*

A Project Report on

DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology
in
Computer Science and Engineering

Submitted by

G HARSHITHA
(19H51A0509)

G VINAY
(19H51A05D4)

J VISHNUKANTH
(20H55A0511)

Under the esteemed guidance of

MR. B. SIVAIAH
(Associate Professor)

EXPLORE TO INVENT

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution under UGC & JNTUH, Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NBA.)
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2019- 2023

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase-1 report entitled **"DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS"** being submitted by **G.HARSHITHA(19H51A0509), G.VINAY(19H51A05D4), J.VISHNUKANTH(20H55A0511)** in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. B. Sivaiah
Associate Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. B. Sivaiah, Associate Professor**, Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academic, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

G. HARSHITHA	(19H51A0509)
G. VINAY	(19H51A05D4)
J. VISHNUKANTH	(20H55A0511)

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. The goal this project is to predict the Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes.

CHAPTER 1
INTRODUCTION

1. INTRODUCTION

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criterion to be taken into consideration for an individual to harbour this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage.

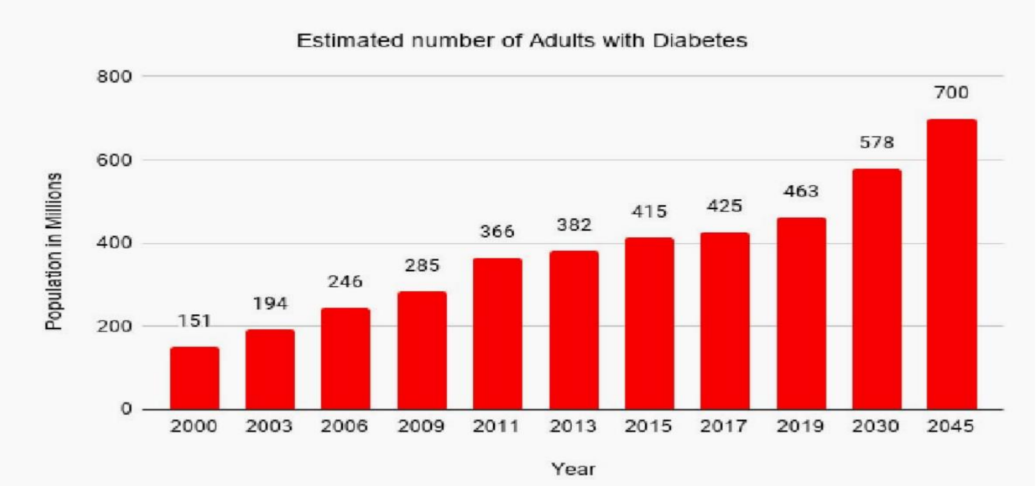


Fig.1. Number of diabetic patients estimated with respect to year

In the year 2019, approx. 463 million adults between the age of 20-79 years had diabetes (International Diabetes Federation- IDF). 79% of the adult population were living in the countries with the low and middle-income groups. It is estimated that by the year 2045 approx. 700 million people will have diabetes (IDF). Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which include unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease. In the Fig. 1 we can see that the number of cases is rising every year and there is not slowing down in the active cases. It is a very crucial thing to worry as diabetes has become one of the most dangerous and fastest diseases to take the lives of many individuals around the globe. Machine Learning is very popular these days as it is used everywhere, where a large amount of data is present, and we need some knowledge from it. Generally, we can categorise the Machine Learning algorithms in two types but not limited to-

• **Unsupervised Learning:** In unsupervised learning, the information is not labelled and also not trained. Here, we just put the data in action to find some patterns if possible.

• **Supervised Learning:** In supervised learning, we train the model based on the labels attached to the information and based on that we classify or test the new data with labels.

With the rise of Machine Learning and its relative algorithms, it has come to light that the significant problems and hindrances in its detection faced earlier, can now be eased with much simplicity, yet, giving a detailed and accurate outcome. As of the modern-day, it is comprehended that Machine Learning has become even more effective and helpful in collaboration with the domain of Medicine. Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The work which will be described in this paper is to perform the diabetes disease prediction using machine learning algorithms for early care of an individual.

