HEALTH MONITORING SYSTEM: A PROJECT REPORT ON DIABETES PREDICTION

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF COMPUTER APPLICATION

by

Sakshi Tyagi
(200029014005795)

Shivani Chauhan
(200029014005799)

Mukul Dhama
(200029014005758)

Under the Supervision of

Mr. Rabi N. Panda
KIET Group of Institutions Ghaziabad

DR. APJ ABDUL KALAM TECHNICAL UNIVERSITY

(Formerly Uttar Pradesh Technical University) LUCKNOW

June, 2022

DECLARATION

We hereby declare that the work presented in this report entitled "Health Monitoring System", was carried out by Sakshi Tyagi. We have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute.

We have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. We have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name : Sakshi Tyagi
Enrollment No. : (200029014005795)
Course : Master of Computer Application

(Candidate Signature)

DECLARATION

We hereby declare that the work presented in this report entitled "Health Monitoring System", was carried out by Shivani Chauhan. We have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute.

We have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. We have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name : Shivani Chauhan
Enrollment No. : (200029014005799)
Course : Master of Computer Application

(Candidate Signature)

DECLARATION

We hereby declare that the work presented in this report entitled "Health Monitoring System", was carried out by Mukul Dhama. We have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute.

We have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. We have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name : Mukul Dhama
Enrollment No. : (200029014005758)
Course : Master of Computer Application

(Candidate Signature)

CERTIFICATE

Certified that **Sakshi Tyagi (200029014005795)** has carried out the project work presented in this report entitled "**Health Monitoring System**" for the award of **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University, Lucknow under my supervision. The report embodies result of original work, and studies are carried out by the student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University.

Mr. Rabi N. Panda **External Examiner**
Associate Professor
Dept. of Computer Applications
KIET Group of Institutions, Ghaziabad

Dr. Ajay Kumar Shrivastava
Professor &
Head Department of
Computer Applications
KIET Group of
Institutions, Ghaziabad

Date:

CERTIFICATE

Certified that **Shivani Chauhan (200029014005799)** has carried out the project work presented in this report entitled "**Health Monitoring System**" for the award of **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University, Lucknow under my supervision. The report embodies result of original work, and studies are carried out by the student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University.

Mr. Rabi N. Panda **External Examiner**
Associate Professor
Dept. of Computer Applications
KIET Group of Institutions, Ghaziabad

Dr. Ajay Kumar Shrivastava
Professor &
Head Department of
Computer Applications
KIET Group of
Institutions, Ghaziabad

Date:

CERTIFICATE

Certified that **Mukul Dhama (200029014005758)** has carried out the project work presented in this report entitled "**Health Monitoring System**" for the award of **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University, Lucknow under my supervision. The report embodies result of original work, and studies are carried out by the student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University.

Mr. Rabi N. Panda **External Examiner**
Associate Professor
Dept. of Computer Applications
KIET Group of Institutions, Ghaziabad

Dr. Ajay Kumar Shrivastava
Professor &
Head Department of
Computer Applications
KIET Group of
Institutions, Ghaziabad

Date:

ABSTRACT

The world's most deadly disease, Diabetes Mellitus, is among the fastest growing. Medical professionals are looking for a technology that can predict diabetes properly. To study data from diverse perspectives and synthesise it into useful knowledge, various machine learning algorithms can be applied. With the application of appropriate data mining techniques to large amounts of data, we will be able to obtain valuable information. The main objective is to find new patterns and then provide relevant and helpful information to consumers by analysing them. Diabetes is linked to cardiovascular disease, renal problems, nerve damage, and blindness. Data mining is critical to deal with while doing diabetes research. Methods and strategies for efficiently classifying and detecting patterns in the Diabetes dataset will be discovered using data mining techniques and methods. The purpose of this study was to predict diabetes by using medical bioinformatics. The WEKA programme was used as a diagnostic mining tool. The University of California at Irvine provided the Pima Indian diabetes database, which was utilised in the research. A model based on the data was developed for accurately predicting and diagnosing diabetes. This study compares the performance of Nave Bayes, Decision Trees, and (KNN) with bootstrapping resampling to improve accuracy.

Diabetes has spread throughout the world, affecting people of all ages, regardless of their age. As diabetes patients increase, it is due to numerous factors such as bacterial or viral infections, chemical or toxic substances in food, auto immune reactions, obesity, poor diet, lifestyle changes, eating habits, pollution, etc. To save a person's life, it is essential to identify diabetes in a timely manner. Data analytics is the process of analysing and identifying hidden patterns in large amounts of data to draw conclusions. This analytical procedure is carried out in health care by employing machine learning algorithms to analyse medical data and develop machine learning models for medical diagnostics. This study explains how to diagnose diabetes using a diabetes prediction algorithm. Furthermore, this research investigates several machine learning algorithms and strategies for improving diabetes prediction accuracy utilising medical data. Diabetes Mellitus (DM) is a group of metabolic disorders that afflict millions of people worldwide. Diabetes must be detected early in order to prevent serious complications. There have been numerous research studies on diabetes diagnosis, the majority of them rely on a single data source, the Pima Indian diabetes data set. A collection of studies on Indian women dating back to 1965 has been collected as part of the Pima Indian data set, and it has a relatively high rate of diabetes onset. The majority of prior research studies focused exclusively on one or two specialised sophisticated procedures for data testing, with no comprehensive research on multiple generic techniques. This study explores a number of Machine Learning techniques (e.g., the KNN algorithm) for the purpose of identifying diabetes and methods used for data pre-processing. Testing the accuracy of this technique will utilize the UCI ML repository data set.

ACKNOWLEDGEMENT

I take this occasion to thank God, almighty for blessing us with his grace and taking our endeavor to a successful culmination. I extend my sincere and heartfelt thanks to our esteemed guide, **Mr. Rabi N. Panda**, for providing me with the right guidance and advice at the crucial junctures and for showing me the right way. I extend my sincere thanks to our respected **Head of the Department Dr. Ajay Kumar Shrivastava**, for allowing us to use the facilities available. I would like to thank the other faculty members also, on this occasion. Last but not the least, I would like to thank my friends and family for the support and encouragement they have given me during our work.

Sakshi Tyagi
Roll No. 2000290140110

Shivani Chauhan
Roll No. 2000290140114

Mukul Dhama
Roll No. 2000290140073

LIST OF FIGURES:

LIST OF TABLES:

CONTENTS

1. Introduction

1.1 INTRODUCTION

Diabetes is characterised by elevated blood sugar levels, which manifest as frequent urination thirst, appetite, and weight loss. Diabetes is diagnosed when blood glucose levels exceed 200 mg/dL two hours after loading, as well as numerous diabetes research studies that necessitate fast call detection. Patients with diabetes typically require ongoing therapy; otherwise, they risk a variety of life-threatening complications.

Early detection and treatment of diabetes can help you avoid complications. For diabetes detection, the suggested method employs machine learning algorithms. The suggested system will be a medical field application that will aid diabetic doctors and patients in recognising diabetes. The suggested approach automates the detection of diabetes using data from previous diabetes patients. People of all ages are affected by diabetes, which is a rapidly spreading disease [1]. The presence of too much sugar (glucose) in the blood causes diabetes. The two types of diabetes are type 1 and type 2. Diabetes type 1 is an autoimmune disease. In this situation, the body destroys the cells required to create insulin and absorb sugar for energy production. Obesity has no bearing on this type of cancer. A person's body mass index defines the obesity which is above his/her normal threshold [2]. Type 1 diabetes can strike at any age, including infancy and puberty. Type 2 diabetes is more likely to develop in obese individuals. The body either opposes or fails to manufacture insulin in this condition. Type 2 is more common in middle-aged and older people [1]. Other causes of diabetes include bacterial or viral infections, toxic or chemical components in food, auto immunity reactions, obesity, inadequate diets, lifestyle changes, exposure to pollution, etc. Diabetes causes a number of ailments, including cardiovascular difficulties, renal problems, retinopathy, and foot ulcers [1]. All over the world, people are affected by diabetes, which is a serious condition. Worldwide, chronic illnesses like this lead to many deaths among adults. Chronic illnesses increase costs as well. A large percentage of government and individual budgets is spent on chronic diseases [3,4]. Diabetes affected 382 million people globally in 2013 [5]. In 2012, it was the eighth biggest cause of mortality for both men and women. Diabetes is more likely to occur in high-income countries [6].

Diabetes diagnosis is regarded as a difficult subject for quantitative research. Due to constraints, the use of numerous criteria such A1c [7], fructosamine, white blood cell count, fibrinogen, and haematological indices [8] are considered ineffective. These criteria were employed in various research investigations to diagnose diabetes [[9], [10], and [11]. Chronic ingestion of booze, salicylates, and drugs have all been linked to an increase in A1C. When measured by electrophoresis, vitamin C consumption may raise A1c levels, but when measured by chromatography, levels may appear to decrease [12].

Available clinical records show that type 1 diabetes is a notable clinical problem worldwide. There are about 2.6 million adults over the age of 18 suffering from diabetes, and the severity of diabetes should increase in Malaysia. Ketones are artificial substances that appear in the body when the muscle-to-fat ratio is used, as opposed to glucose due to urgency. This indicates that the body's cells cannot absorb enough sugar (glucose) from the blood, especially if the body's insulin is too low. Insulin is used by the body to use glucose, which focuses on what is important. In this sense, which screens ketone zookeepers one by one, helps control and screen the condition of diabetics with a vast number of ketones that remain wild in diabetics. Anyway, the two steps were considered

intrusive, horrifying, and anomalous. CH3) 2CO is abstractly recognized by the method of diabetes biomarkers. CH3) 2CO stands for traditional stuff and knows the technology for Ketobetix and the fragrant scent of breathing in diabetics. The combination of breath CH3) 2CO is associated with glucose absorption and lipolysis. This is the method by which attachment to CH3) 2CO respiration is presented in the elevated modern form of diabetes and can be used to study the development of diabetes is called the easiest way and diabetes. Prevents rapid detection of sex ketoacidosis Type 1 diabetes. A method of checking ketone levels is performed using breath estimation. Atem presents a simple portable. Today, there are several chronic illnesses such as heart disease, stroke, cancer, chronic respiratory illness, and diabetes. It is a dangerous disease and has recently become one of the leading causes of death worldwide and requires careful monitoring to maintain patient health. The biggest challenge for diabetics is to raise or lower blood glucose within a specific interval, as diabetes is caused by insulin resistance and inadequate insulin production can lead to level up or level down of blood glucose. If these conditions are not met, some patients may need urgent medical care to avoid exacerbations [1] human management for checking diabetes levels. The strategy presented demonstrated advances in the hardware relationship with the Internet of Things (IoT) system to enhance patient evaluation and individual observation methods. The Arduino board will be used to study the sensor with breath detection capability. Breath observation levels are recorded on the system using distant correspondence. Data collection is linked to the site page. Ketone levels are assessed as the percentage of CH3) 2CO in exhaled breath accumulated when the patient breathes into a mouthpiece containing a gas sensor. This assessment is based on detecting the patient's blood glucose level by separating CH3) 2CO levels from respiration at and sending data to clinical authorities by the WIFI method via a message. These devices are a new way of continuous monitoring. They provide real-time information about blood glucose levels. This article introduces an intelligent diabetes monitoring system using the node MCU and machine learning algorithms. The MCU node is connected to a glucose meter and periodically records the blood glucose levels of diabetics. This collected data can be used by caregivers (patients, researchers, and doctors) to remotely monitor patients. As a result, patients and physicians alike need to process multiple records, interpret vast amounts of data, adjust insulin doses, and bring blood glucose levels as close to normal as possible.

Diabetes is a life-threatening disease that has no cure. If you get this sickness once, it will be with you for the rest of your life. At the same time, having too much glucose in your blood might cause health problems. Kidney illness, heart disease, stroke, vision problems, dental problems, foot problems, and nerve damage are just a few examples. so that you can keep track of your diabetes and avoid complications

Kind 1 diabetes is the most well-known type of diabetes.

Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type

Type 1 diabetes occurs when the body is unable to manufacture insulin. It has a negative impact on children and young adults. It can also affect people of any age. People with this kind of diabetes must take insulin daily.

Type 2 diabetes occurs when the body is unable to manufacture or utilise insulin.

This kind of diabetes primarily affects people in their forties and fifties.

Diabetes During Pregnancy

This kind of diabetes primarily affects women. During pregnancy, this kind of diabetes develops. High blood sugar levels caused by gestational diabetes can harm your pregnancy and your baby's health.

1.2 TOOLS AND TECHNIQUES

1.2.1 SUPPORT VECTOR MACHINE(SVM): SVM stands for supervised knowledge representation model. Furthermore, it is the associated algorithm model that is used to study patient data for regression and classifications.

