PREDICTING DIABETES MELLITUS USING MACHINE LEARNING BASED ENSEMBLE MODEL

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Mamta

(801632023)

Under the supervision of:

Tarunpreet Bhatia

Lecturer

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY

PATIALA – 147004

June 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Predicting Diabetes Mellitus using Machine Learning based Ensemble Model*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Ms. Tarunpreet Bhatia* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Mamta

(Mamta)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Bhatia

(Tarunpreet Bhatia)

Lecturer, CSED

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisor **Ms. Tarunpreet Bhatia,** Lecturer, Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, Patiala. She has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of her. I also thank my supervisor for her time, patience, discussion and valuable comments. I am truly grateful to her for extending his total co-operation and understanding whenever I needed help and guidance for her. I am also heartily thankful to **Dr. Maninder Singh,** Head, Computer Science and Engineering Department and **Dr. Ashutosh Mishra,** PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report. I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia,** Dean Academic Affairs, for making provisions of infrastructure such a library facilities, computers labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection which made my stay at Thapar Institute of Engineering & Technology memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Mamta

ABSTRACT

Diabetes is the fast growing disease among the people even among the youngsters. Diabetes is caused by the increase level of the sugar in the blood. Diabetes is a serious health problem that needs special attention and public health interventions in the 21st century. Diabetes Mellitus is defined as a chronic condition, a disorder of metabolism characterized by increased concentration of blood sugar levels caused by either insufficient secretion of insulin or resistance to insulin action or a combination of both. Diabetes is also the creator of another kind of disease mainly, the eyes, kidneys, blood vessels, nerves and the heart. Machine learning techniques are used in medical predictions. Machine learning allows building models to quickly analyze data and deliver results, leveraging both historical and real-time data. With machine learning, healthcare service providers can make better decisions on patient's diagnoses and treatment options, which leads to the overall improvement of healthcare services. In this work, we have applied different machine learning models such as Decision Tree, Naive Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbors, Adaboost, Linear Model and Neural Network. From these models, performance are evaluated on the basis of sensitivity, specificity, precision, recall, and accuracy. After that top five models are selected to perform ensembling. A Boosting based ensemble model is predicting to improve the accuracy of the dataset. The proposed ensemble model gives better results as compared to the single model in terms of accuracy. The accuracy is improved up to 84.82%. 10-fold cross validation is applied to improve the robustness of the data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

ANN	Artificial Neural Networks
BMI	Body Mass Index
BP	Blood Pressure
CVA	Cardiovascular diseases
DT	Decision Tree
DM	Diabetes Mellitus
FPR	False Positive Rate
GDM	Gestational Diabetes Mellitus
K-NN	K-Nearest Neighbor
LMIC	Low-To-Middle Income Countries
LR	Logistics Regression
LM	Linear Model
ML	Machine Learning
NB	Naïve Bayes
NCD	Non-Communicable Diseases
NN	Neural Network
RF	Random Forest
SVM	Support Vector Machines
TID	Type 1 diabetes
T2D	Type 2 diabetes
TPR	True Positive Rate
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

Diabetes is one of the major chronic non communicable diseases that affect millions globally. Evidence shows that diabetes has become a major epidemic in newly industrialized and developing nations. With a diabetic prevalence of 8.6 %, India is ranked at the top with respect to the burden of diabetes. Diabetes Mellitus (DM) is defined as a chronic condition, a disorder of metabolism characterized by increased concentration of blood sugar levels caused by either insufficient secretion of insulin or resistance to insulin action or a combination of both. Chronic hyperglycemia results in long-term damage of various organs, mainly, the eyes, kidneys, blood vessels, nerves and the heart. Nowadays, diabetes has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. Hence, diagnosing the diabetes is very essential to save the human life from diabetes. Moreover, in 2014, diabetes global prevalence was estimated to be 9% among adults over the age of 18. In 2030 the diabetes will be the 7 prominent cause of death indicated by the World Health Organization (WHO). Due to low income, more than 80% of death occur due to diabetes. In the year of 2017, about 8% of women suffers with diabetes. In India, moreover 30 million peoples have been identified with diabetes. The Crude occurrence rate in the city areas of India is supposed to be 9% [1]. In rural areas, the occurrence is around 3 % of the aggregate population. Now a days, the population of India is more than 1000 million. India is sincerely facing a healthcare issues. The consequences for the Indian healthcare system are vast. Overall adult population in India is 829,491 and total 72,946.4 cases of diabetes in adults. India has achieved not much progress in the health status since gaining independence. There has been significant reduction in infant mortality rate and crude death rates and life expectancy has doubled. The country has seen major transitions in nutritional status, fertility and mortality rates, economic growth, etc. This in turn has affected the health profile of our country. Despite being one of the fastest growing economies of the world, the Indian health indicators are one of the worst among developing countries. Diabetes is a condition in which the blood glucose levels in the body increase. While this is the symptom, the reason behind this is that the glucose in the blood isn't getting transported into the cells.

1.1 Types of Diabetes

1. **Type 1 diabetes:** Juvenile-onset type of diabetes due to cellular mediated β-cell destruction resulting in insulin deficiency. It accounts for about 5-10% of the diabetics.
2. **Type 2 diabetes:** Adult-onset type diabetes due to insulin resistance predominantly with relative insulin deficiency. It accounts for about 90-95% of those with diabetes.
3. **Other types:** This includes those due to genetic defects of β-cell dysfunction, genetic defects of insulin action, disease of exocrine pancreas, endocrinopathies, infections and others.
4. **Gestational Diabetes Mellitus (GDM):** Any degree of glucose intolerance that has its onset or is recognized during pregnancy. Prevalence varies from 1-14% depending on the population studied. Glucose tolerance in women with GDM usually becomes normal post-delivery. However, women who develop gestational diabetes have a higher risk of developing T2DM (Type 2 Diabetes Mellitus) later in life. GDM is more common among certain populations with a higher risk of T2DM.

1.2 Non-Communicable Disease

A non-communicable disease is defined as a medical condition that is neither transmissible nor infectious. They usually have a long duration and a slow progression. It include heart and blood vessel diseases, cancers, asthma, lungs cancer, lung diseases and diabetes. Non-communicable disease are the leading cause of death globally and contribute significantly to the global burden of diseases. In India, 61% of deaths occur from non-communicable disease and 23% risk of premature death from target non-communicable disease. Non-communicable diseases are caused by a combination of modifiable and non-modifiable risk factors. Behavioral risk (modifiable) factors contribute to the increased deaths caused by non-communicable diseases around the world. A large number of non-communicable diseases can be prevented by taking measures against the following four behavioral risks, namely, tobacco use, alcohol abuse, unhealthy diets and decreased physical activity. According to WHO estimates, the prevalence of daily tobacco smoking in India was 25.1% for males and 2% for females. Also, there was a 14% prevalence of physical inactivity; 10.8% in males and 17.3 % in females. Non-communicable diseases cause more deaths than all others causes combined together. Non-communicable diseases are projected to become the most

common cause of death by 2030 surpassing all other causes of death (communicable diseases, maternal, perinatal and nutritional) [2].

Non-communicable diseases are also often associated with older age groups, but 16 million of all deaths were caused by non-communicable diseases before 70 years of age and 82% of the 16 million deaths occurred in low-to-middle income countries (LMIC). Non-communicable diseases also account for about 48% of disability adjusted life years globally. According to WHO estimates, non-communicable diseases will cause about 52 million deaths by 2030 and will also be responsible for three times as many disability adjusted life years in the low-to-middle income countries. Demographic transition and heightened risk for acquiring non-communicable diseases in India have also pushed the out of pocket expenditure for non-communicable diseases from 31.6% to 47.3%. Major portion of the expenditure was used for diagnostic tests for non-communicable diseases, medical equipment's and purchasing medications. It was found that 25% of the households with a member with Cardiovascular diseases and 50% of the families with a member with cancer experienced catastrophic expenditures leading almost 10% and 25% respectively, of the families below poverty line due to healthcare expenditures. Prevention and control of non-communicable diseases require different government sectors to co-operate and find a comprehensive approach to reduce the risks associated with non-communicable diseases. In India, a large low-middle income country, a non-communicable disease have become one of the major contributors to increased burden of diseases, morbidity and mortality. Urbanization in India has increased post-independence. An increase in both urban population and migration to the urban areas coupled with increased prevalence of non-communicable diseases show that there is a possible relationship between urban city and non-communicable diseases risk factors. Exposure to an urbanized environment results in lifestyle modifications such as increased smoking and alcohol consumption, low physical activity due to better transportation systems, modernized and mechanized lifestyle, nutrition imbalance, stress etc. which in turn are significant risk factors in the development of non-communicable diseases. A non-communicable diseases were responsible for 61% of the all causes of deaths in 2017, cardiovascular diseases accounting for 26%; cancers for 7%; diabetes for 3% and respiratory diseases for 13%; and other non-communicable diseases for 12% as shown in Figure 1.