1.1 PROBLEM STATEMENT:

- Diabetes is a most common disease caused by a group of metabolic disorders. It is also known as Diabetic mellitus. It affects the organs of the human body. It can be controlled by predicting this disease earlier. If diabetic’s patient is untreated for a long time, it may lead to increase blood sugar.
- Now a days, Healthcare industries generating large volume of data. Machine Learning algorithms and statistics are used to predict the disease with the help of current and past data. Machine learning techniques help the doctors to predict early stage for diabetics. Diabetes patient medical record and different types of algorithms are added in dataset for experimental analysis.

1.2 OBJECTIVES:

- This research work aims to analyze the Diabetes dataset, design, and implement a Diabetes prediction and recommendation system utilizing machine learning classification techniques. The specific objectives of this project work are:

1. To review existing literature along the area of diabetes diagnosis and prediction.
2. Design and develop a model using machine learning techniques.
3. To analyse the Diabetes dataset and use various machine learning algorithms to develop a prediction model.
4. To identify and discuss the benefits of the designed system along with effective applications.

1.3. PROJECT SCOPE AND LIMITATIONS:

- Diabetes prediction is one of the most important tasks in today's world, especially in light of its serious complications. Diabetes is the leading cause of death in the world. The system model is primarily focused on detecting diabetes using a few parameters.
- Physicians can utilize the system to anticipate diabetes in the early stages. So that patients can receive traditional treatments and remedies. For the prediction, the system employed some approaches such as machine learning (ML) in order to provide more precise findings. For hospitals and clinicians, developing a diabetes illness prediction system is beneficial.
- The system predicts disease at an early stage, allowing clinicians to better treat patients. The proposed model is a real-time application that can be used in various hospitals and predicts disease in a shorter amount of time. We will receive more accurate and efficient findings when we employ machine learning algorithms for disease prediction.

CHAPTER 2

BACKGROUND WORK

2. BACKGROUND WORK

2.1 An Intelligent System for Diabetes Prediction

2.1.1 Introduction:

AUTHORS: Z. Tafa, N. Pervetica, and B. Karahoda "An Intelligent System for Diabetes Prediction" presents a joint Matlab implementation of the SVM and Naïve Bayes methods in a new dataset acquired from the patients examined for diabetes in Kosovo. The developed diagnostic tool enables the intelligent computer-based prediction on diabetes, based on the previously acquired values. The statistical analysis shows the high accuracy of data classification. Also, the proposed joint implementation of two algorithms aims to improve the reliability of the decision by using the power of both algorithms in minimizing their individual weakness.

2.1.2 Merits and Demerits:

- Ensemble of naive bayes and support vector machine has achieved the accuracy of 97.6%.
- When run alone on the dataset naive bayes achieving an accuracy of 94.52 and support vector machine achieving 95.52%.
- The authors have not mentioned any preprocessing technique to filter out any unwanted values from the dataset.

2.1.3 Implementation of An Intelligent System for Diabetes Prediction:

Dataset description:

Dataset consists of 402 instances taken from three different locations in Kosovo. During the data acquisition process, the appropriate importance is given to the patient's data privacy and anonymity. The attributes of the database are: BMI (Body Mass Index), glucose level before meal and after meal, the systolic and diastolic blood pressure, the hereditarily factor, the regular diet, and daily physical activities. The last two attributes are evaluated as follows. Regarding the issue of regular diet, while relying on inputs from the medical clinicians, patients were asked if they took their meals in approximately same equidistant daily intervals at least three times a day and also if their meals were not voluminous. With these answers being positive, we consider that a patient is having the regular diet. On the other hand, according to the U.S. Center for Disease Control and Prevention (CDCP), the adult person is considered to be physically active if he/she conducts the 150-200 min of physical activities a week. With the answers on the family history questions, and in accordance to the above given thresholds, the answers of the examinee regarding the last three questions are mapped into two values:1 and 0.

ATTRIBUTE	Value Range	
	From	To
BMI	15	40
Pre meal glucose	3.5	19
Post meal glucose	4.9	22.8
Diastolic blood pressure	55	110
Systolic blood pressure	90	200
Family history of diabetes	No(0)	Yes(1)
Regular diet	No(0)	Yes(1)
Physical Activites	No(0)	Yes(1)

Table.2.1.3.1. The ranges of the values of all attributes are given in Table.