A Support Vector Machine creates a hyperplane or group of hyperplanes in high layered space, maps all of the models in a guide, and divides the examples by an identifiable hole about as broad as could be expected, with each side presenting one class. We should adjust the regularisation boundary C, which determines the model's complexity, in this strategy.

1.2.2 DECISION TREE: A decision tree is a mechanism that iteratively divides a given dataset into at least two examples. The technique's goal is to predict the class worth of the objective variable. The decision tree will assist in separating the informative collection and constructing the choice model in order to predict the obscure class marks. A decision tree can be built for both parallel and continuous factors. The root hub is ideally observed by a decision tree based on the most significant entropy value.

1.2.3 NAÏVE BAYES: Prior research has shown that Induction algorithms based on Naive-Bayes have shown surprising accuracy in many classification tasks even when the assumption of conditional independence on which they are founded is broken. The majority of studies, however, were conducted on small databases. We demonstrate this in certain larger databases, Naive-Bayes accuracy does not scale up as well as decision trees.

1.2.4 k NEAREST NEIGHBOR (kNN) Algorithm: The k-nearest neighbor algorithm (k-NN) is an important approach to classifying objects based on the nearest training data in the feature space. Although it is the simplest of all machine learning rules, the accuracy of the K-NN rule is harmed by the availability of screaming choices.

Effects of diabetes

Diabetes has an impact on different parts of the body which incorporates pancreas glitch, risk of heart ailments, hypertension, kidney disappointments, pancreatic issues,

nerve harm, foot issues, ketoacidosis, visual unsettling influences, and other eye issues, waterfalls and glaucoma. Diabetes is influenced by different parts of the body which incorporates

- **Cardiovascular disease:** Diabetes dramatically increases the risk of various cardiovascular problems, including coronary artery disease with chest pain (angina), heart attack, stroke and narrowing of arteries (atherosclerosis). If you have diabetes, you're more likely to have heart disease or stroke.
- **Nerve damage (neuropathy):** Excess sugar can injure the walls of the tiny blood vessels (capillaries) that nourish your nerves, especially in your legs. This can cause tingling, numbness, burning or pain that usually begins at the tips of the toes or fingers and gradually spreads upward. You could lose all sense of feeling in the affected limbs. Damage to the nerves related to digestion can cause problems with nausea, vomiting, diarrhea or constipation. For men, it may lead to erectile dysfunction.
- **Kidney damage (nephropathy):** The kidneys contain millions of tiny blood vessel clusters (glomeruli) that filter waste from your blood. Diabetes can damage this delicate filtering system. Severe damage can lead to kidney failure or irreversible end- stage kidney disease, which may require dialysis or a kidney transplant.
- **Eye damage (retinopathy):** Diabetes can damage the blood vessels of the retina (diabetic retinopathy), potentially leading to blindness. Diabetes also increases the risk of other serious vision conditions, such as cataracts and glaucoma.
- **Foot damage:** Nerve damage in the feet or poor blood flow to the feet increases the risk of various foot complications. Left untreated, cuts and blisters can develop serious infections, which often heal poorly. These infections may ultimately require toe, foot or leg amputation.
- **Skin conditions:** Diabetes may leave you more susceptible to skin problems, including bacterial and fungal infections.
- **Hearing impairment:** Hearing problems are more common in people with diabetes.

In order to lower the morbidity and reduce the influence of DM, it is vital for us to focus on a high-risk group of people with DM. According to the latest World

Health Organization (WHO) standard, the definitions of groups with a high risk of DM are as follows:

- Age ≥ 45 and seldom exercising
- BMI $\geq 24\ kg/m^2$
- Impaired glucose tolerance (IGT) or impaired fasting glucose (IFG)
- Family history of DM
- Lower high-density lipoprotein cholesterol or hypertriglyceridemia (HTG)
- Hypertension or cardiovascular and cerebrovascular disease
- Gestation female whose age ≥ 30

Diagnosis of diabetes is done in accordance with fasting blood tests, which is performed after having a fast of eight hours [3]. It requires much effort for testing. Recent improvement in technology has revolutionized various field of society using data mining [5]. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision-making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. Several existing techniques have not provided effective results when different parameters were used for prediction of diabetes. In our study, diabetes is predicted with the assistance of significant attributes, and the association of the differing attributes. In order to research the high-risk group of DMs, we need to utilize advanced information technology. Therefore, data mining technology is an appropriate study field for us.

1.3 Problem Statement

We considered the PIMA Diabetes dataset which is UGC approved repository. The dataset consists of irrelavant and noise data. It also consists of irrelavant features. Handling of irrelavant data and selecting an appropriate model is the main objective of the work. Evaluating the model with the help of evaluation metrics and achieving better accuracy is the post preprocessing work involved.

2. Literature Review

2.1 Literature Review

Diabetes analysis and prediction has become an area of tremendous relevance as the global diabetes rate has increased [13]. The machine learning models were used to The Pima Indian diabetes database by Saru, S., and S. Subashree [14]. With 10 cross-validation runs of their Naive Bayes, Decision Trees, and KNN models, they used bootstrapping resampling to forecast and compare their accuracy. The proposed methodology was determined to have an accuracy of 90.36 percent. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz [15] investigated the accuracy of different data-mining strategies, including decision trees, Naive Bayes, SVM, and hybrid algorithms. With an accuracy of 94% and sensitivity of 91%, hybrid techniques (proposed ensemble SVM + decision tree with 100 iterations) surpassed all other algorithms.

Sneha N. and Tarun Gangil [16] investigated various classification algorithms in order to determine the best diabetes predictor. The study used five classification algorithms: random forest, KNN, decision tree, Naive Bayes, and SVM, which were all obtained from the UCI machine repository archive. The best accuracy was 82.3 percent for Naive Bayes. Aada, A., and Sakshi Tiwari [17] analysed the PIMA Indian diabetes dataset using KNN, Naive Bayes, and decision trees, as well as bootstrapping-like approaches. After bootstrapping, SVM has the highest accuracy of 94.44 percent. On the Pima Indians dataset, Srivastava and Suyash [18] used machine learning algorithms and artificial neural networks to predict diabetes.

The PIMA Indian diabetes dataset was used in a study by Kaur, Harleen, and Vinita Kumari [19] to predict and assess diabetes trends. For prediction, they employed R data processing software and five algorithms: SVM-linear, radial basis function kernel support vector machine, k-nearest neighbor, artificial neural network, and multi-factor dimensions reduction. For diabetes prediction, the SVM-linear model had the highest accuracy of 89 percent. Md Maniruzzaman used the dataset of diabetes from the survey done by the National Health and Nutrition Examination [20], which included 6561 people, 657 of whom were diabetic. For diabetes prediction, logistic regression, Naive Bayes, decision trees, AdaBoost, and random forest were used. With logistic regression as feature selection and random forest for classification, the greatest accuracy of 94.25 percent was achieved.

Naive Bayes, decision tree, J48, and random forest with 10 fold cross-validation were used by Prasad, K.S., Reddy, N.C.S., and Puneeth, B.N. [21]. The most accurate method was Random Forest, whereas Naive Bayes had the lowest mean absolute error and root mean squared error.

“Continuous Glucose Monitoring: Review of an Innovation in Diabetes Management” was a research paper written by Z. Mian and the team in 2019 that predicted the outcome “Continuous glucose monitoring and sensor-enabled pump technology are used. This technology eliminates the need for frequent blood glucose monitoring, which is often inconvenient for patients, and instead offers them a more convenient option.” [1]

“Enabling large-scale biomedical analysis in the cloud” was a research paper written by Lin and the team in 2013 that predicted the outcome “Explains the data-intensive computing system and lists available cloud-based bioinformatics resources. To make a large amount of variety of data understandable and usable for biomedical research, we need to make it easier.” [2]

“Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends” was a research paper written by Mohammed and the team in 2014 that predicted the outcome “The Hadoop platform and the MapReduce programming framework's potential applications. To process large amounts of clinical data in sectors connected to medical health informatics.” [3]

“Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage” was a research paper written by Lee and the team in 2013 that predicted the outcome “The feasibility of three database technologies - NoSQL, XML-enabled, and native XML - for structured clinical data is evaluated. The greatest choice for query performance is a NoSQL database, although XML databases are better in terms of scalability, flexibility, and extensibility, all of which are necessary to deal with the features of clinical data.” [4]

“Leveraging Big Data analytics and Hadoop in Developing India’s Health Care Services” was a research paper written by D. Peter Augustine and the team in 2014 that predicted the outcome “The use of Big Data Analytics and Hadoop illustrates the significance of these technologies in providing healthcare services to everyone at the lowest possible cost.” [5]

“Big data analytics in healthcare: promise and potential” was a research paper written by Wullianallur Raghupathi in 2014 that predicted the outcome “To describe big data analytics' promise and potential in healthcare. For healthcare academics and practitioners, it provides a wide understanding of big data analytics.” [6]

“Quality assurance tasks and tools: The many roles of machine learning.” was a research paper written by Kalet and the team in 2019 that predicted the outcome “Improvements in planning time, plan quality, advanced dosimetric QA, predictive machine maintenance, higher safety checks, and advancements are all important for new adaptive planning-driven QA paradigms.” [7]

“Prevalence of diabetes mellitus amongst hospitalized tuberculosis patients at an Indian tertiary care center: A descriptive analysis” was a research paper written by Tripti and the team in 2018 that predicted the outcome “Age, type of TB, and undernutrition were all found to be significant predictors of TB-DM co-prevalence.” [8]

“A Variable Service Broker Routing Policy for data center selection in cloud analyst.” was a research paper written by Ahmad M. Manasrah in 2017 that predicted the outcome “Variable Service Broker Routing Policy is used to reduce the processing and response time of customer requests while staying within a reasonable cost range. The proposed policy alters the old policy's sorting and selection equations.” [9]

“A Smart Glucose Monitoring System for Diabetic Patients.” was a research paper written by A. Rghioui and the team in 2020 that predicted the outcome “A compact portable device capable of detecting blood glucose levels and body temperature in diabetics. Diabetes disease surveillance would allow doctors to remotely monitor their patients' health using sensors included in smartphones and smart portable devices.” [10]

“Predictive Methodology for Diabetic Data Analysis in Big Data.” was a research paper written by N. M. S. Kumar and the team in 2015 that predicted the outcome “Hadoop/predictive MapReduce's analytical method Reduce the environment to forecast the types of diabetes that are common and the complications that come with it. This approach enables patients to be cured and cared for in a more efficient manner, with improved outcomes like as affordability and accessibility.” [11]

“Effects of External Factors in CGM Sensor Glucose Concentration Prediction.” was a research paper written by Ahmed and the team in 2016 that predicted the outcome “To develop a blood glucose prediction system for usage in conjunction with a continuous glucose monitoring (CGM) device. This technology eliminates the need for frequent blood glucose monitoring, which is often inconvenient for patients, and instead offers them a more convenient option.” [12]

“Predictive Methodology for Diabetic Data Analysis in Big Data.” was a research paper written by Dr. Saravana Kumar and the team in 2015 that predicted the outcome “Hadoop/predictive MapReduce's analytical method Reduce the environment to forecast the types of diabetes that are common and the complications that come with it. This approach enables patients to be cured and cared for in a more efficient manner, with improved outcomes like as affordability and accessibility.” [13]

“An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm.” was a research paper written by Dost Mohammad Khan 1 and team in 2011 that predicted the outcome “To create a more efficient data mining approach employing intelligent agents, we combine the K-means clustering algorithm with the Decision tree (ID3) algorithm. Data mining algorithms are used to uncover hidden patterns and connections across variables in large datasets.” [14]
“Prediction of Diabetes Using Classification Algorithm.” was a research paper written by Deepti Sisodia in 2018 that predicted the outcome “The goal is to create a model that can accurately predict the likelihood of diabetes in people. The designed system can be used to predict or diagnose diabetes using machine learning classification methods.” [15]

3. METHODOLOGY

3.1 Proposed System

3.1.1 System Architecture

In this paper, we have proposed a three-level architecture model for predicting type2 diabetes. It consists of Data pre-processing, Data Modeling, and Evaluation metrics as core methodologies. Fig.1 shows the architecture of the entire system.

In the first phase, Diabetes dataset is taken as input for constructing the models. The dataset initially consists of redundant and useless data. To eliminate such data, Pre-processing techniques like Data cleaning and Feature selection are applied over the dataset to improve the accuracy of the model. It is done to eliminate the problem of over fitting. The output data is then split into two subsets, Training data, and testing data. We considered 75% of our data to be training data and 25% as testing data.