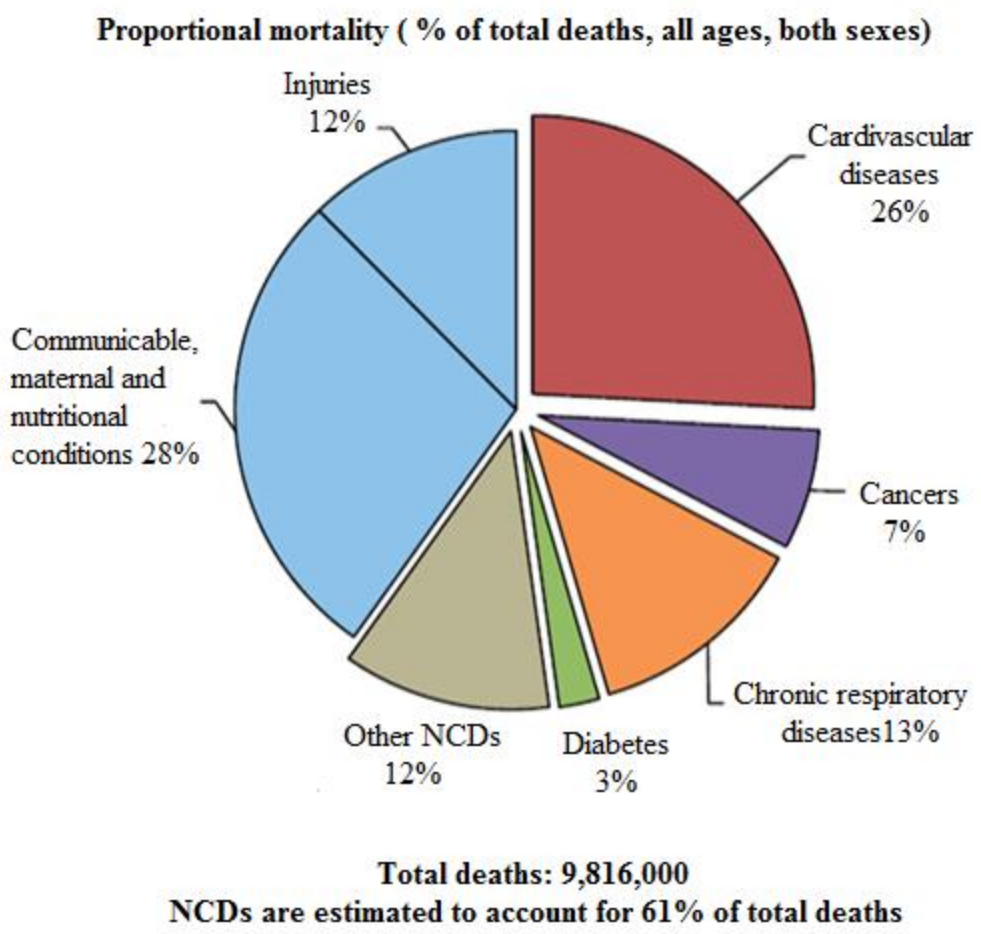

Fig. 1.1 Causes of Death in India [1]

Non-communicable diseases can further complicate issues for the low/middle-income groups. Unmanaged blood glucose levels, obesity, and physical inactivity or leading a sedentary lifestyle are common in today's population. Such "intermediate risk factors" as described by the WHO can be identified by even a basic health center or clinic. It is common knowledge to the literate world that stopping or reducing the dependency on alcohol or tobacco, limiting greasy foods that are high in fats and reducing sugar and salt intake can aid people to reduce the risks associated with non-communicable diseases and cardiovascular diseases in particular. When a potential threat is detected, people must immediately seek medical help and reschedule their lifestyles. Moreover, people should use appropriate medication to control or limit the damage caused by these factors.

1.3 Diabetes Risk Factors

1. **Demographics:** Many studies showed that risk of diabetes increases as the person gets older especially after 45 years of age. According to Diabetes Association, the risk of diabetes in Indians is very high because these populations are more like to have high BP and high BMI. In 2017, Centre for Disease Control survey estimated 86 million pre diabetes cases among population of 20 years or older.
2. **Blood Pressure:** Hypertension is one of the major factors that can worsen the complications of diabetes. Most people with diabetes are diagnosed to have high blood pressure.
3. **Body Measures:** Many researchers found that risk of diabetes increases with the increase in body mass index, overweight people are more likely to have diabetes compared to their counter parts.
4. **Physical Activity:** Physical activity helps in controlling blood glucose, High-density lipoprotein cholesterol, blood pressure and triglycerides resulting into lower risk of diabetes. Many risk factors are directly related to physical activity such as BMI and waist circumference. Thus making it one of the major risk factor involved in many chronic diseases.
5. **Smoking – Cigarette Use:** Research conducted by Julie C Will shows an increase in diabetes rate with the increase in smoking. It shows that men who smoked have 45% higher diabetes rate compare to the men who had never smoked thus making smoking as an important indicator of diabetes.
6. **Alcohol Use:** Alcohol consumption has become an important risk factor for diabetes. Many researchers investigated that moderate intake of alcohol is associated with reduced risk of diabetes, however heavy intake of alcohol increases the risk of diabetes.
7. **Age:** Prevalence of T2DM increases with age. In affluent countries, T2DM is found to develop in middle to older age groups. However, in developing countries like India, T2DM occurs in lower age groups as well. It is found that in the Indian population, T2DM peaks at 60-69 years of age. However, the prevalence of T2DM in persons under age 44 years increased form 25% in 2000 to 46% in 2016. Thus, T2DM seems to be affecting both the younger and older population increasingly by the years.

8. **Family History:** T2DM is associated by strong familial influences. These risks are increased if the family member who is affected had early onset of T2DM.
9. **Obesity:** Being overweight or obese is the main risk factor for developing diabetes. By reducing obesity, there is a 50-75% risk reduction of developing diabetes.

1.4 Symptoms of Diabetes

Diabetes signs vary relying on how much blood sugar is expanded. Few people, especially those people with type 2 diabetes or pre diabetes, may not know the symptoms at initial stage. In type 1 diabetes, signs and symptoms have a tendency to come on quickly and be more intense. There are various symptoms of TID and T2D diabetes that are to be followed increased thirst, extreme starvation, frequent urination, Presence of ketones inside the urine, Irritability, slow-recuperation sores, Blurred vision, unexplained weight reduction, frequent infections, inclusive of gums or pores and skin infections and vaginal infections and Fatigue. TID can develop at any age, though it frequently seems at some point of early life. T2D, can increase at any age, although it is common in people older at the age of 40.

1.5 Role of Machine Learning in Healthcare Industry

ML is utilized as a part of disease identification and analysis of ailments is at the forefront of ML research in clinical drug. Personalized medicine, or extra effective remedy based on character health information paired with predictive analytics, it is closely related to better disease evaluation. Machine learning can help hospital systems identify patients with undiagnosed or misdiagnosed chronic disease, predict the likelihood that patients will develop chronic disease, and present patient-specific prevention interventions.

1.6 Motivation of the thesis

The diagnosis process is a decision making process in which decisions are made by the medical experts with the help of their knowledge and the experience they get with the treatment of patients suffering from same problem and symptoms. Disease diagnosis is a complicated process and may lead to wrong assumption as some factors are associated with many organs. Hence, there is a need to automate the medical diagnosis process and develop a diagnosis system to determine the stroke level of diabetes disease with higher precision and without causing any delay in the proper subsequent action. The number of tests is conducted on the patient for the diagnosis of a disease.

With the use of machine learning technique for predicting the disease, the number of tests can be reduced which saves time and provides quality services at a reasonable cost.

1.7 Thesis Organisation

The proposed work shows an effective methodology for predicting the diabetes. The results of the classification models are examined and final output is based on the accuracy of the model.

CHAPTER 2 describes the literature review and the various methods of machine learning are used for prediction.

CHAPTER 3 identifies the techniques used for our proposed work. It also gives a detailed description of Machine Learning techniques used to build the prediction models

CHAPTER 4 discusses the problem statement. It includes the main aim of carrying this research work and objective of the thesis.

CHAPTER 5 provides an overview of the research methodology used in order to solve the problem stated. It includes the description of work flow.

CHAPTER 6 shows the results and discussion to find the best model. It include experimental setup, results and observations.

CHAPTER 7 discusses the conclusion and possible future work in this field of study.

CHAPTER 2

LITRATURE REVIEW

Loannis et al. [6] used support vector machines (SVM), Logistic Regression (LR) and Naïve Bayes (NB) using 5 fold cross validation to predict different/varies medical datasets including diabetes dataset. The researchers compare the accuracy and the performance of the algorithm based on their result and the researchers conclude that SVM provides best accuracy than the other algorithm. Francesco et al. [7] multilayer perceptron, random forest and Decision Tree machine learning Algorithms for prediction. The researchers concluded that multilayer perceptron (MP) provides high accuracy.

Kandhasamy and Balamurali [8] used Artificial Neural Networks (ANN), K-Nearest Neighbor (K-NN), NB, J48 classification techniques. From those algorithm, Naïve Bayes provide better accuracy in diabetes dataset in this study. The two algorithms KNN and ANN provide high accuracy in other datasets on their study. Meng et al. [9] proposed different data mining techniques to predict the diabetic diseases. In this, three techniques ANN, LR, and j48.The outcome illustrations that the j48 machine learning technique provide better accuracy.

Zhang et al. [10] initialized different machine learning techniques to predict the diabetic diseases i.e. Naïve Bayes Random forest and Adaboost. Researchers conclude that Random forest was better than other models. Francesco et al. [11] uses j48, multilayer perceptron and Random Forest. j48 was provide better accuracy i.e. 77.5% in comparison to others.

Rani and Jyothi [12] used Naïve Bayes and ANN on different datasets including diabetes. Naive Bayes provided high accuracy with the accuracy value of 77.01%. Saravananathana and Velmurugan [13] used various ML algorithms such as CART, SVM, and K-NN. In their study, CART gives 62.28% accuracy, SVM 65.04% and K-NN 53.39%.