Support Vector Machine(SVM):

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane .The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

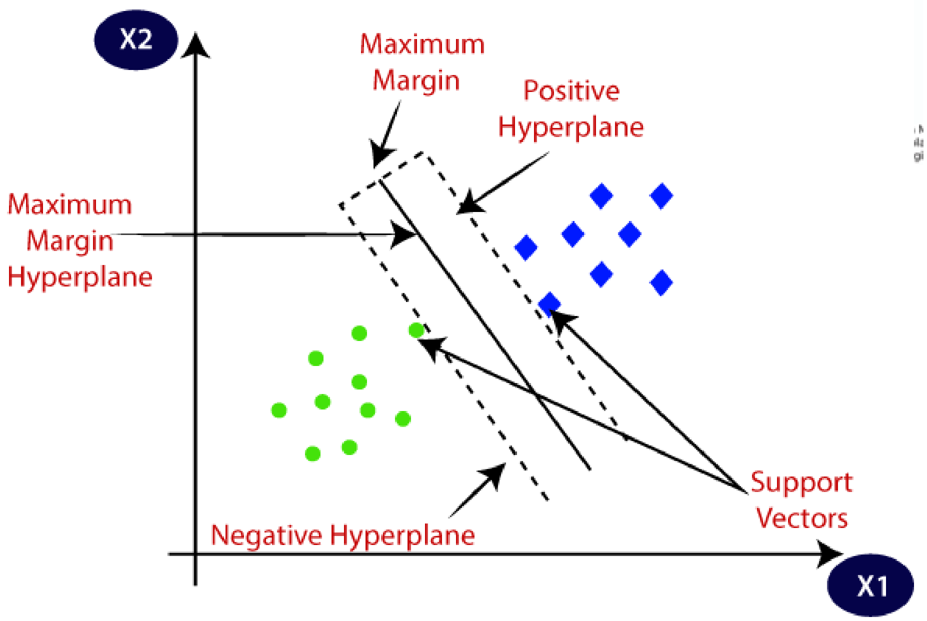


Fig.2.1.3.2.Support Vector Machine Diagram

Naive Bayes:

The Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It is mainly used in *text* classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions**.** It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are Spam filtration, Sentimental analysis, and classifying articles**.**

Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

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

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

The Naïve Bayes working can be explained on the basis of the below algorithm:

STEP 1: Convert the given dataset into frequency tables.

STEP 2: Generate a Likelihood table by finding the probabilities of given features.

STEP 3: Now, use the Bayes theorem to calculate the posterior probability.

Gaussian Naive Bayes:

Real-valued attributes are estimated by assuming a Gaussian distribution. Easiest to work with, only need mean and std from training data and calculate mean and std of input values(X) for each class to summarize the distr.

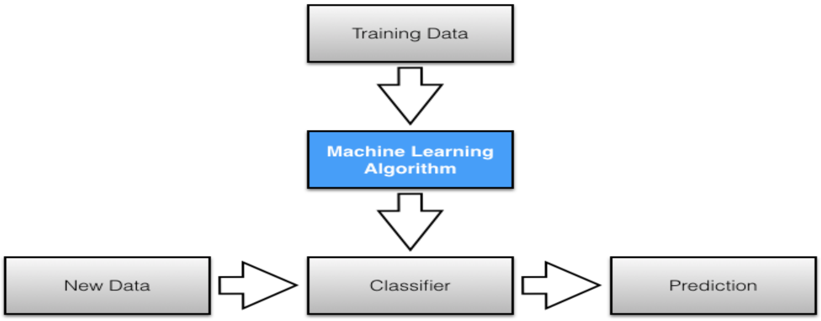


Fig.2.1.3.3.Naive Bayes Model Diagram

2.2. Prediction of Diabetes using Classification Algorithms.

2.2.1. Introduction:

- In this model the dataset is collected from Pima Indian Diabetes Dataset.
- The model was proposed using a combination of Naive Bayes, Decision Tree and support vector machine algorithms for diabetes prediction.
- Eight attributes were present inside the dataset, and it consisted of 768 patient's data.
- When run alone on the dataset, that is,

  Naïve Bayes achieving an accuracy of 76.30%

  Support vector machine achieving 65.10%

  Decision Tree 73.82%.

2.2.2. Merits and Demerits:

- Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm.

- They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

2.2.3. Implementation of Prediction of Diabetes using Classification Algorithms:

Naive Bayes Classifier:

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with unbalancing problems and missing values. Naive Bayes [24] is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability $P(C|X)$ can be calculated from P(C),P(X) and $P(X|C)$ [23].