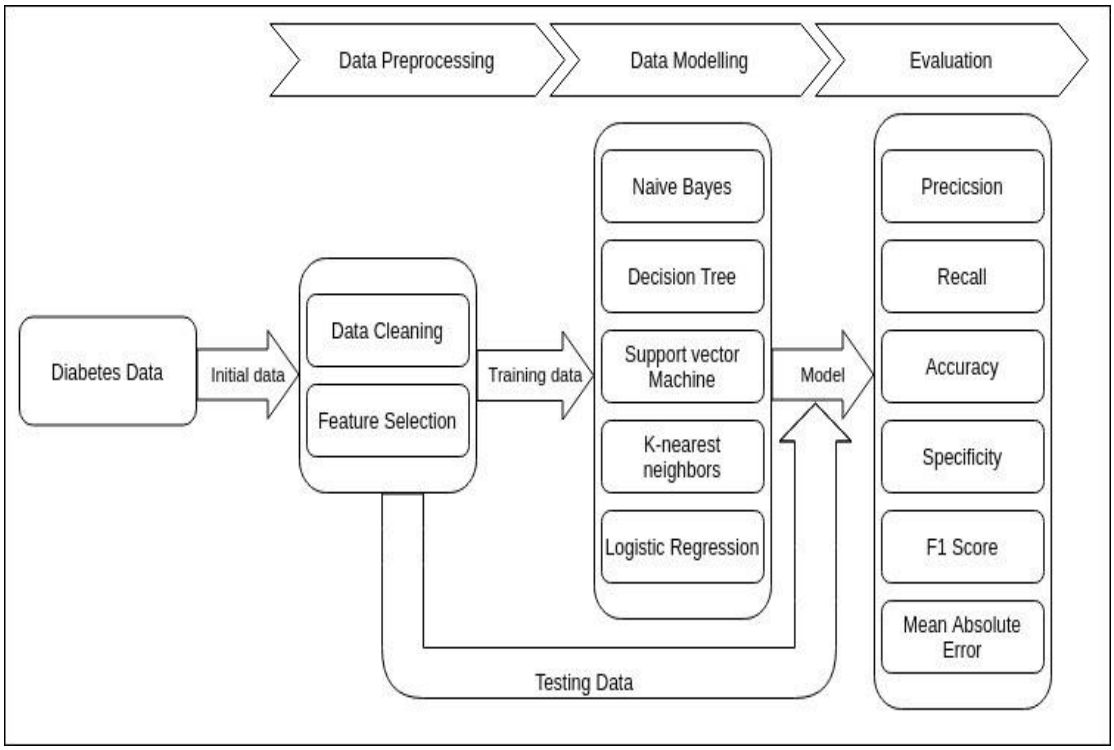


Fig 3.1 System Architecture

In the second phase, different algorithms like Naive Bayes, Decision Tree, Support Vector Machine, K-nearest neighbors, and Logistic Regression are applied

over the training dataset. Models are constructed from different algorithms. The constructed model is then tested with the testing dataset.

In the third phase, Evaluation metrics like Precision, Recall, Specificity, Accuracy and Mean Absolute Error are calculated to predict the accuracy of the model. Evaluation metrics of all the models is compared. In the next stage, Cross Validation is performed to improve the accuracy of the model.

3.2 Dataset

The quality of data, to a major extent, affects the predicted result. The accuracy depends mainly on the data considered. We used the "Pima Indians Diabetes Dataset" standard which was supported by the UCI machine learning repository [12]. This is a standard dataset that has drawn the values from the real instances.

Dimensions of the dataset: (768, 9)

It consists of 768 instances and each instance is associated with 9 attributes which are all numeric values. The table shows the names, description and value range of each attribute.

Sample Dataset:

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

Fig 3.2 Sample dataset

S.NO	Name	Description	Unit	Value Range
1	Pregnancy	No of times pregnant	Numeric value	0-9
2	Glucose	Glucose content	Numeric value	0-199
3	Blood Pressure	Diastole blood pressure	mmHg	0-122
4	Skin	Triceps skin fold thickness	Mm	0-99
5	Insulin	2-hours serum insulin	Mu/Uml	0-846
6	BMI	Body mass index	Weight in kg Height in m	0-67.1
7	Pedigree	Pedigree function	Numeric value	0.08-2.42
8	Age	Age	Numeric value	21-81
9	Class	Diabetes mellitus type 2	Numeric value	Positive=1 Negative=0

Table 3.1 Dataset description, units and value range.

3.3 Data Pre-processing

Data Preprocessing is the process of collecting and modifying the data into desired format. The collected data may consist of redundant, inconsistent and noisy data lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Preprocessing of data helps in resolving such type of data.

The Bioinformatics datasets collected from the field, are very raw and have the tendency of following characteristics. This data has to be processed before analysing through Data Mining Techniques.

a. Incomplete

When collecting the data of any domain or bioinformatics data from the field, there is a possibility of lacking in attribute values or certain attributes of interest, or containing only aggregate data.

b. Missing data

Seen particularly for tuples with missing values for some attributes, may need to be inferred.

c. Noisy

Noisy data means, the data in the tuples containing errors, or outlier values that deviate from the expected.

d. Incorrect data

It may also produce the result from inconsistency in naming conventions or data codes used, or inconsistent formats for input fields, such as date. Hence it is necessary to use some techniques to replace the noisy data.

e. Inconsistent

Inconsistent means, the data source containing discrepancy between different data items. Some attributes representing the given concept may have different names in different databases, causing inconsistency and redundancy. Naming inconsistency may also occur in attribute values. Therefore, the inconsistency in data needs to be removed.

f. Aggregate Information

It would be useful to obtain aggregate information such as the bioinformatics data sets something that is not a part of any pre-computed data cube in the data warehouse.

g. Enhancing Mining Process

Large number of data sets may make the data mining process slow. Hence, reducing the number of data sets to enhance the performance of the mining process is important.

h. Improve Data Quality

Data Pre-processing techniques can improve the quality of the data, thereby help to improve the accuracy and efficiency of the subsequent mining process. Data Pre-processing is an important step in the knowledge discovery process, because quality decisions is based on the quality data. The detecting data become anomalies and rectifying them can lead to improve the accuracy and efficiency of the data analysis.

3.3.1 Data Preprocessing Methods

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to improve the quality of the data and the mining results, raw data is pre-processed so as to improve the efficiency and ease the mining process. Data Pre-processing is one of the most critical steps in data mining process which deals with the preparation and transformation of the initial dataset. Data pre-processing methods are divided into the following categories. They are:

1. Data Cleaning.

2. Data Integration.

3. Data Transformation.

4. Data Reduction.

3.3.1.1 Data Cleaning

Data is analysed by data mining techniques which may incomplete, noisy, and inconsistent. Real-world data tend to be incomplete, noisy, and inconsistent. International Journal of Pure and Applied Mathematics Special Issue 788 Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. This section, gives basic methods for data cleaning. Incomplete, noisy, and inconsistent data are the common properties of real-world databases and data warehouses. Data can be noisy, having incorrect attribute values. Owing to the following, the data collection instruments used may be fault. There may be human or computer errors occurred at data entry. Errors in data transmission can also occur. There may be technological limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also produce the result from inconsistency in naming conventions or data codes. Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistency. Dirty data can cause confusion for the mining procedure. Although most mining routines have some procedures, they deal incomplete or noisy

data, which are not always robust. Therefore, a useful pre-processing step is to run the data through some data cleaning routines.

3.3.1.2 Data Integration

It is likely that the data analysis task involves in data integration, which combines the data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are number of issues to consider during data integration. Schema integration is tricky. How can real world entities from multiple data sources be 'matched up'? This is referred to as the entity identification problem. For example, how can the data analyst or the computer be sure that customer id in one database, and cust_number in another refer to the same entity? databases and data warehouses typically have metadata - that is, data about the data. Such metadata is used to help avoid errors in schema integration. Redundancy is another important issue. An attribute may be redundant, if it is derived from another table, such as annual revenue. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

3.3.1.3 Data Transformation

In data transformation, the data are transformed or consolidated into appropriate forms for mining. Data transformation involves the following:

1. In Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.

2. Smoothing works remove the noise from the data. Such techniques include binning, clustering, and regression.

3. In Aggregation, summary or aggregation operations are applied to the data.

3.3.1.4 Data Reduction

Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data reduction techniques is helpful in analyzing the reduced representation of the dataset without compromising the integrity of the original data and yet producing the qualitative knowledge. The concept of data reduction is commonly understood as either reducing the volume or

reducing the dimensions (number of attributes). There are number of methods that facilitate in analyzing the reduced volume or dimension of data and yield useful knowledge. Certain partition-based methods work on partition of data tuples. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following:

1. In Data cube aggregation, aggregation operations are applied to the data in the construction of a data cube.

2. In Dimension reduction, irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

3. Data compression - encoding mechanisms are used to reduce the data set size. The methods are used for data compression are wavelet transform and principle component analysis.

4. Numerosity reduction - the data is replaced or estimated by alternative, smaller data representations such as parametric models (which store only the model parameters instead of the actual data e.g. regression and log-linear models), or nonparametric methods such as clustering, sampling, and the use of histograms.

3.3.2 Data Cleaning

In the dataset, we can see some missing values. Most of the inaccurate experimental results were caused by these meaningless values. These values can be replaced by the average of values either mean, median, mode of the attribute. For example, in the original dataset, the values 0, indicates that the real value was missing. We replace them by using the mean of the attributes.

3.3.3 Feature Exploration

Feature selection also called as attribute selection is the process of gathering relevant features for constructing the model. The features are selected based on the correlation between the attributes. It helps in reducing the irrelevant data and improve the prediction accuracy. It is also used to combine different features to produce more sophisticated features. We have used random forest classifier for selecting the features.

Random Forest Classifier

Random Forest is a multi-tree classifier that can be used for both Regression and Classification problems. Decision trees form the basic building blocks of the random forest model. It splits the data into different samples in a random fashion and constructs a decision tree for each data sample. The predictions of each tree are put for voting. In classification problems, the class having the highest number of votes will be considered. In Regression problems, the average of all the class predictions will be considered. The accuracy of the model mainly depends upon the number of trees constructed.

The logic behind the random forest is bagging technique to create random sample features. The difference between the decision tree and the random forest is the process of finding the root node and splitting the feature node will run randomly.

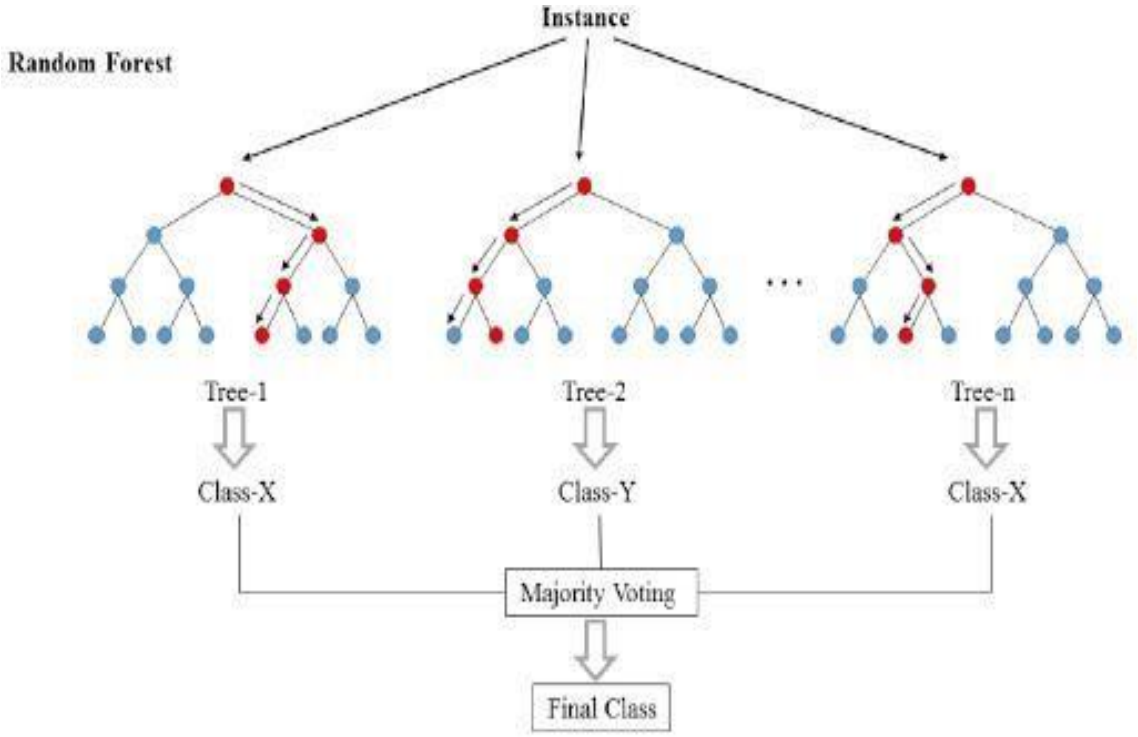


Fig 3.3 Random Forest Classifier

The Steps are given below

1. Load the data where it consists of “m” features representing the behavior of the dataset.

2. The training algorithm of random forest is called bootstrap algorithm or bagging technique to select n feature randomly from m features, i.e. to create random samples.