Yasodha and Kannan [14] applied ML classifiers on diverse types of datasets to decide if a person is diabetic or not. In this study, the implementation was done by using WEKA to classify the data and the data is assessed by means of 5-fold cross validation approach, as it performs very well on

small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Sumbaly et al. [15] discovered solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The experimental results shows that j48 algorithm gives better accuracy rate of 74.8% as compared to Naive Bayes. Gupta et al. [16] aimed to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyze the results of several classification methods in WEKA. The result shows that Random Forest shows highest accuracy i.e. 81.3%, sensitivity is 59.7% and specificity is 81.4%.

Chikh et al. [17] used enhanced Artificial Immune Recognition System 2 (AIRS2) called modified Artificial Immune Recognition System 2 (MAIRS2) to increase the diagnostic accuracy of diabetes diseases. K-nearest neighbor's algorithm swap with the fuzzy K-nearest neighbors to enhance the diagnostic accuracy of diabetes diseases. The authors attained a good tradeoff between classification accuracy and data reduction. The propose system (MAIRS2) that performed better than classical AIRS2. The authors achieved highest classification accuracy by MAIRS2 is 80.10%.

Sharmila and Manickam [18] aimed to analyze the data in predicting the diabetes from medical record of the patients. This study is analyzing the diabetes from huge medical records by using decision trees algorithm using R tool. The accuracy obtained by decision tree algorithm is 79.33%.

Lavanya et al. [19] focused on a prediction model by analyzing the algorithm in Hadoop/Map Reduce environment to predict the widespread types of diabetes and the related problems and also the treatment. The suggested design of predictive analysis system is constructed on numerous levels e.g. data collection, warehousing, predictive analysis, processing analyzed reports.

Tiwari and Diwan [20] has given an automatic and hidden approach to identify, patterns that are hidden, of cancer disease. The given system used data mining techniques such as association rules and clustering. Attribute based clustering for feature selection is an important task in this paper. In this method they have done vertical fragmentation in the data set. Here the data set is divided into two clusters, one cluster has all the relevant attributes and the other cluster has all the irrelevant attributes.

Khaleel et al. [21] concentrated on examining data mining strategies which are required for medical information mining particularly to find frequent infections, for example, heart sicknesses, lung malignancy, breast cancer etc. The data mining techniques have been applied to medical data include Apriori and FPGrowth and unsupervised neural networks, linear programming, Association rule mining. The association rule mining discovers frequently occurring items in the give dataset. The medical mining yields required business intelligence to support well informed diagnosis and decisions.

Chaurasia and Pal [22] initialized different machine learning techniques to predict the diabetic diseases. Information mining device WEKA is utilized which contains an arrangement of machine learning algorithms with the end goal of mining. For this study NB, j48 and bagging are utilized. j48 gives 84.35% of accuracy. 85.03% of accuracy is accomplished by Bagging. Bagging gives the better classification factor on this dataset.

Parthiban and Srivatsa [23] had done a work on finding of coronary illness in diabetic patients. For this machine learning techniques were utilized. Naive Bayes and SVM methods are utilizing by WEKA. The highest accuracy of 94.60 is achieved by utilizing SVM. Sumbaly et al. [24] had done a work to foresee diabetes illness by utilizing two methods decision tree and Naive Bayes. Cross Validation and Percentage Split (PS) separately by j48. Naive Bayes gives 79.5652% exactness by utilizing PS. Algorithms gives most elevated accuracy by utilizing rate split test.

Alic et al. [25] worked on comparative analysis of most commonly used disease prediction techniques that are ANN and Bayesian Networks. In which higher accuracy is achieved by Artificial Neural Networks with 89.78% as compared to Bayesian Network 80.43% due to independent relation between observed nodes.

Murthy et al. [26] anticipated the liver illness utilizing Support vector machine (SVM) and Naive Bayes Classification algorithms. Data set comprises of 560 occurrences and it additionally comprises of 10 attributes. Comparison is done in light of the precision and time taken for execution. Naive Bayes gives 61.28% exactness in 1670 ms. 79.66% accuracy is obtained inside 3210 ms by utilizing SVM. SVM gives most elevated accuracy when contrasted with the Naive Bayes for the expectation of liver disease. Regarding time taken for execution, Navies Bayes takes less time when contrasted with the SVM.

Baby and Vital [27] developed a model that takes kidney disease data set of patients and developed a model that can predict the type of kidney disease. The model used several classification algorithms like random forests, Decision Trees, j48, K-means algorithms and compared the results upon statistics that showed RF gives better result than the other algorithms.

Razia and Rao [28] developed a framework model to diagnose the thyroid disease using machine learning techniques. The unsupervised learning and supervised learning are used to diagnose the thyroid disease and compared with the decision tree model ultimately the framework model is outperformed than the decision tree model.

Alehegn et al. [29] initialized different machine learning techniques to predict the diabetic diseases i.e. SVM, Naïve Net, Decision Stump, and Proposed Ensemble method. Proposed ensemble method was provide better accuracy i.e. 90.36%. In comparison to others. Kang et al. [30] used AdaboostM1 algorithm along with random committee. The prediction accuracy obtained is 81.0%.

CHAPTER 3

MACHINE LEARNING

3.1 Machine Learning

Machine Learning is a branch of computer science that consists of algorithms that can learn from data. It provides set of techniques that can detect patterns in the data and use the patterns to generate future predictions. ML deals with the problem of how to build computer programs that automatically improve with experience. Much of a computer's activity was, and continues to be, designed to perform a set of predetermined instructions. A design is developed to solve a question or automate a process and the software does so in a deterministic manner. That brings up the question of why machines need to learn at all. Certain tasks are too computationally intensive, or too mathematically complex, to be performed in a deterministic manner. Often, in these cases, a process based on machine learning that provides good enough results can be found. In addition, machine learning can also help find quite usable solutions for problems that may not be trivial to solve (through regression, optimization etc.). The learning comes into play because a machine learning algorithm first needs to train on a set of data that helps adapt the algorithm to a desirable state. This helps in several ways, including working with datasets that are too large for human beings to comprehend in their entirety. Also, subtle changes in large datasets that continue to grow can be easily missed by human beings; however, they can be quickly caught by a machine learning algorithm. Machine learning as a field encompasses many other fields and areas that in conjunction make learning possible. The most important of these are computer science, statistics, and probability theory. Together, these fields can be used to solve problems that have become increasingly more relevant in modern society as more and more data is collected and stored related to a number of different areas. One of the challenging and most exciting parts of machine learning has been the vast array of application areas that have come up since its inception i.e. astronomy, oil and gas exploration, web-user activity analysis, page ranking, collaborative filtering, translation, etc.

Machine learning methods are used for a variety of tasks, the primary ones being classification, regression, clustering and optimization. As mentioned previously, many of these tasks necessitate machine learning simply because understanding and developing a mathematical model for the

underlying system is not efficient. Hence, a good approximation provided by a machine learning algorithm is preferred.

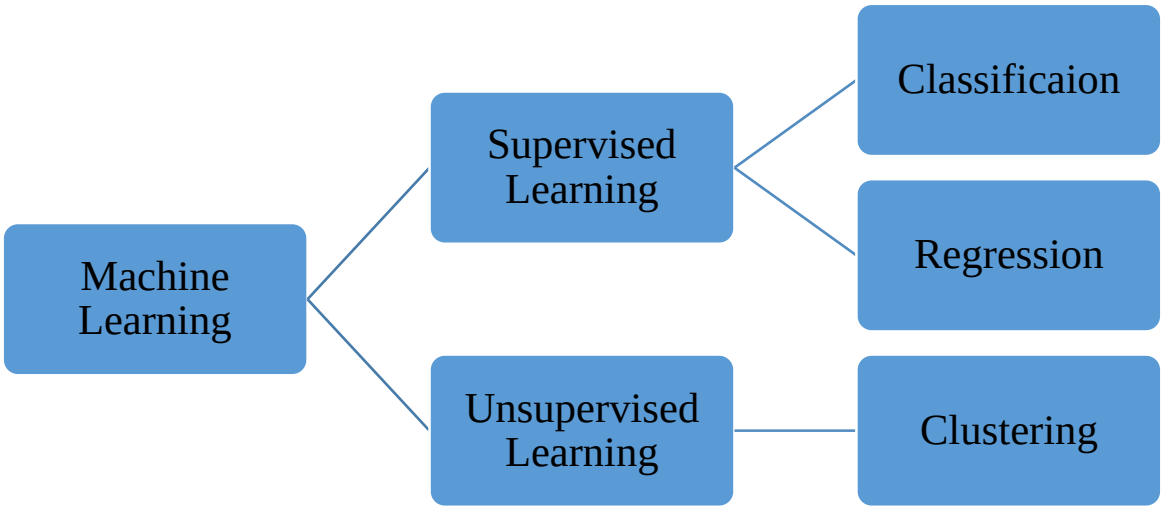

Fig. 3.1 Machine Learning Techniques

Machine learning is divided into two main types supervised and unsupervised learning.

3.1.1 Supervised Learning: We have a data set that includes the target values (the values we wish to predict). We try to learn a function that correctly predict the target values from the other features, which can then be used to make predictions about other examples. Typical examples: classification, regression. Classification refers to the task of identifying what group a given object belongs to, given information related to its attributes. The attribute data can be discrete or continuous, while the classes are discrete. Each object typically belongs to only one class. A machine learning method can be used to create a classifier, given training data made up of a number of objects that already have class assignment. Some commonly used classifiers are DT, NN and SVM. Each classifier has its own strengths and weaknesses. Machine learning algorithms help us identify classes by generalizing data. Specifically, domain knowledge that help classify the data are essential. In cases where classes are given, supervised learning is used most often. Regression is a supervised learning method. It predicts continuous valued output. The regression is used to predict the numeric data instead of labels. It can also identify the distribution trends based on the available data or historic data. Predicting a person's income from their age, education is example of regression task. In other way, target attribute containing continuous values requires a regression

technique. The most commonly used regression type is linear regression. In this, the line that minimizes the average distance among all the points from the line i.e. that best fits the data is calculated.