Therefore, $P(C|X) = (P(X|C)\ \mathrm{P(C)})/P(X)$

Where, $P(C|X)$ = target class's posterior probability.

$P(X|C)$ = predictor class's probability.

P(C) = class C's probability being true.

P(X) = predictor's prior probability. The evaluated performance of Naive Bayes algorithm using Confusion Matrix is as follows:

	A	**B**
A-Tested Negative	422	78
B-Tested Positive	104	164

Table.2.2.3.1. Confusion Matrix of Naive Bayes

Decision Tree Classifier:

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes. The evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

	A	**B**
A-Tested Negative	407	93
B-Tested Positive	108	164

Table 2.2.3.2. Confusion Matrix of Decision Tree

Support Vector Machine (SVM):

SVM is one of the standard set of supervised machine learning model employed in classification. Given a two class training sample the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes The Accuracy of the experiment is evaluated using WEKA interface. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by $w^T\ x + b = -1$ and the hyperplane defined by $w^T\ x + b = 1$

This distance is equal to $\frac{2}{||w||}$. This means we want to solve max $\frac{2}{||w||}$ Equivalently we want min $\frac{||w||}{2}$. The SVM should also correctly classify all x(i), which means $y^i(w^T x^i + b) >= 1, \forall i \in \{1,$ ¢¢, $N\}$. The evaluated performance of SVM algorithm for prediction of Diabetes [16], [30] using Confusion Matrix is as follows:

	A	**B**
A-Tested Negative	500	0
B-Tested Positive	268	0

Table.2.2.3.3. Confusion Matrix of SVM

Model Diagram:

Proposed procedure is summarized in figure-1 below in the form of model diagram. The figure shows the flow of the research conducted in constructing the model.

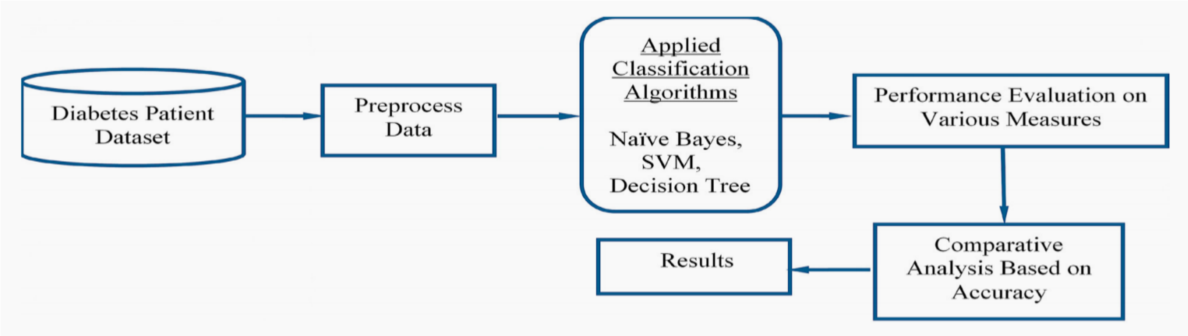

Diabetes Patient Dataset
Preprocess Data
Applied Classification Algorithms
Naïve Bayes, SVM, Decision Tree
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Results

Fig.2.2.3.4.Proposed Model Diagram

Dataset Used:

In this work WEKA tool is used for performing the experiment. WEKA is a software which is designed in the country New Zealand by University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements. The main aim of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database PIDD. Table-4 shows a brief description of the dataset.

Database	No. of Atributes	No. of Instances
PIDD	8	768

Table.2.2.3.5.Dataset Description

PIDD-Pima Indians Diabetes Dataset:

The proposed methodology is evaluated on Diabetes Dataset namely (PIDD), which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class ’0’ treated as tested negative for diabetes and value of another class ’1’ is treated as tested positive for diabetes.

Accuracy Measures: Naive Bayes, SVM and Decision Tree algorithms are used in this research work.

S.No	Attribute	Abbreviation of Attributes
1	Number of times pregnant	Pr
2	Plasma glucose concentration	Pl
3	Diastolic blood pressure (mm Hg)	Pr
4	Skin fold thickness (mm)	Sk
5	2-Hour serum insulin (mu U/ml)	In
6	BMI (weight in kg/(height in m)^2)	Ma
7	Diabetes pedigree function	Pe
8	Age in years	Ag
9	Class ’0’ or ’1’	cl

Table.2.2.3.6.Accuracy Description:

S.No	Measures	Definitions	Formula
1	Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2	Precision (P)	Classifiers correctness/accuracy is measured by Precision.	P = TP / (TP+ FP)
3	Recall (R)	To measure the classifiers completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4	F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)
5	ROC	ROC(Receiver Operating Curve) curves are used to compare the usefulness of tests.	