3. Calculate the node d using the best split. Split the node into sub-nodes.

4. Repeat the steps, to find n number of trees.

5. Calculate the total number of votes of each tree for the predicting target. The highest voted class is the final prediction of the random forest.

3.4 Model Selection

Model selection or algorithm selection is the process of selecting a model that better classifies the dataset. Model selection is the task of selecting a statistical model from a set of candidate models, given data. In the simplest cases, a pre-existing set of data is considered. However, the task can also involve the design of experiments such that the data collected is well-suited to the problem of model selection. Model construction is completely dependent on the algorithm. There are also some algorithms which do not construct any model and mostly relies on the dataset. In conclusion, we select a model that gives the highest accuracy. The classifiers selected are Naïve Bayes, Decision Tree, Support Vector Machine, Logistic Regression and K- Nearest Neighbor.

3.4.1 Naive Bayes Algorithm

Naïve Bayes is a classification algorithm, a probabilistic classifier that is based on Bayes theorem with naïve independence assumptions. Naïve Bayesian method takes the dataset as input, performs analysis and predicts the class label using Bayes' Theorem. It calculates a probability of class in input data and helps to predict the class of the unknown data sample. It is a powerful classification technique suitable for large datasets. It assumes that the features are independent of the given class. Bayes Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Flow Chart for Naïve Bayes Classifier is represented in next page. Bayes' Theorem is stated as:

$$p(h \not d) = p(d \not h) * \frac{p(h)}{p(d)}$$

Where

$p(h \not| d)$ is the probability of hypothesis h given the data d. This is called the posterior probability.

$p(d \not| h)$ is the probability of data d given that the hypothesis h is true

$p(h)$ is the probability of hypothesis h being true (regardless of the data) This is called the prior probability of h

$p(d)$ is the probability of the data (regardless of the hypothesis)

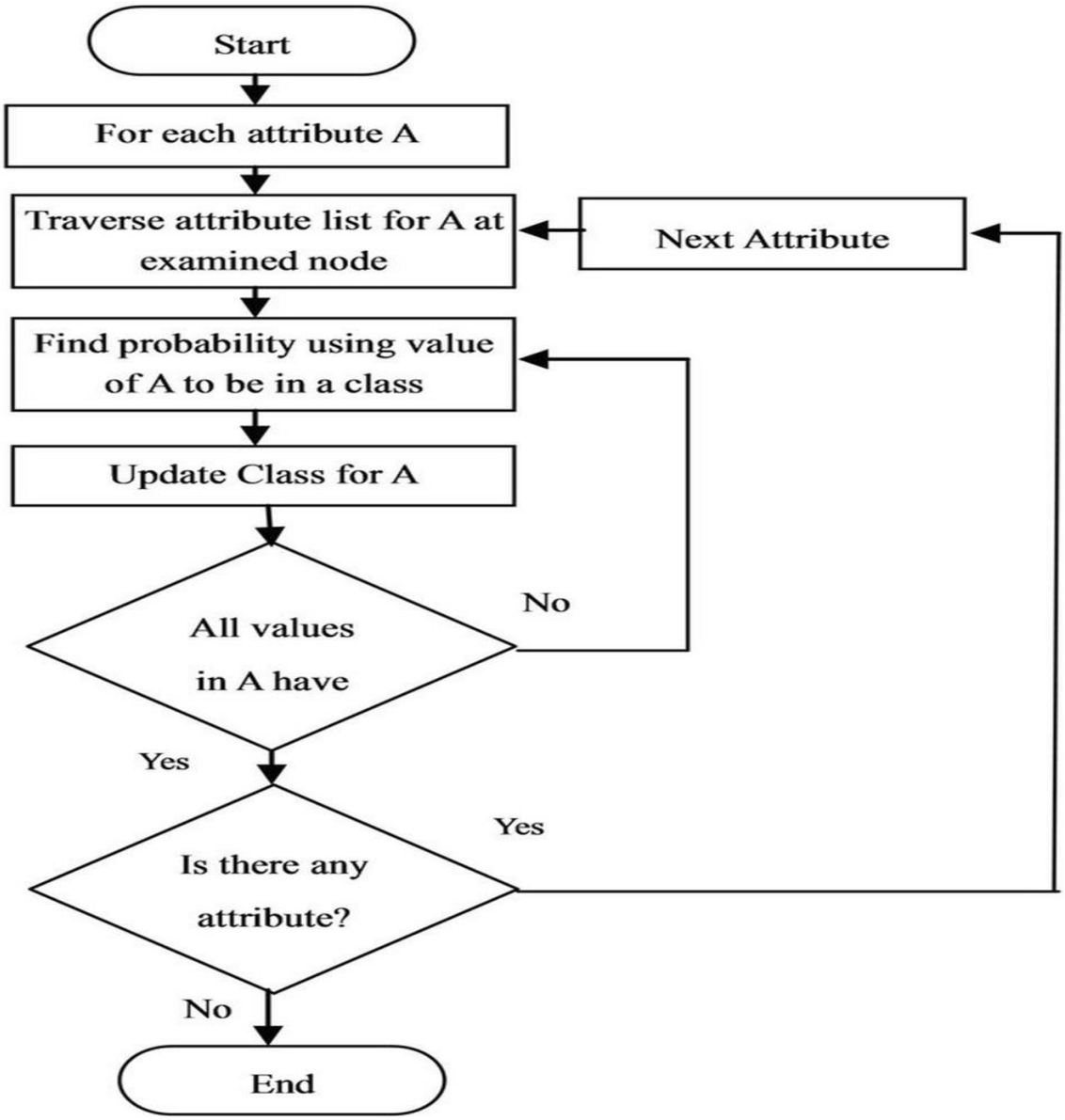


Fig 3.4 Naive Bayes Algorithm

3.4.2 Support Vector Machine

SVM is a classification algorithm that maps the data item into points in the n-dimensional space. The main objective is to find a linear decision boundary (hyperplane) that segregates the data into two classes. The optimal hyperplane is said to be optimal based on maximum marginal boundary i.e., Euclidean distance between the support vectors is maximum. SVM mainly emphasizes on risk minimization. Using of SVM can help in achieving greater performance as it uses significantly less data when compared to other classifiers.

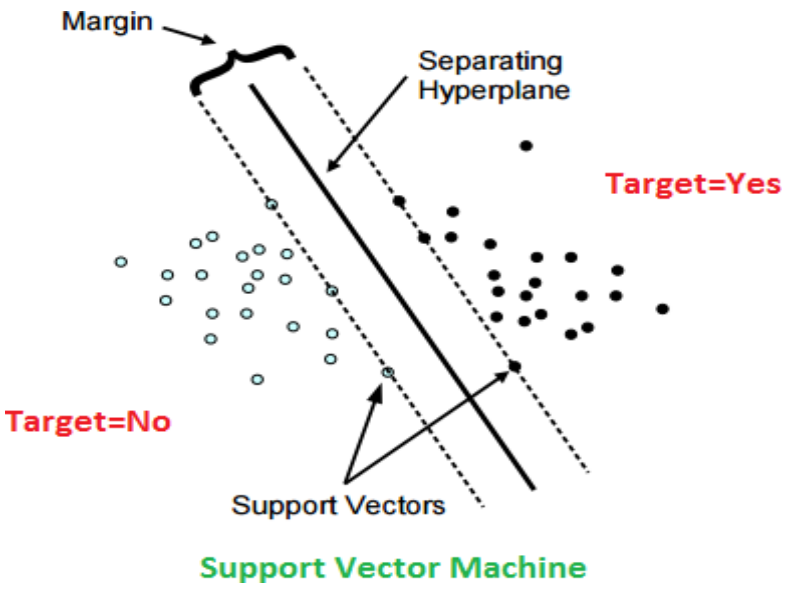


Fig 3.5 Support Vector Machine

3.4.3 Logistic Regression

Logistic Regression is a classification technique that works on a probability basis. It is much similar to Linear Regression which gives continuous value. It is named after the logistic function which is the core method. The Logistic function [13], also called a sigmoid function is an S-shaped curve that takes a real number and transforms it into a value that lies between 0 and 1. The sigmoid function is mathematically represented as

$$\mathbf{F(x)} = \frac{1}{1+e^{-x}}$$

When using linear regression, we used a formula of the hypothesis

$$H(x) = \beta_0 + \beta_1 X$$

For logistic regression, we apply sigmoid function to the hypothesis of linear regression.

$$\sigma(Z) = \sigma(\beta_0 + \beta_1 X)$$

To transform these values to a discrete class, we select a threshold value, above which we will classify values into class A and below which we classify values into class B.

$$p >= 0.5, \text{ class}=1$$

$$p < 0.5, \text{ class}=0$$

In conclusion, we decided to use the logistic regression as one part of our proposed model.

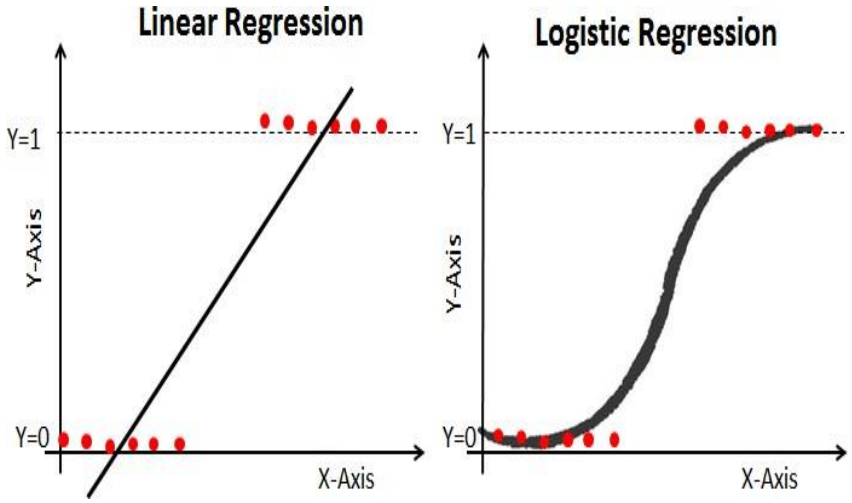


Fig 3.6 Linear and Logistic Regressions

3.4.4 KNN Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN is an instance-based learning algorithm that uses the entire training data for prediction. It does not require any learning as it doesn't have any model.

It is also called as a lazy algorithm as it does not perform any generalization. It is based on the theory of feature similarity**.** In this method, each sample should be

classified similarly to the surrounding samples. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. Classification is done based on minimum distance from the new point to the K nearest neighbors i.e., Euclidean or Manhattan distance.

The performance of a KNN classifier is primarily determined by the choice of K as well as the distance metric applied. With increase in the value of K, the new point is calculated more accurately as more neighbors are included. The steps for KNN are given below:

1. Training phase of the algorithm consists of only storing the feature sample and class label of training sample.

2. Classification phase: the user has to define a “k” value for the classification of the undefined sample for the k number of the class labels, so the unlabeled sample can be classified into the defined class based on the feature similarity.

3. Majority of voting classification occurs for unlabeled class. The value of the k can be selected by various techniques like heuristic technique.

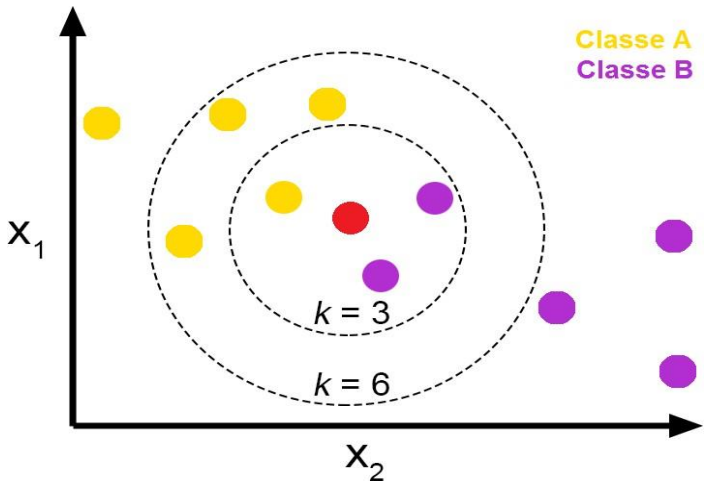


Fig 3.7 KNN Classifier

3.4.5 Decision Tree

A decision tree is a pictorial representation of all the attributes in the dataset to achieve a more generalized conclusion. It is a prediction algorithm that is used for classification purposes. This methodology is more commonly known as learning

decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. The dataset is divided into different subsets based on the most prominent attribute.