3.1.2 Unsupervised Learning: In this, we have a dataset but there is no target to be predicted. Rather, we want to learn a model that might have generated that set. Typical examples: clustering, density estimation, noise reduction. Clustering is frequently performed when the training data does not come with class labels. In fact, the number of classes may not be known either, though often, an assumed number of classes is used in clustering algorithms. Unsupervised learners find similarities in data to identify multiple clusters of points. Each cluster can be assigned a class of its own or several clusters may be merged to form a super cluster and given one class label. One of the more common clustering algorithms is the K-means algorithm which clusters points by proximity to mean of each cluster. The mean is repeatedly evaluated and each point is assigned to the closest cluster, until the change in clusters goes below a certain threshold.

3.2 Sample Data and Building Models

A prerequisite for learning is that there is something to learn from, samples are needed. For machines these samples are sets of data. For a ML algorithm to derive a stable and durable model, it is necessary that the data is suitable. A correlation or a context between the sample data, or a transformation of the sample data and the desired outcome is needed. There may be several problems with the sample data which makes it more difficult to derive a sustainable model for prediction. The data should to be stationary, meaning that the distribution does not change over time. The degree of random noise in the sample data should not be too high and equal inputs should yield equal outputs. When data is gathered it is usually divided into Training and Test set. The data that you apply to build the model is called training set. Training set is used to fit the ML model by pairing the input with the expected output, and trying to find relationship between the variables and minimizing the error. After fitting the model, the test set is put to use. Since the Test set is unseen by the model we can use it to get a final measure on how well the ML model fits the data; this measure should indicate how well the model will perform on real-world data.

3.3 Machine Learning Models

3.3.1 Support Vector Machines (SVM)

Support Vector Machines belong to the area of supervised learning methods and therefore need labeled, known data to classify new unseen data. This technique uses associated learning algorithms for highlighting patterns and understanding data in order to use them for classification and regression analysis. It is calculating the maximal margin between all dimensions i.e. it is creating the largest distance between instances, which is reducing the generalization error. The basic approach to classify the data, starts by trying to create a function that splits the data points into the corresponding labels with (a) the least possible amount of errors or (b) with the largest possible margin. This is due to the fact that larger empty areas next to the splitting function result in fewer errors, because the labels are better distinguished from one another.

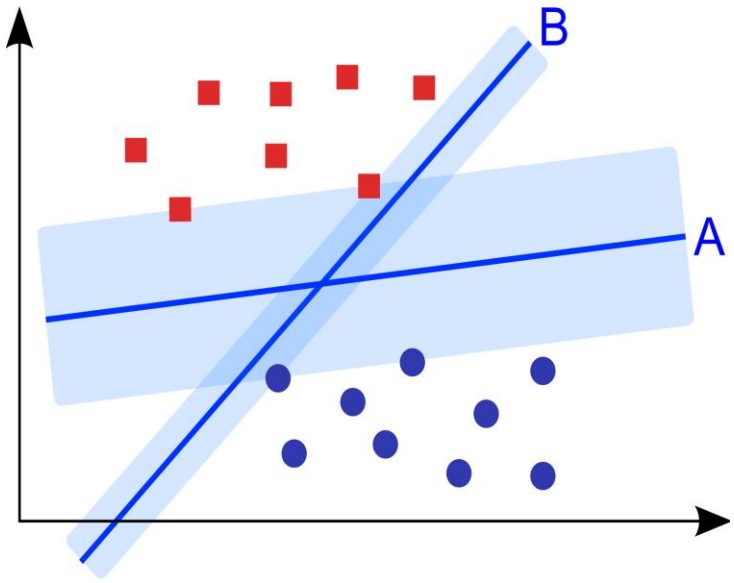

Fig. 3.2 Support Vector Machine [4]

The margin around a separating function is being used as an additional parameter to evaluate the quality of the separation. In this case the separation A is the better one, since it distinguishes the two classes in a more precise manner. Formally, SVM create one or multiple hyper planes in n-dimensional space. The first attempt in the process of splitting the data is always to try to linearly separate the data into the corresponding labels. In this SVM radial basis function is used to find set weights for a cure fitting problem. The learning helps to find out the surface in high dimensional

space which provide best fit to the training data. The hidden layers supports a set of functions that comprises an arbitrary basis for input basis, such function is known as radial basis functions.

3.3.2 Decision tree

A Decision Tree is a classification technique that focuses on an easily understandable representation form and is one of the most common learning methods. Decision trees use data sets that consist of attribute vectors, which in turn contain a set of classification attributes describing the vector and a class attribute assigning the data entry to a certain class. A decision tree is built by iteratively splitting the data set on the attribute that separates the data as well as possible into the different existing classes until a certain stop criterion is reached. The representation form enables users to get a quick overview of the data, since decision trees can easily be visualized in a tree structured format. One of the first algorithms concerning decision tree training were the Iterative Dichotomiser3 (ID3) and its successor the C4.5 algorithm, both developed by Ross Quinlan in 1986 and 1993 [31]. These algorithms formed the basis for many further developments. Decision trees are directed trees, which are used as a decision support tool. They represent decision rules and illustrate successive decisions. In decision trees, nodes can be separated into the root node, inner nodes, and end nodes, also called leafs. The root node represents the start of the decision support process and has no incoming edges. The inner nodes have exactly one incoming edge and have at least two outgoing edges. They contain a test based on an attribute of the data set. If the splitting attribute is of numeric type there is no possibility to split the records into all outcomes of the attribute. This is one of the main upgrades of the C4.5 DT compared to the ID3. The C4.5 is additionally able to calculate the best splitting points for numeric attributes as well and split them by using greater than, equal or smaller than operators.

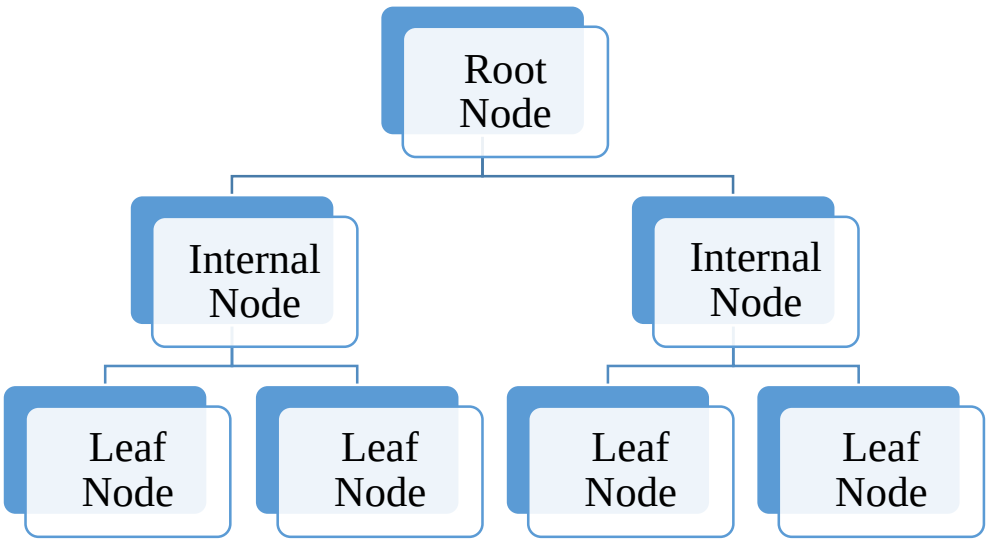

Fig. 3.3 Decision Tree

3.3.3 K-Nearest Neighbor

K-Nearest Neighbor model is also a type of supervised learning algorithm. It is the simplest and easy than other machine learning techniques. This algorithm is representative of lazy algorithms. It is based on assumption that records within a dataset are generally having the same properties. Consequently, in labeled data, when new instances are coming for labeling, the model is finding the closest k neighbors in many dimensional spaces and classifies new one according to the mode in case of categorical label or the average in case of continuous label. K-NN algorithm is relatively slow in classification of new instances coming into model, but fast during training process. Also, this algorithm is very sensitive to noise in dataset.

3.3.4 Random Forest

Random forest tree is based on decision trees. Random Trees lies in one of those class of ML algorithms which does ensemble classification. The random forest model could be a form of additive model that produces predictions by combining choices from a sequence of base models. Random forest or random decision forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision.

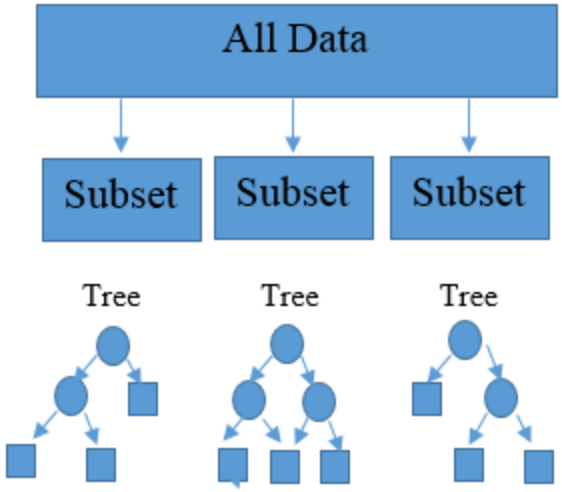

Fig. 3.4 Random Forest

3.3.5 Naive Bayes

Naïve Bayes is a classification technique based on Bayes' theorem. This classifier is widely using in text estimation. For instance, many spam filters are using it in order to divide acceptable content from unacceptable. Usually, the accuracy of this method is relatively low in contrast with other approaches. However, an advantage of this technique is very high speed of classification and also very good level of tolerance to missing values. Additionally, Naïve Bayes algorithm characterized by low tolerance to redundant attributes. Continuous features are not permitted here.