Table.2.2.3.7.Accuracy Measures

Classification Algorithms	Precision	Recall	F-Measure	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Tree	0.735	0.738	0.736	73.82	0.751

Table.2.2.3.8.Comparative Performance of Classification Algorithms on Various Measures

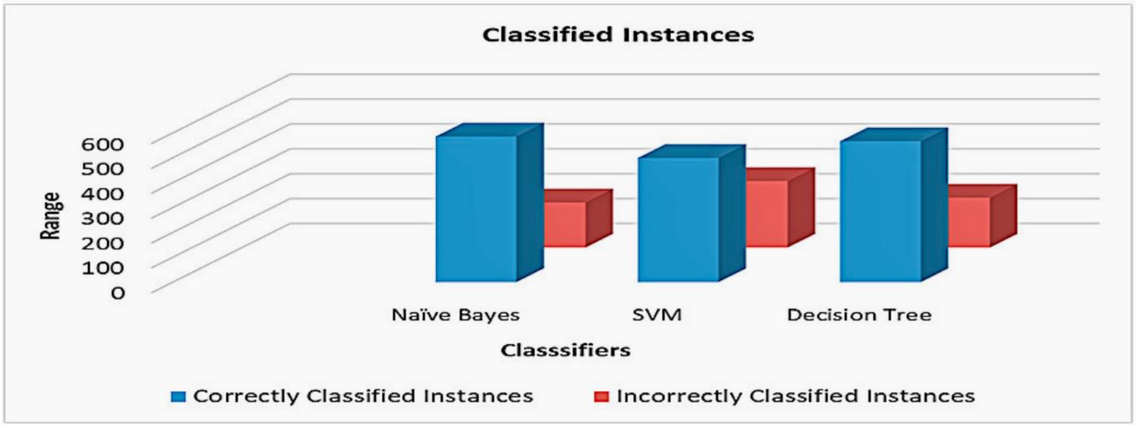


Fig.2.2.3.9.Classified Instances

2.3 Diabetes disease prediction using machine learning on big data of healthcare

2.3.1 Introduction:

Healthcare domain is a very prominent research field with rapid technological advancement and increasing data day by day. In order to deal with large volume of healthcare data we need Big Data Analytics which is an emerging approach in Healthcare domain. Millions of patients seek treatments around the globe with various procedure. Analyzing the trends in treatment of patients for diagnosis of a particular disease will help in making informed and efficient decisions to improve the overall quality of healthcare. Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. This paper aims at building a classifier model using WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. The research hopes to recommend the best algorithm based on efficient performance result for the prediction of diabetes disease. Experimental results of each algorithm used on the dataset was evaluated. It is observed that Support Vector Machine performed best in prediction of the disease having maximum accuracy.

2.3.2 Merits and demerits:

- The experimental has performed by using Naive Bayes, Support Vector Machine, Random Forest and Simple CART classifiers on the diabetes patient dataset.
- Selected the best algorithm after performing the performance evaluation of all the four classifies.
- WEKA is only suitable for small datasets.