Several algorithms like Hunt's, ID3, and CART algorithms were used to implement the decision tree. The way the dataset is split is completely dependent on the algorithm. A decision tree is drawn upside down with its root at the top. Each node except the leaf node consists of two branches, yes and No. Based on the decision, the tree is constructed. It is done until all the attributes are included in the tree. A decision tree can be used to visually and explicitly represent decisions and decision making.

Tree Pruning method is done at the end to eliminate the unwanted data. This helps in improving the accuracy of the model. For example, in the ID3 algorithm, the attributes with highest information gain is selected as root node. The entropy of each attribute is also calculated. Mathematical Expressed as follows

$$\text{Entropy(S)} = \sum_{i=1}^{n} -p_i \log_2 p_i$$

$$\text{Gain (S, J)} = \text{Entropy(S)} - \sum x \in values(J) \frac{|S_x|}{|S|} \text{Entropy } (s_x)$$

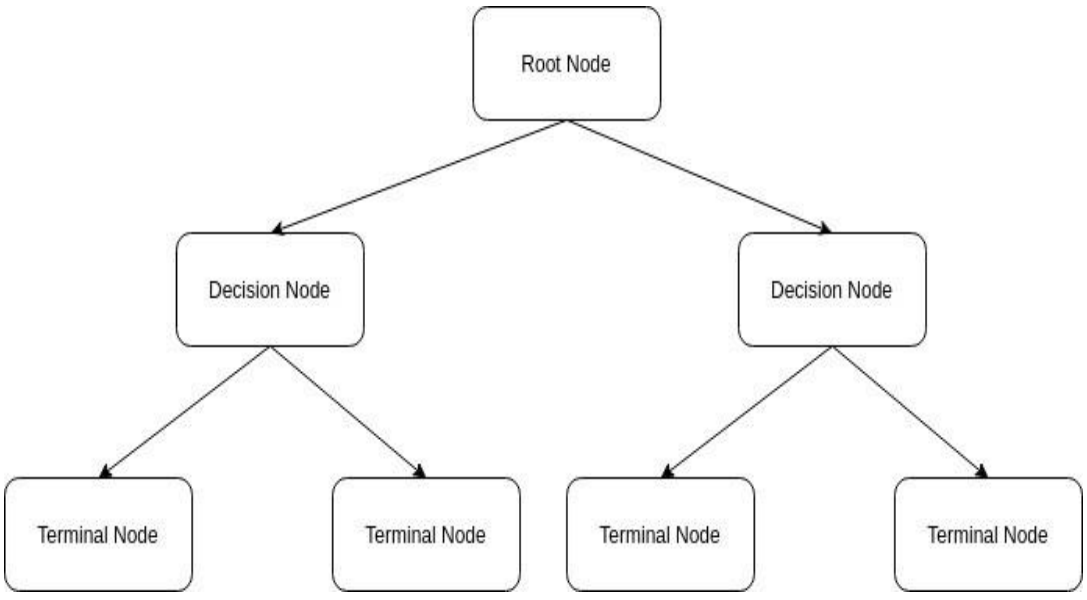


Fig 3.8 Decision Tree

4. Experimental analysis and results

4.1 System configuration

4.1.1 Software Requirements:

Operating System	:	Windows XP Service Pack 2 or above.
Software	:	JUPYTER IDE, WEKA TOOL.

4.1.2 Hardware Requirements:

Input Devices	:	Keyboard and Mouse
RAM	:	512 MB
Processor	:	Intel Pentium4 or above
Storage	:	100 MB of HDD space.

4.2 Screenshots

4.2.1 Cross validation

```
                        Accuracy
Naive Bayes             0.770830
Support Vector Machine  0.769481
Logistic Regression     0.766900
K Nearest Neighbor      0.765602
Decision Tree           0.705622
```

4.2.2 Decision tree

```
0.7402597402597403
0.7012987012987013
confusion matrix: [[167  34]
 [ 46  61]]
precision 0.8308457711442786
recall 0.784037558685446
tnr 0.6421052631578947
accuracy 0.7402597402597403
f1_score 0.8067632850241545
mean_absolute_error 0.2597402597402597
```

4.2.3 Support Vector Machine

```
0.7662337662337663
0.7564935064935064
confusion matrix: [[177  24]
 [ 48  59]]
precision 0.8805970149253731
recall 0.7866666666666666
tnr 0.7108433734939759
accuracy 0.7662337662337663
f1_score 0.8309859154929576
mean_absolute_error 0.23376623376623376
```

4.2.4 Naive Bayes

```
0.7694805194805194
0.75
confusion matrix: [[174  27]
 [ 44  63]]
precision 0.8656716417910447
recall 0.7981651376146789
tnr 0.7
accuracy 0.7694805194805194
f1_score 0.8305489260143197
mean_absolute_error 0.2305194805194805
```

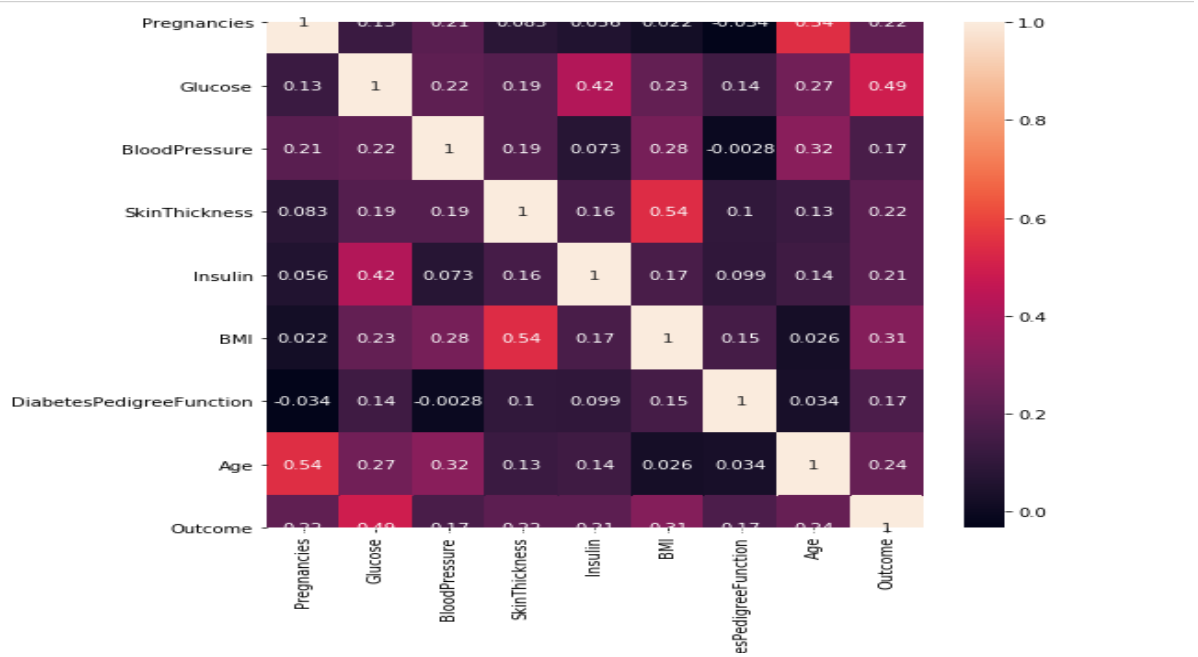
4.2.5 Logistic Regression

```
0.7564935064935064
0.7597402597402597
confusion matrix: [[179  22]
 [ 53  54]]
precision 0.8905472636815921
recall 0.771551724137931
tnr 0.7105263157894737
accuracy 0.7564935064935064
f1_score 0.8267898383371824
mean_absolute_error 0.2435064935064935
```

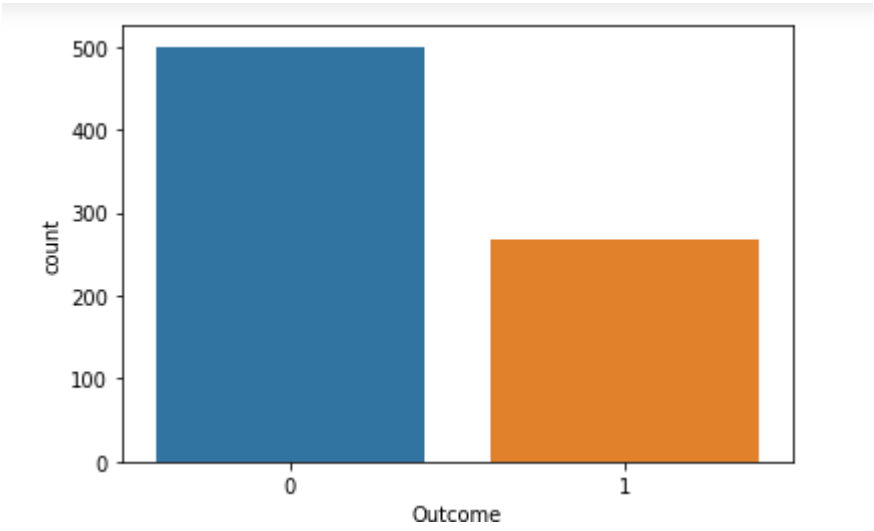
4.2.6 K-Nearest Neighbors

```
0.7694805194805194
0.7532467532467533
confusion matrix: [[184  17]
 [ 54  53]]
precision 0.9154228855721394
recall 0.773109243697479
tnr 0.7571428571428571
accuracy 0.7694805194805194
f1_score 0.8382687927107062
mean_absolute_error 0.2305194805194805
```

4.2.7 Heat Map



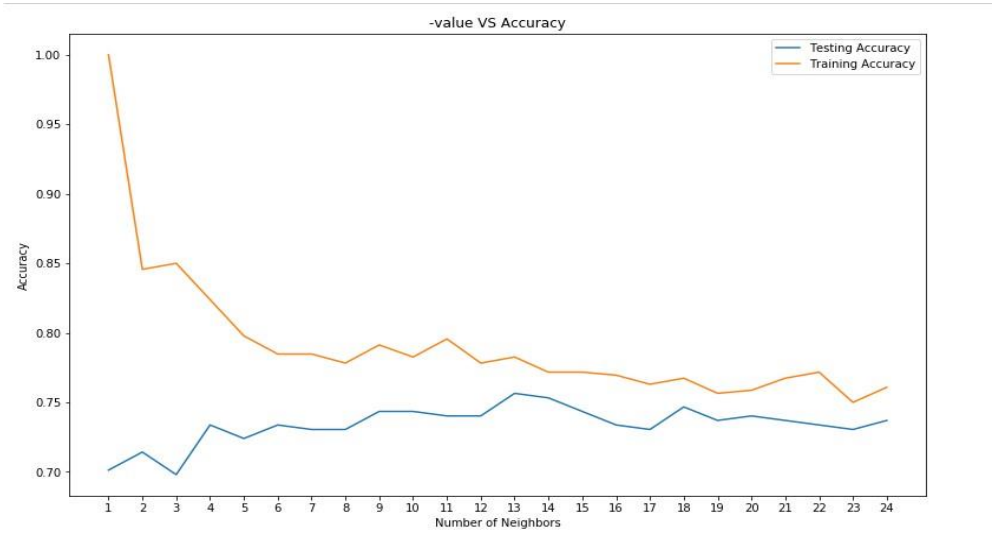
4.2.8 Outcomes count



4.2.9 Data Cleaning

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

4.2.10 K- Nearest Neighbors




```
Best accuracy is 0.7564935064935064 with K = 13
```

4.2.11 Dataset Description

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.435949	12.096346	8.790942	85.021108	6.875151	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.153420	155.548223	32.400000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	155.548223	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

4.3 Experimental Analysis/Testing

The idea of building machine learning models works on a constructive feedback principle. You build a model, get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results.

Simply building a predictive model is not your motive. It's about creating and selecting a model which gives high accuracy on out of sample data. Hence, it is crucial to check the accuracy of your model prior to computing predicted values. In our industry, we consider different kinds of metrics to evaluate our models. The choice of metric completely depends on the type of model and the implementation plan of the model.