Naïve Bayes works on conditional probability P (A|B).

$$P\,(A|B) = \frac{P\,(B|A) * P(A)}{P\,(B)}$$

Where,

A and B are events.

P(X): Is the probability of event X

P (A|B): Probability of event A if B is already occurrence.

3.3.6 Neural Network

A neural network is a supervised machine learning algorithm, which as the name indicates, is modeled after the neurons in brain. The goal of the algorithm is to assign each input to one of the many predefined output classes based on a prediction value. The processing in a neural network occurs when input values pass through a series of batches of activation units. These batches are called layers. A layer of activation units takes the outputs of the previous layer as inputs and processes them simultaneously to generate outputs that are that passed on to the next layer. This process continues until the last layer generates the final cumulative predictions.

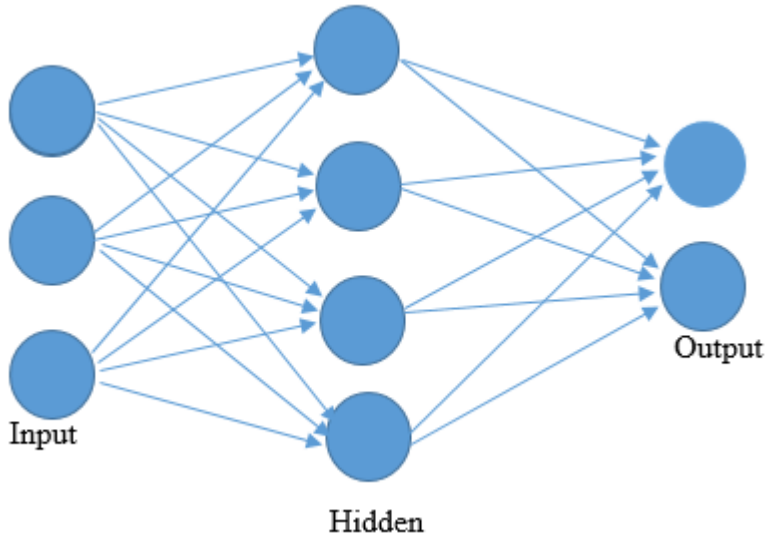

Fig. 3.5 Neural Network

Input layer takes its inputs directly from the features. The intermediate layer of activation units takes response from the input layer and feeds the output to the output layer; hence it's hidden from test inputs and final predictions of our network. The output layer has one-to-one correspondence with the final predictions.

3.3.7 Adaboost

Adaboost is a machine learning meta-algorithm. Adaboost is used in binary classification problem. It is also a boosting algorithm. It is a basic algorithm for understanding boosting. It can be used for boosting the performance of the algorithms. It works well with weak learners.

3.3.8 Linear Model

It uses linear models to carry out regression, single stratum analysis of variance and analysis of covariance without weight or the modified classification [32].

Table 3.1 Classification Algorithm comparison

Feature	**Naïve Bayes**	**SVM**	**Decision Tree**	**Random Forest**	**K-NN**	**Neural Network**	**Ada boost**	**Linear Model**
Training Speed	Very fast	Fast	Fast	Slow	Fast	Slow	Fast	Fast
Accuracy	High	High	Low	Low	Low	High	High	Low
Problem Type	Classification	Either	Either	Either	Either	Either	Either	Either
Prediction Speed	Fast	Fast	Fast	Moderate	Depends on n	Slow	Fast	Slow
Performs well with small number of observations?	Yes	Yes	No	No	No	No	No	Yes

3.4 Cross-Validation

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Methods of Cross Validation

1. **Holdout method:** Holdout method is a simplest kind of cross validation. The data set is divided into the training and testing set.
2. **Leave one out Cross Validation:** In this method, we perform training on the whole data-set but leaves only one data-point of the available data-set and then iterates for each data-point. It has some advantages as well as disadvantages also. An advantage of using this method is that we make use of all data points and hence it is low bias. The major drawback of this method is that it takes a lot of execution time as it iterates over 'the number of data points times.
3. **K- Fold Cross Validation**: In this method, firstly, we split the data-set into k number of subsets then we perform training on the all the subsets but leave one (k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

CHAPTER 4

PROBLEM STATEMENT

4.1 Problem Statement

Diabetes is the leading worldwide reason of death. The count of people dying every year from diabetes disease is increasing drastically. Worryingly, diabetes is now being shown to be associated with a spectrum of complications and to be occurring at a relatively younger age within the country. In India, the steady migration of people from rural to urban areas, the economic boom, and corresponding change in life-style are all affecting the level of diabetes. Yet despite the increase in diabetes there remains a paucity of studies investigating the precise status of the disease because of the geographical, socio-economic, and ethnic nature of such a large and diverse country. The diagnosis process is a complicated process in which doctors make a decision with the help of their knowledge and the clinical data available which may lead to a wrong assumption due to the complex association among various factors. The disease is now highly visible across all sections of society within India, there is now the demand for urgent research and intervention - at regional and national levels - to try to mitigate the potentially catastrophic increase in diabetes that is predicted for the upcoming years. There is a need to detect and diagnose the disease at an early stage so that considerable life can be saved. The work presented in this thesis is intended to automate the medical diagnosis process and develop a prediction system to detect the diabetes disease using machine learning with higher accuracy.

4.2 Objectives

The objective of our thesis are as follows:

1. In the literature review we study the various ML models and techniques.
2. To predict the diabetes disease by considering eight classification models.
3. To evaluated the performance on the basis of Sensitivity, Specificity, Precision, Recall, and Accuracy.
4. To propose ensemble technique on top five models.
5. Compare the performance of proposed model with existing models.

CHAPTER 5

PROPOSED METHODOLOGY

5.1 Proposed Approach

The first and foremost step is to collect the dataset required for the study. The methodology is applied to a factual data having information about the patient who suffering with the diabetes. The workflow of the research is shown in figure 5.1.

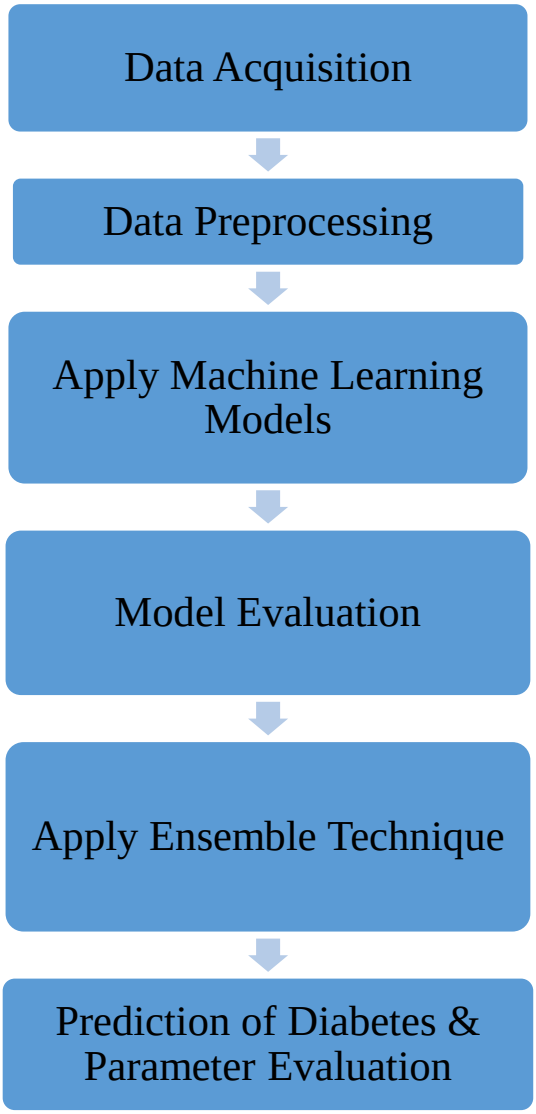

Fig. 5.1 Workflow of Proposed Approach

The pseudo code of proposed approach is shown in this figure 5.2:

Initialize startTime to currentTime.

Include library

Initialize modelName = "svm"

Initialize InputDataFileName to the data file to be used.

Initialize training percentage to 70.

Read the inputDataFileName Load the dataset from inputDataFileName to dataset.

Count total number of dataset rows (totalDataset) after shuffling the dataset row wise.

Choose target variable.

Choose input variable.

Select training dataset trainDataset.

Select testing dataset testDataset.

Show top 6 records of testDataset and total number of testDataset.

Assign formula = as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))

Build model for training from ksvm passing formula in params,

model = ksvm(formula, trainDataset, kernel="rbfdot", prob.model=TRUE)

Predict the data using model and testDataset.

Assign the rounded off value of the prediction to Predicted.

Extract actual results Actual <- as.double(unlist(testDataset[target]))

//Model Evaluation start

Create the confustion matrix.

Evaluate the parameters using round(HMeasure(Actual,PredictedProb)$metrics,3)

Calculate accuracy, accuracy = round(mean(Actual==Predicted) *100,2)

Set totalTime = proc.time()[3] - startTime.

Plot the curves.

a. ROC and ROCH Curves.

b. H Measure Curve.

c. AUC Curve.

d. SmoothScoreDistribution Curve.

Save evaluation result to EvaluationsParameters.

//Model Evaluation ends.

