and P. Topale, "Diabetes prediction using feature extraction and Machine Learning Models," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021.
- [17] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018.
- [18] R. Akula, N. Nguyen, and I. Garibay, "Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes," 2019 SoutheastCon, 2019.
- [19] B. Sridhara MURTHY and J. SRILATHA, "Comparative analysis on diabetes dataset using machine learning algorithms," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021.
- [20] M. Rady, K. Moussa, M. Mostafa, A. Elbasry, Z. Ezzat, and W. Medhat, "Diabetes prediction using Machine Learning: A Comparative Study," 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2021.
- [21] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

- [22] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in Healthcare," 2018 24th International Conference on Automation and Computing (ICAC), 2018.
- [23] R. Priyadarshini, N. Dash, and R. Mishra, "A novel approach to predict diabetes mellitus using modified extreme learning machine," 2014 International Conference on Electronics and Communication Systems (ICECS), 2014.
- [24] "What is diabetes" National Institute of Diabetes and Digestive & Kidney Diseases. [Online]. Available: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>. [Accessed: 1-Feb-2022].
- [25] "Home," International Diabetes Federation (IDF) - Home. [Online]. Available: <https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html>. [Accessed: 1-Feb-2022].
- [26] N. E. Costea, E. V. Moisi, and D. E. Popescu, "Comparison of machine learning algorithms for prediction of diabetes," 2021 16th International Conference on Engineering of Modern Electric Systems (EMES), 2021.
- [27] S. K. Reddy, T. Krishnaveni, G. Nikitha, and E. Vijaykanth, "Diabetes prediction using different machine learning algorithms," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021.
- [28] S. Perveen, M. Shahbaz, T. Saba, K. Keshavjee, A. Rehman, and A. Guergachi, "Handling irregularly sampled longitudinal data and prognostic modeling of Diabetes Using Machine Learning Technique," IEEE Access, vol. 8, pp. 21875–21885, 2020.
- [29] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018.
- [30] Neha Prerna Tiggaa, and Shruti Garga, "Prediction of type2 diabetes using machine learning classification methods," Procedia Computer Science. 167:706716, 2020.
- [31] L. Lei, "Prediction of score of diabetes progression index based on logistic regression algorithm," 2020 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 2020.
- [32] N. Mohan and V. Jain, "Performance analysis of Support Vector Machine in diabetes prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020.

[33] "Healthline : Diabetes in India" 2021. [Online]. Available : <https://www.healthline.com/health/diabetes/diabetes-in-india>. [Accessed: 15-Nov-2021].

[34] "Pima Indians Diabetes Dataset" 2018. [Online]. Available: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>. [Accessed: 15-Nov-2021].

[35] Yang Guo, Guohua Bai and Yan Hu, "Using Bayes Network for Prediction of Type-2 diabetes," 2012 International Conference for Internet Technology and Secured Transactions, 2012, pp. 471-472.

List of Publications

[1] Tarun Kharkwal, Shweta Meena, “Diabetes Prediction using Machine Learning”. Accepted at the **International Conference on Big Data, Machine Learning and Iot Applications (ICBMLIA)**. Further selected for publishing in ESCI (Wos) Journal.

Indexed by Web of Science, Scopus.

Journal Name : INT-JECSE

ISSN : 1308-558

Paper-Id : ICBMLIA19

Abstract- Diabetes is a disease which occurs when the glucose level increases in blood. It is a persistent disease that occurs mainly in two ways: First one is, if the adequate insulin is not produced by the pancreas and the second one, if insulin is not used by the body effectively. Insulin Hormone is responsible for regulating Blood sugar. Diabetes can harm our body parts too like eyes, kidneys, nerve, heart and blood vessels. Therefore Predicting diabetes in the primitive phase is very essential to control the diabetes and to save lives. In this research we will study various ML algorithms to predict diabetes. The different algorithms that we have used in this research are Logistic Regression, K-Nearest Neighbour, Decision Tree, Naive Bayes, Random Forest, Decision Tree and SVM classifier. After these techniques we did hyper-parameter tuning. And then we used k fold cross validation and bagging classifier. and also we compare accuracy of each model using SMOTE and without using SMOTE. We balanced our dataset using SMOTE and then we achieved accuracy of 81.74% using the bagging technique with Random Forest classifier as base model which was earlier 77.82%.

[2] Tarun Kharkwal, Shweta Meena, "Prediction Markets using Machine Learning". Accepted at the **International Conference on Advances in Data Science and Computing Technologies (ADSC-2022)**. Further Selected for publishing in Springer Scopus Indexed LNEE series.

Indexed by Scopus.

Paper-ID: ADSC83

Abstract- The introduction of a brand-new product line predicting future stock prices is an important part of financial decision-making and investment since stock values fluctuate often. Although the value of the share may reduce due to market movements, there is still a danger of losing money. The stock price and trade volume are affected by these swings, making the forecast even more difficult. There are a wide range of methods and techniques that may be used to anticipate the stock market's behaviour, and these methods and techniques can help investors respond more quickly and accurately to know when to purchase or sell the stock therefore a tremendous diversity of strategies have been created. Even though a number of strategies have been developed, none of them reliably anticipate stock prices. Stock price prediction concerns are being solved via data mining and evolutionary strategies. In data mining, the extraction of a large amount of information from a big database is called "mining". Data mining techniques are aimed to assist investors in uncovering hidden patterns from the historical data that includes plausible forecasting capability in the stock market because of the enormous volume of data. Stock market robots have been created by combining predictive analytics with data mining. Prediction models are built using historical data, which helps investors find patterns in the data and anticipate future returns. The evolutionary algorithm, on the other hand, is critical in properly projecting stock values. Evolutionary strategies have been found to outperform other parametric approaches in a number of studies. Evolutionary approaches may be used to improve more formal procedures since they are simple to apply and comprehend. They also do not suffer from the negative consequences of dimensionality.