Evaluation metric plays a critical role in achieving the optimal classifier during the classification training. Thus, a selection of suitable evaluation metric is an important key for discriminating and obtaining the optimal classifier. Generally, many generative classifiers employ accuracy as a measure to discriminate the optimal solution during the classification training. However, the accuracy has several weaknesses which are less distinctiveness, less discriminability, less informativeness and bias to majority class data**.**

Classification Metrics

Binary classification is one of the most frequent studies in applied machine learning problems in various domains, from medicine to biology to meteorology to malware analysis. Many researchers use some performance metrics in their

classification studies to report their success. However, the literature has shown a widespread confusion about the terminology and ignorance of the fundamental aspects behind metrics.

In binary classification, there are two possible output classes. In multi-class classification, there are more than two possible classes. There are many ways of measuring classification performance like Accuracy, Confusion Matrix, Precision, Recall, F1 score and Mean Absolute Error.

Performance metrics for binary classification are designed to capture trade-offs between four fundamental population quantities: true positives, false positives, true negatives and false negatives. Despite significant interest from theoretical and applied communities, little is known about either optimal classifiers or consistent algorithms for optimizing binary classification performance metrics beyond a few special cases.

We will be evaluating the model by splitting the data set into two portions, training set, and testing set. The training set is used to train the model and the testing set is used to test the model. After being processed by classification algorithms, we evaluate the accuracy of the model.

4.3.1 Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. The information about the actual and predicted classification system is held by the Confusion matrix. It demonstrates the accuracy of the solution to a classification problem.

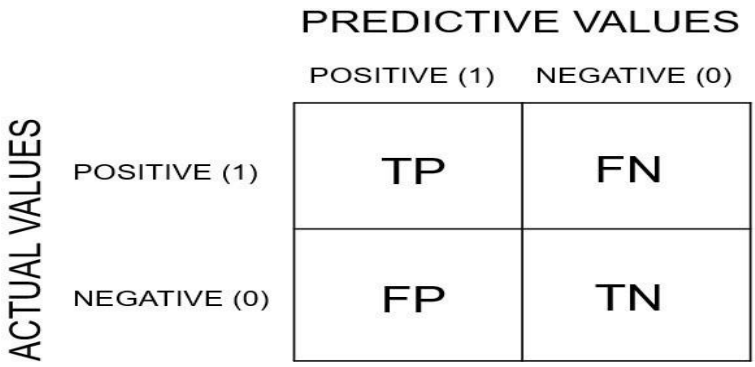


Fig 4.1 The Confusion Matrix for a two class Classifier

Above figure shows the confusion matrix for a binary classifier. The entries in the confusion matrix have the following meaning in the context of our study.

- TP is the number of correct predictions that an instance is positive.
- FN is the number of incorrect predictions that an instance is negative.
- FP is the number of incorrect predictions that an instance is positive and
- TN is the number of correct predictions that an instance is negative.

4.3.2 Precision

Precision is the ratio of correctly predicted true positive events to the total number of predicted positive events. It predicts the percentage of events that are relevant among all the predicted events.

$$\boldsymbol{Precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fp}}$$

4.3.3 Mean absolute error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} |\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{y}|_{\boldsymbol{i}}$$

4.3.4 Recall / True positive rate / Sensitivity

Sensitivity is the ratio of correctly predicted true positive events to the total number of actual positive events. The recall is also known as sensitivity or true positive rate. The sensitivity of the test is the proportion of people who test positive for the disease among those who have the disease. Mathematically, this can be expressed as:

$$\boldsymbol{Recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp}+\boldsymbol{fn}}$$

4.3.5 True negative rate / Specificity

Specificity is the ratio of correctly predicted true negative events to the total number of actual negative events. Specificity is also known as a True negative rate. Specificity of a test is the proportion of healthy patients known not to have the disease, who will test negative for it. Mathematically, this can be expressed as:

$$\boldsymbol{Specificity} = \frac{\boldsymbol{tn}}{\boldsymbol{tn}+\boldsymbol{fp}}$$

4.3.6 Accuracy

The ratio of correctly classified samples to the total number of samples is known to be Accuracy (AC). It shows the overall effectiveness of the classifier. Accuracy can be a misleading metric for imbalanced data sets. It works well only if there are equal number of samples belonging to each class.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{tp} + \boldsymbol{tn}}{\boldsymbol{tp} + \boldsymbol{tn} + \boldsymbol{fp} + \boldsymbol{fn}}$$

4.3.7 F1 Score

F1 Score is the harmonic mean of precision and recall. It gives a better measure of the incorrectly classified cases than the Accuracy Metric. F1-score is a better metric when there are imbalanced classes.

$$\boldsymbol{F1} = \mathbf{2} * \frac{\boldsymbol{precision} \times \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}}$$

4.4 Cross validation

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing the model accuracy. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset and validating the analysis on the other subset. To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined over the rounds to give an estimate of the model's predictive performance. In summary, cross-validation combines (averages) measures of fitness in prediction to derive a more accurate estimate of model prediction performance.

Using this method, we split our dataset into ‘n’ equal parts. “n-1” parts are used in training while one part is used for testing the data. All possible combinations are used for both testing and training of data. Different accuracy scores are obtained for different combinations of data. The mean of these scores gives a generalized accuracy. Cross Validation table is as follows:

Classifier	**Accuracy**
Support vector machine	0.77
Naïve Bayes	0.75
Decision tree	0.68
K nearest neighbors	0.74
Logistic regression	0.76

Table 4.1 Comparison of accuracy of classifiers.

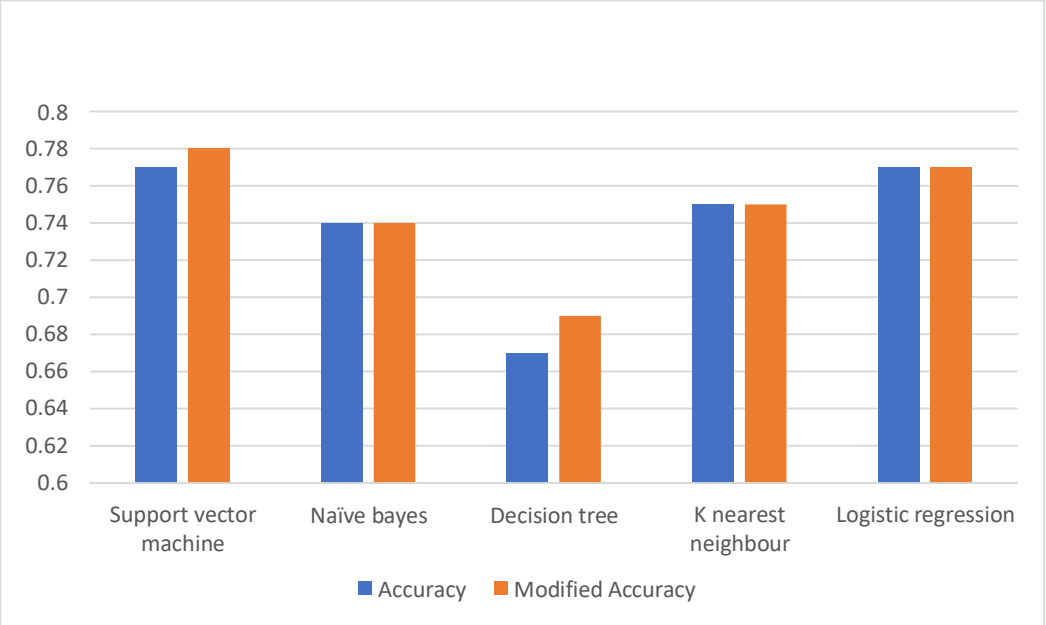
4.5 Comparison of classification algorithms

Below table provides the accuracy of classification algorithms before and after performing feature selection. We have considered only four attributes from the original dataset to increase the accuracy. By comparing the classifiers, we find SVM as the best classifier that better classifies the dataset which has achieved an accuracy of 78%.

Classifier	**Accuracy**	**Modified Accuracy**
Support vector machine	0.77	0.78
Naïve Bayes	0.74	0.74
Decision tree	0.67	0.69
K nearest neighbors	0.75	0.75
Logistic regression	0.77	0.77

Table 4.2 Comparison of different algorithms.

Fig 4.1 Graphical representation of the models based on their accuracy.



Below table provides the statistical information of models in terms of evaluation measures like Precision, Recall, Specificity, Accuracy, and Mean Absolute error.

Algorithm	**Precision**	**Recall**	**Specificity**	**Average**	**F1**	**Mean Absolute Error**
Support Vector Machine	0.90	0.78	0.75	0.77	0.84	0.22
Naïve Bayes	0.88	0.77	0.70	0.75	0.82	0.24
Decision tree	0.77	0.71	0.50	0.65	0.74	0.34
K-Nearest Neighbors	0.78	0.83	0.63	0.75	0.80	0.24
Logistic Regression	0.84	0.77	0.64	0.73	0.80	0.26

Table 4.3 Evaluation Measures of different algorithms.

Graphical representation of the above table gives an insight into the various machine learning models and their predictive accuracy in terms of performance.

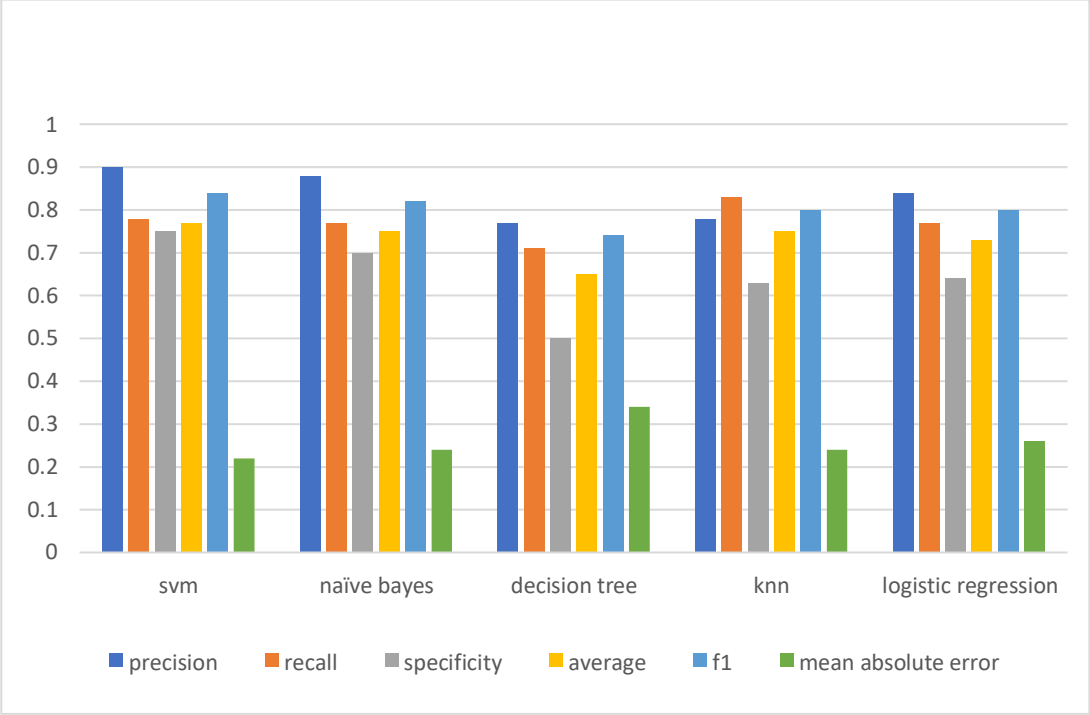


Fig 4.2 Graphical representation of evaluation measures.

5. Conclusion and Future Work

5.1 Conclusion

To support the lives of the people all over the world, we are trying to detect and prevent the complications of diabetes at the early stage through predictive analysis by improving the classification techniques. Our proposed work also performs the analysis of the features in the dataset and selects the optimal features based on the correlation values. The paper is aimed to provide a model that better classifies the instances of the dataset. Techniques like Data cleaning and Feature selection has helped to improve the potentiality of the dataset. All the Classifiers have achieved an accuracy of above 67%. SVM has achieved an accuracy of 77% which stands at the top of the order, while Decision Tree has achieved an accuracy of 67% which stands at the bottom among all the classifiers. Cross-validation is performed on each combination to get the mean accuracy of each model. SVM has achieved an accuracy of 78% while Logistic Regression has achieved 77%. Apart from Cross Validation, we also used Machine Learning tool kit, in order to make a valid comparison with others' results, it was necessary to conduct this model using the WEKA toolkit and use the same Pima Indian Diabetes Dataset. All the classifiers has achieved better accuracy when compared to WEKA tool. SVM and naïve Bayes are comparatively better in terms of evaluation metrics like precision, recall, f1 score, and specificity. SVM has less mean absolute error when compared to other models. By the comparative analysis, we specify SVM as the best model that fits the dataset concerning the diabetic and non-diabetic persons.