Write the evaluation results to a csv file.

Write Actual and prediction data results to another csv file.

Save the model.

Display total time taken for the process.

Fig. 5.2 Pseudo code of purposed Approach

5.2 Description of workflow

5.2.1 Data Acquisition

The data is acquired from Kaggle known as Diabetes dataset, it contains 768 instances and 9 attributes such as plasma glucose, diastolic BP, BMI, age etc. [33]. The datasets includes 500 tested negative while 268 of them were tested positive.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1

Fig. 5.3 Dataset for Diabetes [33]

In diabetes dataset all the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. Second is the value of glucose concentration a 2 hours in an oral glucose tolerance test and then is the diastolic blood pressure, fourth in line is the skin thickness, then is the 2-Hour serum insulin, sixth is Body mass index (BMI) and then seventh is the diabetes pedigree function (DPF) and the second last value is the that of the age in years. The ninth column is that of the class variable (0 or 1), 0 for no diabetes and 1 for the presence.

5.2.2 Data Preprocessing

Once data is acquisition, it is transform into required form for the machine learning process, which is known as pre-processing phase. Data preprocessing is done to reduce the noise and missing values.

5.2.3 Apply Machine Learning Models

Different machine learning models are applied to the given dataset such as SVM, DT, RF, K-NN, Naïve Bayes, NN, Adaboost, and LM. Dividing dataset into training and testing part i.e. 70% and 30%.

1) Support Vector Machines

Support Vector Machines belong to the area of supervised learning methods and therefore need labeled, known data to classify new unseen data. In this SVM radial basis function is used to find set weights for a cure fitting problem. The learning helps to find out the surface in high dimensional space which provide best fit to the training data. The hidden layers supports a set of functions that comprises an arbitrary basis for input basis, such function is known as radial basis functions as mention in equation 5.1

$$model \ <- ksvm(formula, trainDataset, kernel = "rbfdot", prob.model = TRUE) \quad (5.1)$$

Where,

Ksvm: ksvm implementation in the R package 'kernlab'

Formula: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + Age

Kernel: The kernel function used in training and predicting. It computes a dot product between two arguments.

prob.model: If it set to TRUE builds a model for calculating class probability.

2) Decision tree

In decision tree use split method for splitting the data set on the attribute that separates the data as well as possible into the different existing classes until a certain stop criterion is reached. DT is additionally able to calculate the best splitting points for numeric attributes as well and split them by using greater than, equal or smaller than operators as mention in equation 5.2

$$model \ <- rpart(formula, trainDataset, method = "class", parms = list(split = "information"), control = rpart.control(usesurrogate = 0, maxsurrogate = 0)) \quad (5.2)$$

Where,

rpart: Is the implementation package for decision tree.

Usesurrogate: It is used in the splitting process. 0 means display the missing value, I means split subject missing the primary variable. If all surrogate are missing its means no split in missing observation.

Maxsurrogate: the number of surrogate splits retained in the output. If this is set to 0 the compute time will be reduce.

Split: number of observations that must exist in a node in order for a split to be attempted.

3) K-Nearest Neighbor

In K-NN we used kknn implementation package as mention in equation 5.3

$$model \ <- train.kknn(formula, trainDataset, trace = FALSE, maxit = 1000) \quad (5.3)$$

Where,

Kknn: Is the implementation package for K-NN.

Train: Matrix or data frame of training set cases.

maxit: containing the mean and variance of the misclassification error, the absolute and the squared distances.

Trace: TRUE/FALSE if additional information about the imputation process should be printed.

4) Random Forest

Random forest or random decision forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision as mention in equation 5.4

$$model <- randomForest(formula, trainDataset, ntree = 500, mtry = 2) \quad (5.4)$$

Where,

Random forest: is the implementation package for random forest.

ntree: Number of trees to grow.

mtry: variables randomly sampled at each split.

5) Naive Bayes

The goal of the algorithm is to assign each input to one of the many predefined output classes based on a prediction value. The processing in a neural network occurs when input values pass through a series of batches of activation units. These batches are called layers as mention in equation 5.5

$$model <- train(formula, trainDataset, method = "nb") \quad (5.5)$$

Where,

Method: nb method used in the Naïve Bayes classifier.

6) Neural Network

In neural network nnet method is used for predicting the results. In this Input layer takes its inputs directly from the features. The intermediate layer of activation units takes response from the input

layer and feeds the output to the output layer; hence it's hidden from test inputs and final predictions of our network. The output layer has one-to-one correspondence with the final predictions as mention in equation 5.6

$$model \ <- nnet(formula, trainDataset, size = 10, linout = TRUE, skip = TRUE, MaxNWts = 10000, trace = FALSE, maxit = 100) \quad (5.6)$$

Where,

nnet: Is the implementation package for neural network.

Linout: linear output units by default it is TRUE

Size: Number of units in the hidden layer it can be 0 if there are skip-layer units

skip: skip-layer connections from input to output.

MaxNWts: maximum allowable number of weights. Increasing MaxNWts are very slow and time-consuming.

trace: Switch for tracing optimization by default it is FALSE.

Maxit: Parameter for weight decay. Default 0.

7) Adaboost

In adaboost model ada package is used for implementation. AdaBoost is used in binary classification problem. It is also a boosting algorithm. It is a basic algorithm for understanding boosting. It can be used for boosting the performance of the algorithms. It works well with weak learners as mention in equation 5.7

$$model \ <- ada(formula, trainDataset) \quad (5.7)$$

Ada: Is the implementation package for adaboost.

8) Linear model

In linear model stats package and lm method used for implementation as mention in equation 5.8

$$model \quad <- multinom(formula, trainDataset, trace = FALSE, maxit = 15000) \qquad (5.8)$$

Where,

trace: Switch for tracing optimization. Default FALSE.

Maxit: Parameter for weight decay. Default 0

5.3 ROC (Receiver operating characteristic) curve

ROC graphs are representations of the sensitivity i.e. called the true positive rate on the Y axis and 1-specificity i.e. called the false positive rate on the X axis, corresponding to each possible cut-off point. The area under ROC Curve gives the probability that, when one draws one majority and one minority class example at random, the decision function assigns the higher value to the minority class than the majority class sample. AUROC is not sensitive to the class distributions in the dataset. Accuracy is shown as the area under the curve. The greater the area under the curve, the more accurate the test. A perfect test has an area under the ROC curve of 1. The random guess is a diagonal line drawn at 45 degrees from the x-axis. The diagonal line in a ROC curve represents perfect chance. In figure 5.7 point C shows the best predictive among A and B. The result of B lies on the random guess line.

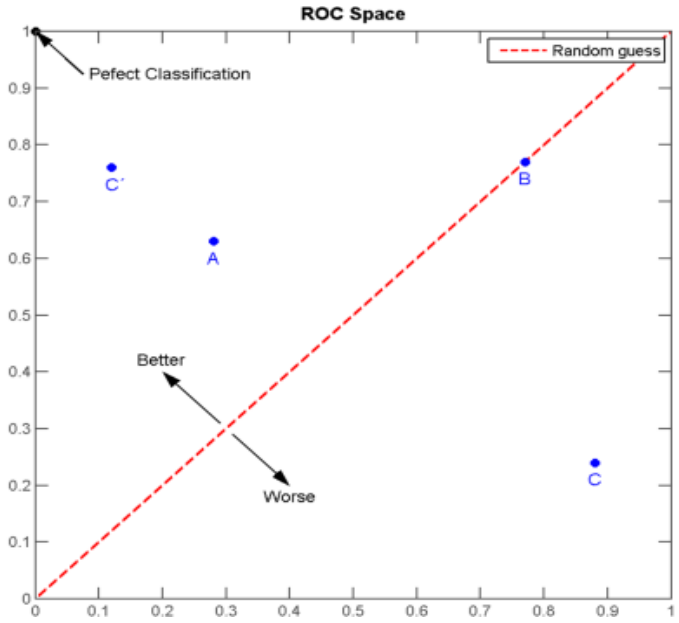

Fig. 5.4 ROC Space Detail [5]

5.4 Proposed Ensemble Technique

Ensemble technique is the process of combining the multiple models to improve the accuracy of the models. Ensemble models perform better as compared to single individual model. An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. Ensemble is done to improve the accuracy of the models simply by combining the different models.

In this boosting technique is used for the combining the top five models i.e. SVM, Adaboost, Naïve Bayes, Decision tree and Neural network models and evaluate the performance of combining models in terms of accuracy. Ensembling is done on the basis of actual and the predicted value of the models. With Boosting, the ensemble is trained in serial i.e. the next classifier is trained based on the error of the previous classifier during several iterations. In each iteration the weights of the examples are changed based on the error of the trained classifier, and more weights are assigned to the misclassified data. Then, in the next iteration, a weighted sample of the data is extracted based on the weights of the data that have been assigned in the previous iteration, and a new classifier is trained on the samples. The final label of the data is calculated by weighted averaging of different classifiers outputs.

CHAPTER 6

RESULTS AND DISCUSSION

The experimental study is implemented by using R language. It is an open source programming language, run on 64 bit operating system with 4 GB RAM. The Diabetes dataset is used that consists of 9 attributes and 768 instances. The SVM classifier with radial basis function is implemented using 70-30 rule and K-fold cross validation using 90-10 rule. The performance of proposed ensemble method is improved by implemented different machine learning classifiers.

6.1 Performance metrics

The performance of the system is evaluated in terms of accuracy, precision and recall using the below parameters:

Specificity: Specificity is the ability of a test to detect absence of a disease when no disease is present.

Sensitivity: Sensitivity is the ability of a test to detect disease when disease is present.