2.3.3 Implementation of **Diabetes disease prediction using machine learning on big data of healthcare:**

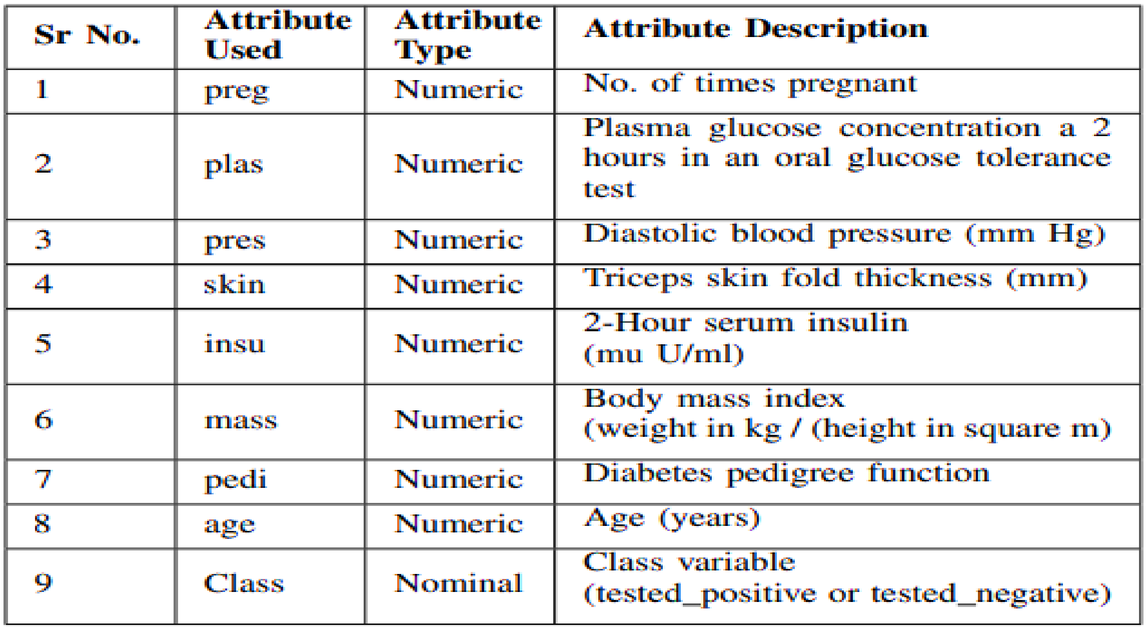
WEKA Tool Description**:** WEKA [Waikata Enviroment for Knowledge Analysis]

- It is a very popular machine learning and data mining toolkit for conducting data driven researches.
- Developed in New Zealand at the University of Waikato
- The collection of machine learning and data mining algorithms present are written in Java
- The version of WEKA used for experimentation in this paper is WEKA Version 3.82 the

research made use of WEKA tool as it helps in performance evaluation and performing comparison of various machine learning techniques conveniently on real time data.

Diabetes Disease Dataset:

The name of the dataset that has been considered is Pima Indians Diabetes Database which is collected from National Institute of Diabetes and Digestive and Kidney Diseases. The total No. of Instances are 768 and the size is 37 KB. The total no. of attributes are 9 including the target class attribute. The name of two target classes are tested positve and tested negative. The no. of instances for tested positive are 268 and the no. of instances for tested negative are 500. The data pre-processing is automatically performed by WEKA tool.



Sr No.	Attribute Used	Attribute Type	Attribute Description
1	preg	Numeric	No. of times pregnant
2	plas	Numeric	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	pres	Numeric	Diastolic blood pressure (mm Hg)
4	skin	Numeric	Triceps skin fold thickness (mm)
5	insu	Numeric	2-Hour serum insulin (mu U/ml)
6	mass	Numeric	Body mass index (weight in kg / (height in square m)
7	pedi	Numeric	Diabetes pedigree function
8	age	Numeric	Age (years)
9	Class	Nominal	Class variable (tested_positive or tested_negative)

Table.2.3.3.1. Dataset Attributes

In this model specifically considered diabetes disease and takes input of the dataset is processed using four machine learning algorithms that are Naive Bayes, SVM, Random Forest, Simple CART and for each algorithm respective classifier model is trained and tested and the results are gathered. Based on the experimental results the best performing algorithm can be determined which will help in accurate prediction of the disease.

Flow chart:

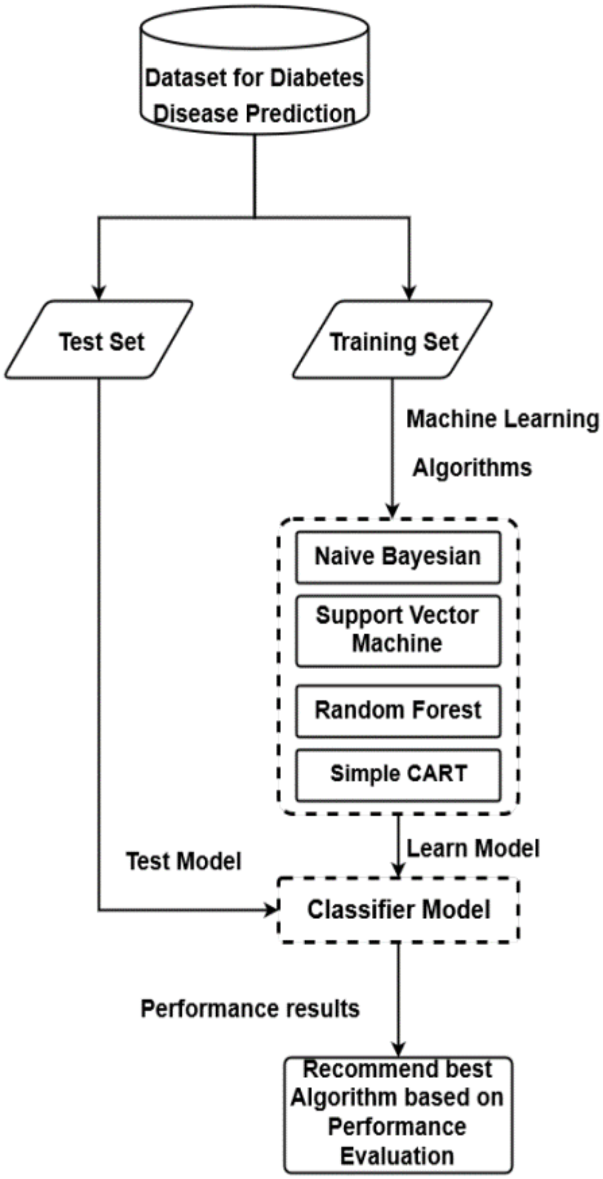


Fig.2.3.3.2.Flow chart

Procedure:

- Step 1 :- Preprocess the input dataset for diabetes disease in WEKA tool.
- Step 2 :- Perform percentage split of 70% to divide dataset as Training set and Test set .
- Step 3 :- Select the machine learning algorithm i.e. Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm.
- Step 4 :- Build the classifier model for the mentioned machine learning algorithm based on

training set.

- Step 5 :- Test the Classifier model for the mentioned machine learning algorithm based on test set.
- Step 6 :- Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
- Step 7 :- After analyzing based on various measure conclude the best performing algorithm.

Performance metrics:

Classification Accuracy Results:

The following table represents the experimental classification accuracy results of Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. The table displays the Training time, testing time and Accuracy value of each algorithm.

Algorithm	Training Time	Testing Time	Accuracy Value
Naïve Bayes	0.03 sec	0.02 sec	0.77
Support Vector Machine	0.14 sec	0.03 sec	0.7913
Random Forest	0.67 sec	0.06 sec	0.765
Simple Cart	1.38 sec	0.02 sec	0.765

Table.2.3.3.3.Performance metrics

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure
Naïve Bayes	0.770	0.317	0.767	0.770	0.768
Support Vector Machine	0.791	0.345	0.784	0.791	0.782
Random Forest	0.765	0.326	0.756	0.765	0.758
Simple Cart	0.765	0.364	0.762	0.763	0.446

Table.2.3.3.4.Major Accuracy Measure Values

CHAPTER 3

RESULTS AND DISCUSSION

RESULTS AND DISCUSSION

3.1. Comparison of Existing Solutions:

S.No	Proposed by	No of datasets used	Name of the dataset	Data size	Classifier used	No of features used	Classification accuracy	Published year
1	Z. Tafa, N. Pervetica, and B. Karahoda	3	Collected from Kosova	402	Naive Bayes, SVM	8	SVM=95.52%, Naive Bayes=94.53%	2015
2	Deepti Sisodiaa, Dilip Singh Sisodia	1	UCI repository	768	Decision tree	8	SVM=65.10% Naive Bayes= 76.30% Decision tree=73.82%	2018
3	A. Mir and S. N. Dhage	1	UCI repository	768	Naive Bayes, SVM, Random Forest and Simple CART.	9	Not Measured	2018

Table.3.1.1. Comparison of Existing Solutions

3.2. Data Collection and Performance metrics:

Data Collection and Performance metrics of An Intelligent System for Diabetes Prediction:

ATTRIBUTE	Value Range	
	From	To
BMI	15	40
Pre meal glucose	3.5	19
Post meal glucose	4.9	22.8
Diastolic blood pressure	55	110
Systolic blood pressure	90	200
Family history of diabetes	No(0)	Yes(1)
Regular diet	No(0)	Yes(1)
Physical Activites	No(0)	Yes(1)

Table.3.2.1. Data Collection of An Intelligent System for Diabetes Prediction

PERFORMANCE METRICS:

Algorithm	No. of records	Train set/Test set	No. of correctly classified	No. of incorrectly classified records	Classifier performance (mean value)
SVM	402	201/201	192	9	95.52%
Naïve Bayes	402	201/201	190	11	94.53%

Table.3.2.2. Classification and the average accuracy

Algorithm	Precision (Class NO)	Recall (Class NO)	Precision (Class YES)	Precision (Class YES)
SVM	0.97	0.975	0.892	0.868
Naïve Bayes	0.981	0.951	0.814	0.921

Table.3.2.3.Precision and Recall

Data Collection and Performance metrics of Prediction of Diabetes using Classification Algorithms

S.No	Attribute	Abbreviation of Attributes
1	Number of times pregnant	Pr
2	Plasma glucose concentration	Pl
3	Diastolic blood pressure (mm Hg)	Pr
4	Skin fold thickness (mm)	Sk
5	2-Hour serum insulin (mu U/ml)	In
6	BMI (weight in kg/(height in m)^2)	Ma
7	Diabetes pedigree function	Pe
8	Age in years	Ag
9	Class ’0’ or ’1’	cl

Table.3.2.4. Data Collection of Prediction of Diabetes using Classification Algorithms

PERFORMANCE METRICS:

Classification Algorithms	Precision	Recall	F-Measure	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Tree	0.735	0.738	0.736	73.82	0.751

Table.3.2.5.Comparative Performance of Classification Algorithms on Various Measures

Total no of instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
768	Naive Bayes	586	182
	SVM	500	268
	Decision Tree	567	201

Table.3.2.6.Classifier’s Performance on The Basis of Classified Instances

Data Collection and Performance metrics of Diabetes disease prediction using machine learning on big data of healthcare.

S.No	Attribute	Abbreviation of Attributes
1	Number of times pregnant	Pr
2	Plasma glucose concentration	Pl
3	Diastolic blood pressure (mm Hg)	Pr
4	Skin fold thickness (mm)	Sk
5	2-Hour serum insulin (mu U/ml)	In
6	BMI (weight in kg/(height in m)^2)	Ma
7	Diabetes pedigree function	Pe
8	Age in years	Ag
9	Class ’0’ or ’1’	cl

Table.3.2.7. Data Collection of Diabetes disease prediction using machine learning on big data of healthcare.

PERFORMANCE METRICS:

Algorithm	Training Time	Testing Time	Accuracy Value
Naïve Bayes	0.03 sec	0.02 sec	0.77
Support Vector Machine	0.14 sec	0.03 sec	0.7913
Random Forest	0.67 sec	0.06 sec	0.765
Simple Cart	1.38 sec	0.02 sec	0.765

Table.3.2.8.Classification Accuracy results

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure
Naïve Bayes	0.770	0.317	0.767	0.770	0.768
Support Vector Machine	0.791	0.345	0.784	0.791	0.782
Random Forest	0.765	0.326	0.756	0.765	0.758
Simple Cart	0.765	0.364	0.762	0.763	0.446

Table.3.2.9.Major Accuracy measure values

CHAPTER 4
CONCLUSION

4.1 CONCLUSION

Machine learning can help doctors identify and cure diabetes. In this report we have discussed existing solutions for prediction of diabetes. We will conclude that improving the accuracy of the classification will help the machine learning models perform better. The performance analysis is in terms of accuracy rate among all the classification techniques such as K-nearest neighbors, Naive Bayes, Decision Tree Classifier, Support Vector Machine, random forest. Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. The main aim is to design and implement diabetes prediction using machine learning methods and performance analysis of that method. The proposed method approach uses SVM, KNN, logistic regression, and random forest. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

CHAPTER 5
REFERENCES

REFERENCES

[1]. Z. Tafa, N. Pervetica, and B. Karahoda, “An Intelligent System for Diabetes Prediction,” IEEE Explore, in Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382, Budva, Montenegro, June 2015.

[2]. D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science.[Online].
Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

[3]. Mir and S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.

[4]. P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in *2018 4th International Conference on Computing Communication and Automation (ICCCA),* 2018, pp. 1–4.

[5]. Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference*, St´efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.

[6]. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R’ıo, M. Wiebe, P. Peterson, P. G’erard-Marchant, K.Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[7]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and ´ Edouard Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, p. 28252830, 2011.

[8]. Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference*, St’efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.