5.2 Future Work

Future work includes the usage of other sophisticated algorithms to increase the accuracy. Naïve Bayes consists of Gaussian, multinomial and Bernoulli's classifiers. We have implemented only the Gaussian naïve Bayes classifier. The other types can also be implemented for the comparative study of Bayes classifier. SVM consists of linear, non-linear, polynomial, sigmoid and other kernels. We implemented linear kernels as it best classifies our dataset. The other kernels can also be implemented for better classification. Future selection is achieved by using Random Forest Classifier. The other selection methods like wrapped, embedded, filters can also be used to increase the accuracy of the model. Developing an application that will

provide reasonable and rational health suggestions to the high-risk group. It is necessary to bring in hospital's real and latest patients' data for continuous training and optimization of our proposed model. The dataset seems to be imbalanced. The quantity of the dataset should be large enough for training and predicting. Using of technologies like Neural Networks could also achieve better accuracy in classifying the dataset.

6. References

[1] Kaveeshwar, S.A., and Cornwall, J., 2014, "**The current state of diabetes mellitus in India**". AMJ, 7(1), pp. 45-48.

[2] Dean, L., McEntyre, J., 2004, "**The Genetic Landscape of Diabetes [Internet]. Bethesda (MD): National Center for Biotechnology Information (US);.** Chapter 1, Introduction to Diabetes. 2004 Jul 7.

[3] D. Falvo, B.E. Holland **Medical and psychosocial aspects of chronic illness and disability Jones & Bartlett Learning** (2017)

[4] J.S. Skyler, G.L. Bakris, E. Bonifacio, T. Darsow, R.H. Eckel, L. Groop, *et al.* **Differentiation of diabetes by pathophysiology, natural history, and prognosis** Diabetes, 66 (2017), pp. 241-255

[5] Z. Tao, A. Shi, J. Zhao **Epidemiological perspectives of diabetes** Cell Biochem Biophys, 73 (2015), pp. 181-185

[6] W.H. Organization **World health statistics 2016: monitoring health for the SDGs sustainable development goals** World Health Organization (2016)

[7] L. Cobos **Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat)** Endocr Pract, 24 (2018), pp. 43-44

[8] B. Dorcely, K. Katz, R. Jagannathan, S.S. Chiang, B. Oluwadare, I.J. Goldberg, *et al.* **Novel biomarkers for prediabetes, diabetes, and associated complications** Diabetes, Metab Syndrome Obes Targets Ther, 10 (2017), p. 345

[9] P.P. Singh, S. Prasad, B. Das, U. Poddar, D.R. Choudhury **Classification of diabetic patient data using machine learning techniques** Ambient communications and computer systems, Springer (2018), pp. 427-436

[10] A. Negi, V. Jaiswal **A first attempt to develop a diabetes prediction method based on different global datasets** 2016 fourth international conference on parallel, distributed and grid computing, PDGC) (2016), pp. 237-241

[11] N. Murat, E. Dünder, M.A. Cengiz, M.E. Onger **The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels** Biomed Res, 29 (2018), pp. 1370-1375.

[12] M.S. Radin **Pitfalls in hemoglobin A1c measurement: when results may be**

misleading J Gen Intern Med, 29 (2014), pp. 388-394

[13]. Zhou, Bin, et al."**Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants**." The Lancet 387.10027 (2016): 1513-1530.

[14] . Saru, S., and S. Subashree. "**Analysis and prediction of diabetes using machine learning.**" International Journal of Emerging Technology and Innovative Engineering 5.4 (2019). \

[15]. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz. "**COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS**." (2019).

[16]. Sneha, N., and Tarun Gangil. "**Analysis of diabetes mellitus for early prediction using optimal features selection.**" Journal of Big data 6.1 (2019).

[17]. Aada, A., and Sakshi Tiwari. "**Predicting diabetes in medical datasets using machine learning techniques.**" Int. J. Sci. Eng. Res 5.2 (2019).

[18]. Srivastava, Suyash, et al."**Prediction of Diabetes Using Artificial Neural Network Approach.**" Engineering Vibration, Communication and Information Processing. Springer, Singapore, 2019. 679-687

[19]. Kaur, Harleen, and Vinita Kumari. "**Predictive modelling and analytics for diabetes using a machine learning approach.**" Applied computing and informatics (2020).

[20]. Maniruzzaman, Md, et al."**Classification and prediction of diabetes disease using machine learning paradigm.**" Health Information Science and Systems 8.1 (2020).

[21]. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. **A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms**. SN COMPUT. SCI. 1, 101 (2020).

[22] Isha Vashi, Prof. Shailendra Mishra,"**A Comparative Study of Classification Algorithms for Disease Prediction in Health Care**", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

[23] K. Priyadarshini1 , Dr.I.Lakshmi2 "**A Survey on Prediction of Diabetes Using Data Mining Technique**" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 6, Special Issue 11, September 2017.

[24] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput, "**ASurvey on naive Bayes Algorithm for Diabetes Data Set Problems**", International journal for research in Applied Science & Engineering Technology (IJRASET),Volume 3 issue XII,December 2015.

[25] WEBSOURCE: https://www.researchgate.net/publication/273023827**_PREDICTION_OF_DIABETES_MELLITUS_USING_DATA_MININ G_TECHNIQUES_A_REVIEW**

Health Monitoring System: A Research on Diabetes Prediction and Providing Medical Assistance to the Patients Using Big Data and Cloud IoT.

Mr. R.N. Panda
KIET Group of Institutions, Ghaziabad
rn.panda@kiet.edu

Mr. Ankit Verma
KIET Group of Institutions, Ghaziabad

Sakshi Tyagi
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
sakshi.2023mca1049@kiet.edu

Shivani Chauhan
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
shivani.2023mca1008@kiet.edu

Mukul Dhama
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
mukul.2023mca1019@kiet.edu

ABSTRACT:

The world's most deadly disease, Diabetes Mellitus, is among the fastest growing. Medical professionals are looking for a technology that can predict diabetes properly. To study data from diverse perspectives and synthesise it into useful knowledge, various machine

learning algorithms can be applied. With the application of appropriate data mining techniques to large amounts of data, we will be able to obtain valuable information. The main objective is to find new patterns and then provide relevant and helpful information to consumers by analysing them. Diabetes is linked to cardiovascular disease, renal problems, nerve damage, and blindness. Data mining is critical to deal with while doing diabetes research. Methods and strategies for efficiently classifying and detecting patterns in the Diabetes dataset will be discovered using data mining techniques and methods. The purpose of this study was to predict diabetes by using medical bioinformatics. The WEKA programme was used as a diagnostic mining tool. The University of California at Irvine provided the Pima Indian diabetes database, which was utilised in the research. A model based on the data was developed for accurately predicting and diagnosing diabetes. This study compares the performance of Nave Bayes, Decision Trees, and (KNN) with bootstrapping resampling to improve accuracy.

Diabetes has spread throughout the world, affecting people of all ages, regardless of their age. As diabetes patients increase, it is due to numerous factors such as bacterial or viral infections, chemical or toxic substances in food, auto immune reactions, obesity, poor diet, lifestyle changes, eating habits, pollution, etc. To save a person's life, it is essential to identify diabetes in a timely manner. Data analytics is the process of analysing and identifying hidden patterns in large amounts of data to draw conclusions. This analytical procedure is carried out in health care by employing machine learning algorithms to analyse medical data and develop machine learning models for medical diagnostics. This study explains how to diagnose diabetes using a diabetes prediction algorithm. Furthermore, this research investigates several machine learning algorithms and strategies for improving diabetes prediction accuracy utilising medical data. Diabetes Mellitus (DM) is a group of metabolic disorders that afflict millions of people worldwide. Diabetes must be detected early in order to prevent serious complications. There have been numerous research studies on diabetes diagnosis, the majority of them rely on a single data source, the Pima Indian diabetes data set. A collection of studies on Indian women dating back to 1965 has been collected as part of the Pima Indian data set, and it has a relatively high rate of diabetes onset. The majority of prior research studies focused exclusively on one or two specialised sophisticated procedures for data testing, with no comprehensive research on multiple generic techniques. This study explores a number of Machine Learning techniques (e.g., the KNN algorithm) for the purpose of identifying diabetes and methods used for data pre-processing. Testing the accuracy of this technique will utilize the UCI ML repository data set.

Keywords—Machine Learning, Diabetes, Classification, K-nearest neighbours, Decision Trees, Naive Bayes.

1. INTRODUCTION

Diabetes is characterised by elevated blood sugar levels, which manifest as frequent urination thirst, appetite, and weight loss. Diabetes is diagnosed when blood glucose levels exceed 200 mg/dL two hours after loading, as well as numerous diabetes research studies that necessitate fast call detection. Patients with diabetes typically require ongoing therapy; otherwise, they risk a variety of life-threatening complications.

Early detection and treatment of diabetes can help you avoid complications. For diabetes detection, the suggested method

employs machine learning algorithms. The suggested system will be a medical field application that will aid diabetic doctors and patients in recognising diabetes. The suggested approach automates the detection of diabetes using data from previous diabetes patients. People of all ages are affected by diabetes, which is a rapidly spreading disease [1]. The presence of too much sugar (glucose) in the blood causes diabetes. The two types of diabetes are type 1 and type 2. Diabetes type 1 is an autoimmune disease. In this situation, the body destroys the cells required to create insulin and absorb sugar for energy production. Obesity has no bearing on this type of cancer. A person's body mass index defines the obesity which is above his/her normal threshold [2]. Type 1 diabetes can strike at any age, including infancy and puberty. Type 2 diabetes is more likely to develop in obese individuals. The body either opposes or fails to manufacture insulin in this condition. Type 2 is more common in middle-aged and older people [1]. Other causes of diabetes include bacterial or viral infections, toxic or chemical components in food, auto immunity reactions, obesity, inadequate diets, lifestyle changes, exposure to pollution, etc. Diabetes causes a number of ailments, including cardiovascular difficulties, renal problems, retinopathy, and foot ulcers [1]. All over the world, people are affected by diabetes, which is a serious condition. Worldwide, chronic illnesses like this lead to many deaths among adults. Chronic illnesses increase costs as well. A large percentage of government and individual budgets is spent on chronic diseases [3,4]. Diabetes affected 382 million people globally in 2013 [5]. In 2012, it was the eighth biggest cause of mortality for both men and women. Diabetes is more likely to occur in high-income countries [6].

Diabetes diagnosis is regarded as a difficult subject for quantitative research. Due to constraints, the use of numerous criteria such A1c [7], fructosamine, white blood cell count, fibrinogen, and haematological indices [8] are considered ineffective. These criteria were employed in various research investigations to diagnose diabetes [[9], [10], and [11]. Chronic ingestion of booze, salicylates, and drugs have all been linked to an increase in A1C. When measured by electrophoresis, vitamin C consumption may raise A1c levels, but when measured by chromatography, levels may appear to decrease [12].

2.MATERIALS AND METHODS

Following are the two main types in which diabetes can be further divided, i.e, Type-1 Diabetes & Type-2 Diabetes. Insulin-dependent and insulin-independent diabetes are two different types of diabetes.

2.1 TYPE-1

The pancreas is assaulted by the immune system in diabetes type 1. Because type 1 diabetes patients cannot produce their own insulin, they must take insulin injections throughout their lives.

Type 1 diabetes primarily affects children and adolescents, while it can sometimes affect adults. As a result of the immune system attacking pancreatic beta cells, type 1 diabetes causes them to stop producing insulin

There is no way to prevent Type 1 diabetes as it is genetic. The affect Type-1 diabetes have on people with diabetes is just about 5 to 10 percent.

2.2 TYPE-2

Insulin does not act correctly or the body does not produce enough insulin in type 2 diabetes. Diabetes type 1 patients must take medication to keep their blood sugar levels under control. At any age, a person can develop type-2 diabetes.

The risk of type 2 diabetes has increased as people grow older, but it can still

cause serious health problem in children. While insulin is produced by the pancreas, it is not effectively utilized by the body. According to it, lifestyle factors influence its evolution. The vast majority of people with diabetes have type 2 diabetes.
s

3. LITRATURE REVIEW

Diabetes analysis and prediction has become an area of tremendous relevance as the global diabetes rate has increased [13]. The machine learning models were used to The Pima Indian diabetes database by Saru, S., and S. Subashree [14]. With 10 cross-validation runs of their Naive Bayes, Decision Trees, and KNN models, they used bootstrapping resampling to forecast and compare their accuracy. The proposed methodology was determined to have an accuracy of 90.36 percent. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz [15] investigated the accuracy of different data-mining strategies, including decision trees, Naive Bayes, SVM, and hybrid algorithms. With an accuracy of 94% and sensitivity of 91%, hybrid techniques (proposed ensemble SVM + decision tree with 100 iterations) surpassed all other algorithms.