Accuracy: It is the percentage of correctly predicted samples to the total number of samples.

Precision: It is the fraction of retrieved samples that are predicted correctly.

Recall: It is the percentage of correctly predicted samples that are retrieved.

Table 6.1 Evaluation of Machine Learning Models with 70-30 rule

Models	Sensitivity	Specificity	Precision	Recall	Accuracy
DT	0.506	0.886	0.710	0.506	78.79
K-NN	0.612	0.868	0.71	0.612	77.92
NB	0.63	0.785	0.611	0.63	79.22
NN	0.537	0.919	0.786	0.537	78.35
RF	0.58	0.873	0.712	0.58	77.06

SVM	0.809	0.929	0.78	0.809	80.52
AdaBoost	0.767	0.893	0.771	0.767	80.02
Linear Model	0.6	0.884	0.75	0.672	75.92
Proposed Ensemble Model	0.968	0.947	0.852	0.968	84.82

ROC graphs are bi-dimensional representations of the sensitivity [called the true positive rate (TPR) on the Y axis] and 1-specificity [called the false positive rate (FPR) on the X axis], corresponding to each possible cut-off point (classifying value). Both the TPR and the FPR ranges are from 0 to 1. In figure 6.1, the curve shows the accuracy of our model. The accuracy obtained by these models are 80.52%, 77.92 %, 79.92 %, 78.79%, 78.35%, 77.06%, and 80.02% 75.92 % of SVM, K-NN, DT, NN, RF, Adaboost, and Linear model respectively.

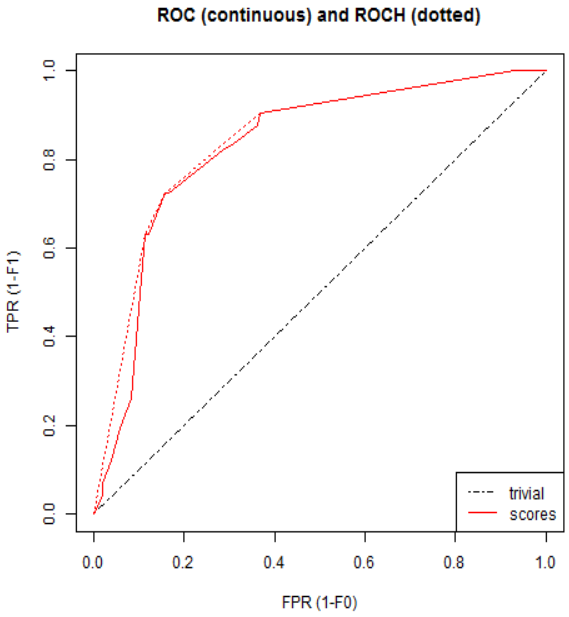

Fig. 6.1 SVM ROC Curve

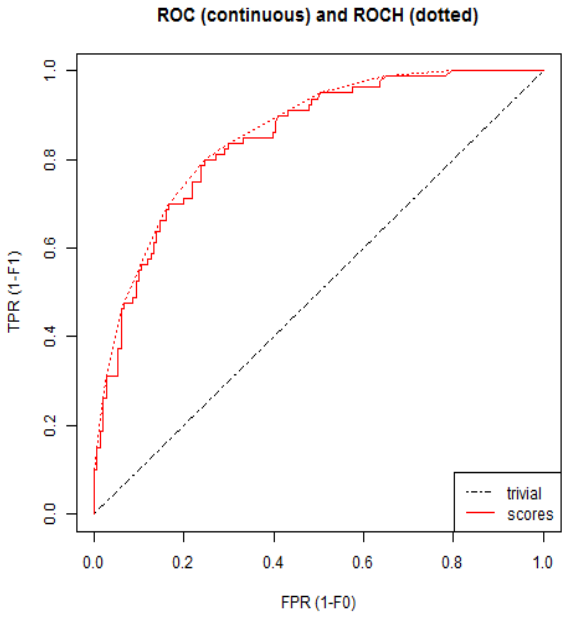

Fig. 6.2 K-NN ROC Curve

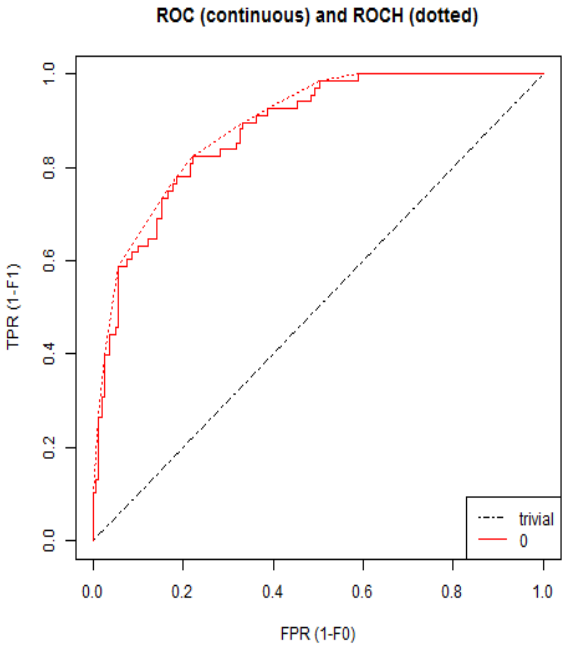

Fig. 6.3 Naïve Bayes ROC curve

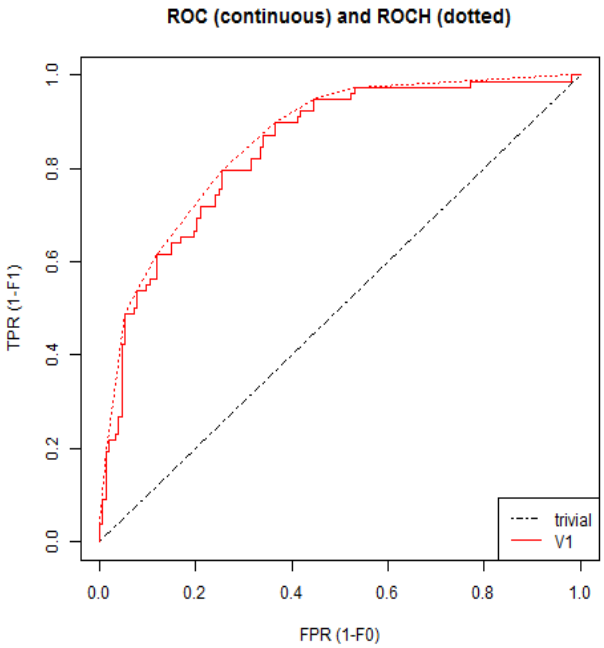

Fig. 6.4 Decision Tree ROC curve

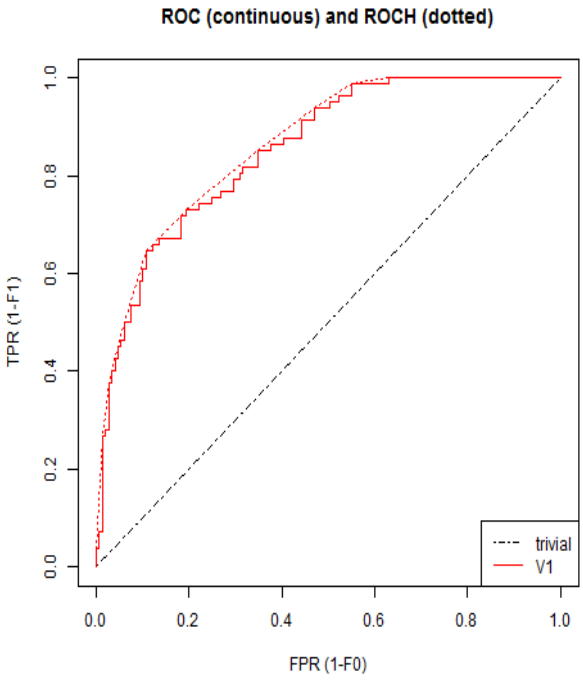

Fig. 6.5 Neural Network ROC curve

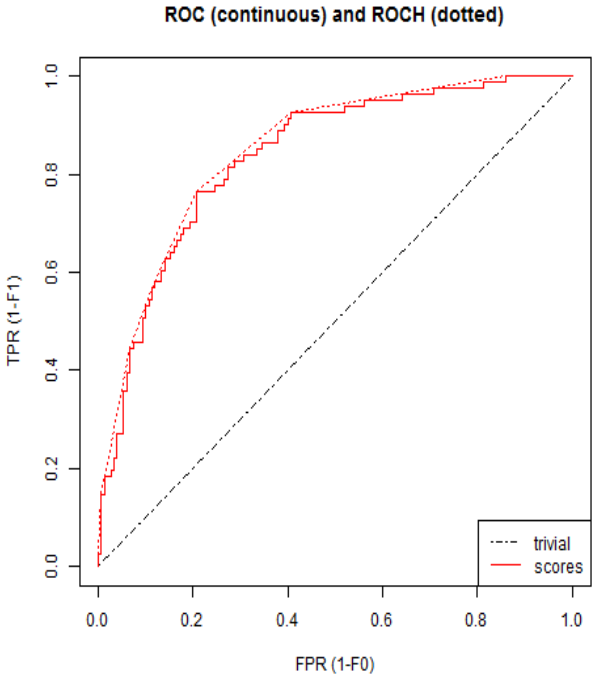

Fig. 6.6 Random Forest ROC curve

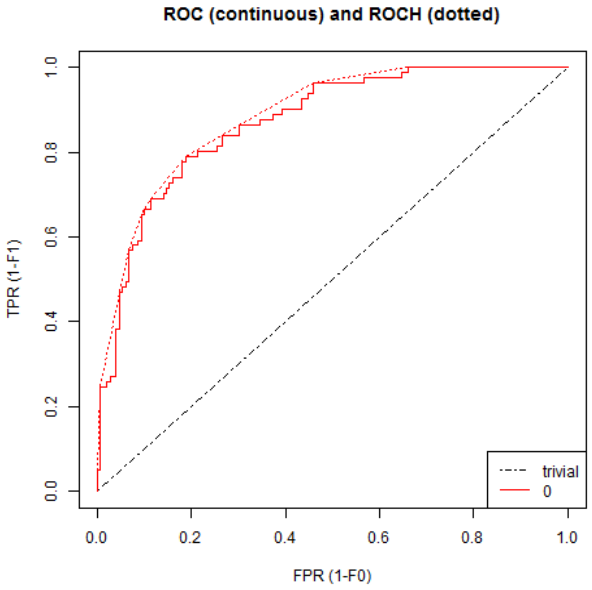

Fig. 6.7 adaboost ROC curve

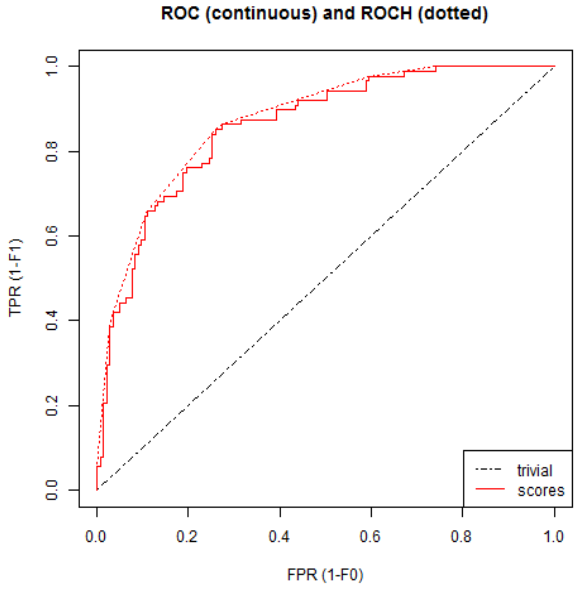

Fig. 6.8 Linear Model ROC curve

6.2 ROC Curve of Ensemble Model

In figure 6.9 shows that the ensemble gives the better results. The accuracy of the proposed ensemble model is high i.e. 84.82%.

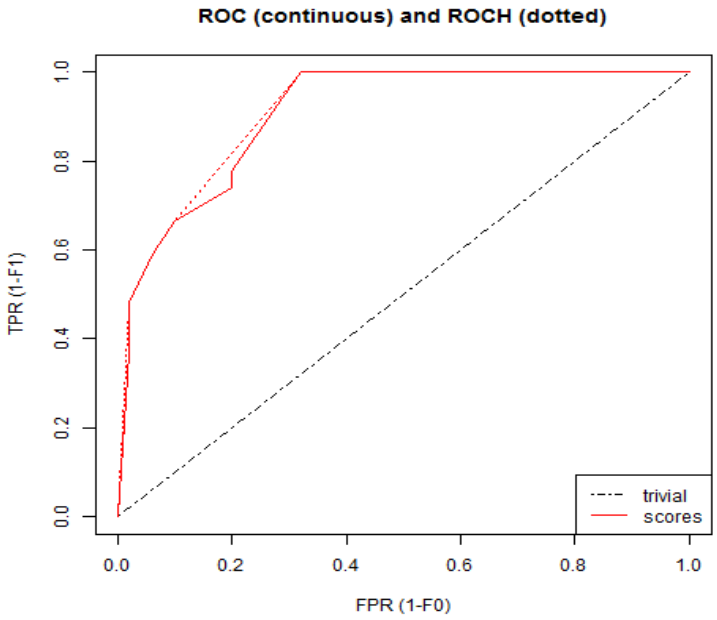

Fig. 6.9 Ensemble ROC Curve

6.3 Cross-Validation Result

10-fold cross validation is applied to improve the robustness of the data. The dataset is partitioned in to training-testing experiment are set to 90% and 10% separately. Cross Validation describes the process of splitting the whole data set into k parts and using each one of them sequentially as the test data set while combining the others to the training data. Afterwards, the performance indicators are averaged over all validation processes. The result shows that the proposed ensemble model gives better results in terms of accuracy as compared to the different machine learning model. The accuracy obtained is 85.82%.

Table 6.2 10-Fold cross validation results with 90-10 rule

Models	Sensitivity	Specificity	Precision	Recall	Accuracy
DT	0.65	0.892	0.72	0.65	79.22
K-NN	0.69	0.885	0.77	0.691	78.03
NB	0.64	0.871	0.78	0.604	80.92
NN	0.56	0.865	0.79	0.569	79.32
RF	0.68	0.908	0.75	0.689	78.56
Adaboost	0.78	0.912	0.781	0.785	81.95
SVM	0.86	0.933	0.78	0.66	82.42
Linear model	0.88	0.845	0.81	0.88	76.23
Proposed Ensemble Model	0.96	0.947	0.852	0.968	85.82

CHAPTER 7

CONCLUSION AND FUTURE WORK

Diabetes is the fast growing disease among the people even among the youngsters. Diabetes is caused by the increase level of the sugar in the blood. It is one of the major chronic non communicable diseases that affect millions globally. Diabetes is also the creator of another kind of disease mainly, the eyes, kidneys, blood vessels, nerves and the heart. In this work, we have applied different machine learning model such as decision tree, Naive Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbors and Neural Network. The experimental results shows that proposed ensemble model gives better results as compared to the single model in terms of accuracy i.e. 84.82%. 10-fold cross validation is applied to improve the robustness of the data.

Future work

Optimization techniques will be applied to the diabetes dataset for obtaining optimal results and for proper prediction more data needed to make it more accurate.

REFERENCES

[1] http://www.who.int/diabetes/en/

[2] https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/non-communicable-diseases-cause-61-of-deaths-in-india-who-report/articleshow/60761288.cms

[3] World Health Organisation. Country Profiles 2017. World Heal Organ. 2017.

[4] W. Commons, "Zwei mögliche trenngeraden mit verschiedenen randgrößen," 2010.

[5] https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

[6] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2010). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.

[7] Mercaldo, F., Nardone, V., & Santone, A. (2010). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Computer Science, 112(C), 2519-2528.

[8] J. Pradeep Kandhasamy, S. Balamurali (2011). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.

[9] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu (2011). Comparison of three data mining models for predicting diabetes or pre diabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[10] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2011, February). Risk prediction of type II diabetes based on random forest model. In Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2011 Third International Conference on (pp. 382-386). IEEE.

[11] Mercaldo, F., Nardone, V., & Santone, A. (2012). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Computer Science, 112(C), 2519-2528.

[12] Rani, A. S., & Jyothi, S. (2012, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2012 3rd International Conference on (pp. 1584-1589). IEEE.

[13] Saravananathan, K., & Velmurugan, T. (2013). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).

[14] P.Yasodha and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WekaTool", *International Journal of Scientific & Engineering Research*, vol. 2, no. 5, 2013.

[15] A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", *IJDKP*, vol. 5, no. 1, pp. 01-14, 2014.

[16] N. NiyatiGupta,A.Rawal, and V.Narasimhan,"Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", *IOSR Journal of Computer Engineering*, vol. 11, no. 5, pp. 70-73, 2014.

[17] M. Chikh,M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor," *Journal of medical systems,*vol.36, no.5, pp. 2721-2729, 2015.

[18] K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from bigdata using R,"*International Journal of Advanced Engineering Research and Science*, vol. 2, Sep 2015.

[19] GyorgyJ.Simon, Pedro J.Caraballo,Terry M. Therneau,Steven S.Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," *IEEE Transanctions on Knowledge and Data Engineering* ,vol 27,No.1,January 2015.

[20] Vikas Tiwari, T. (20l6). Design and implementation of an efficient relative model in cancer disease recognition". IJARCSSE.

[21] Khaleel, M. A. (20l6). A Survey of Data Mining Techniques on Medical Data for finding frequent diseases. IJARCSSE.

[22] Chaurasia.V, P. (20l6). Data Mining Approach to Detect Heart Disease. IJACSIT, 56-66.

[23] Parthiban.G, S. (20l6). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. IJAIS, 25-3O.

[24] Iyer.A, S. (20l7). Diagnosis of Diabetes Using Classification Mining Techniques. IJDKP, l-l4.

[25] Almir Badnjević, Lejla Gurbeta, Berina Alić (20l7). Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases. Mediterranean Conference on Embedded Computing.

[27] Baby, P. (20l7). Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms. IJERT.

[28] Shaik Razia and M.R.NarasingaRao "A Neuro computing frame work for thyroid disease diagnosis using machine learning techniques", Vol.95. No.9. Pages 1996-2005).

[29] Alehegn, Minyechil, and Rahul Joshi& Dr Preeti Mulay. "Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm." *International Journal of Pure and Applied Mathematics*, No. 9 (2018).

[30] Ali, Rahman, et al. "Prediction of diabetes mellitus based on boosting ensemble modeling." *International conference on ubiquitous computing and ambient intelligence*. Springer, Cham, 2014.

[31] Hssina, Badr, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. "A comparative study of decision tree ID3 and C4. 5." *International Journal of Advanced Computer Science and Applications* 4, no. 2 (2014).

[32] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE Transactions on Information Theory, 57(7):4689–4708, 2011.

[33] https://www.kaggle.com/pima-diabetes-dataset.