Sneha N. and Tarun Gangil [16] investigated various classification algorithms in order to determine the best diabetes predictor. The study used five classification algorithms: random forest, KNN, decision tree, Naive Bayes, and SVM, which were all obtained from the UCI machine repository archive. The best accuracy was 82.3 percent for Naive Bayes. Aada, A., and Sakshi Tiwari [17] analysed the PIMA Indian diabetes dataset using KNN, Naive Bayes, and decision trees, as well as bootstrapping-like approaches. After bootstrapping, SVM has the highest accuracy of 94.44 percent. On the Pima Indians dataset, Srivastava and Suyash [18] used machine learning algorithms and artificial neural networks to predict diabetes.

The PIMA Indian diabetes dataset was used in a study by Kaur, Harleen, and Vinita Kumari [19] to predict and assess diabetes trends. For prediction, they employed R data processing software and five algorithms: SVM-linear, radial basis function kernel support vector machine, k-nearest neighbor, artificial neural network, and multi-factor dimensions reduction. For diabetes prediction, the SVM-linear model had the highest accuracy of 89 percent. Md Maniruzzaman used the dataset of diabetes from the survey done by the National Health and Nutrition Examination [20], which included 6561 people, 657 of whom were diabetic. For diabetes prediction, logistic regression, Naive Bayes, decision trees, AdaBoost, and random forest were used. With logistic regression as feature selection and random forest for classification, the greatest accuracy of 94.25 percent was achieved.

Naive Bayes, decision tree, J48, and random forest with 10 fold cross-validation were used by Prasad, K.S., Reddy, N.C.S., and Puneeth, B.N. [21]. The most accurate method was Random Forest, whereas Naive Bayes had the lowest mean absolute error and root mean squared error.

4. EXISTING DATA MINING TECHNIQUES

4.1 Decision Tree:

The most widely used data mining technique is the decision tree. One of the most popular and straightforward classifiers is the decision tree. Data mining uses decision trees as predictive

models. In this study, a decision tree is used with a classification algorithm to predict disease using patient data.

Decision trees are simple to design and understand. Decision tree-based prediction is well-organized. It is capable of handling large amounts of extent data. It is better suited to knowledge discovery searches. Finally, the Decision Tree findings are easier to understand and read [22].

4.2 Naive Bayes:

The Naive Bayes algorithm is based on the Bayes theorem and involves conjecture between predictors. The Naive Bayes allows you to easily develop models with predictive capabilities as well as a new way of navigating and analysing data. When using Naive Bayes to develop a predictive model, this technique can be applied to predictive analysis. All input attributes must be comparably independent for this to work. A Naive Bayesian model is simple to build and does not have any complex iterative parameters. Naive Bayes is an effective predictor. For really big data sets, this strategy is quite beneficial. Before and after reducing of characteristics, Naive Bayes performs consistently with the same representation of construction time. The visualisations for Naive Bayes are straightforward [23].

4.3 k-nearest neighbour algorithm (k-NN):

The k-nearest neighbour algorithm (k-NN) is an important approach to classifying objects based on the nearest training data in the feature space. Although it is the simplest of all machine learning rules, the accuracy of the K-NN rule is harmed by the availability of screaming choices.

5. Classification

5.1 Clustering:

Diabetes can be predicted using classification as one of the methods. Classification is the most important aspect of data mining. Classification is common in large amounts of corporate and medical data sets. Classification is a data mining function that divides a collection of items into desired categories. Classification is the process of dividing situations into groups based on a predictable characteristic. Each case has a number of attributes, one of which is the class attribute, also known as the predictable attribute. Finding a model that defines the class attribute as a function of input attributes is part of the work. Classification measures with data mining tools with detecting the problem by separating the aspects of diseases between patients and diagnosing or predicting which algorithm performs best [24]. Data is classified in order to determine the class that each case belongs to accurately. Before storing the data into the classifying model, this technique can be utilised as a pre-processing step. To locate a set of data, we must cluster them in order to collect everything based on their features. And grouping them together based on their similarities. Segmentation is another term for clustering. It's used to find unprocessed case groupings based on a set of criteria. Cases with more or less similar attribute values inside the same category. Clustering is a data mining activity that is done individually. It means that by using a single attribute you cannot model the training process. All input attributes are given equal weight. Before clustering, the attribute values must be defined to prevent higher value characteristics from impacting lower value attributes.

5.2 Neural Network:

Artificial neural networks are well adapted to issues that human are not adept at addressing, such as pattern recognition and prediction. In the medical field, neural networks have been

applied to diagnosis, picture interpretation, signal interpretation, and drug development [25].

6. METHODOLOGY

The mechanized diabetes testing system is, which is used to consistently record the blood glucose levels of diabetics, especially in the intensive care unit. The observation frame records glucose levels via a blood glucose sensor. Information will be sent to the Specialist via WIFI. This frame is the moment, which indicates the blood glucose level regardless of whether glucose is falling or rising. The framework provides accurate results and patient information is updated daily in the cloud.

ARDUINO UNO ATMEGA 328P:

Arduino microcontroller is something however tough to utilize, open supply, and its device is smart period. The Arduino ATMEGA 328 is a well-known microcontroller chip created with the aid of using Atmel. It is an 8-piece microcontroller that has 32K of glimmer memory, 1K EEPRON and 2K of indoors SRAM of it has 14 superior information/yield pins wherein 6 may be applied as PWN yields and 16MHz earthenware resonator, an ICSP header, the USB association, 6 easy statistics sources, a force.

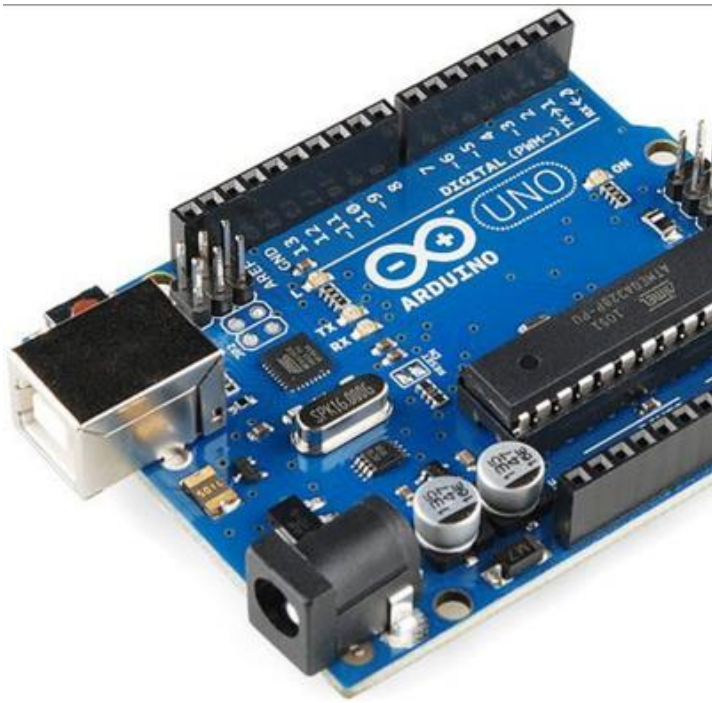


Figure 1: ARDUINO UNO ATMEGA 328

7. OUTCOME

7.1 HUMIDITY IN ADDITION TO TEMPERATURE SENSOR:

This is a focal temperature and viscosity sensor-driven with very little effort. Capacitive humidity sensors and thermistors are used to check for trapped air. Then it outputs a screen signal that passes through the material pin. BMP 280 Pressure Sensor: This sensor has high accuracy and ease of use, making it the perfect solution for accurate pressure estimation. Weight varies with height and the weight estimate is very accurate.

7.2 BMP 280 PRESSURE SENSOR:

With its high accuracy and ease of use, this sensor is the perfect solution for accurate pressure estimation. Weight varies with height and the weight estimate is very accurate.

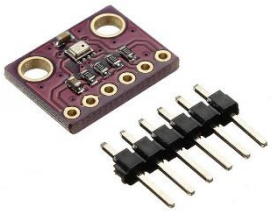


Figure 2: BMP 280 PRESSURE SENSOR

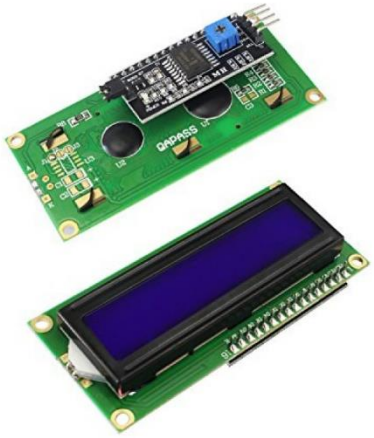


Figure 3: LCD DISPLAY

7.3 GAS SENSOR (FIGARO TGS822):

It is a electricity display case module that uses Liquid Gem to provide a clear image to the. This is usually a required module Used in DIY and circuits. 16 * 2 uses a platform with 16 fonts for each of these two lines.

Figure 4: GAS SENSOR (FIGARO TGS822)

8. WORKING:

The gas sensor detects CH3) 2CO from the respiration of the human body. This is sent to the data distribution unit of the effort module where the Arduino is located. By, Arduino yields will be proven to LCD or distant zone authorities. The BMPP 280 Weight Sensor measures temperature, weight, and humidity. The sensor gets the yield because the simple sign is the yield. This yield was switched to automation and was displayed on the LCD using the Arduino ATMEGA328P.

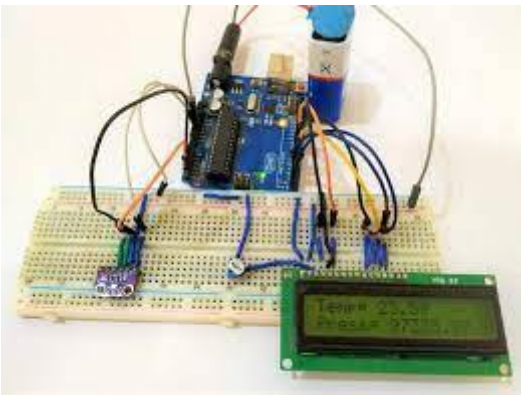


Figure 5: DISPLAYING THE PROJECT MODEL

9. CONCLUSION

The purpose of this paper is to examine symptoms of diabetes and gather information that can be used to assist healthcare professionals in detecting diabetes early and detecting its onset before it becomes a major problem. Data mining uses techniques such as feature selection to analyse data. Classification. All of these are used to analyse trends and forecast symptom severity. diabetes. ANOVA, Mutual Information, and other feature selection techniques to improve accuracy and reduce overhead, genetic algorithms were applied. The model's training time. Naive Bayes, SGD Classifier, KNN, Logistic Regression the Algorithms Random Forest, Decision Tree, and Support Vector Machine were used to diabetes prediction. A comparison of all the implemented algorithms was conducted by Random Forest had the highest accuracy of 93.95 percent when using Genetic Algorithm as a feature selection technique, with Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip as features and a random forest depth of 5. The ability to predict diabetes at an early stage depends on regular monitoring of patients' cholesterol, glucose, cholesterol/HDL, systolic blood pressure, weight, and hip. Diabetic complications can be prevented by early treatment of any abnormality in a number of the following measurements.

In today's world, diabetes prediction is critical, especially given its serious complications. Nowadays most people die due to diabetes. The System model focuses on detecting diabetes using only a few parameters. Physicians can utilise the system to anticipate diabetes in the early stages. So that patients can receive traditional treatments and remedies. The system used techniques like machine learning (ML) for prediction to give more exact results. There has been a lot of research regarding the diabetic imprint. It is important for hospitals and clinics to develop a prediction system to predict diabetes.

REFRENCES

[1] Kaveeshwar, S.A., and Cornwall, J., 2014, "**The current state of diabetes mellitus in India**". AMJ, 7(1), pp. 45-48.

[2] Dean, L., McEntyre, J., 2004, "**The Genetic Landscape of Diabetes [Internet].**
Bethesda (MD): National Center for Biotechnology Information (US);.
Chapter 1, Introduction to Diabetes. 2004 Jul 7.

[3] D. Falvo, B.E. Holland **Medical and psychosocial aspects of chronic illness and disability Jones & Bartlett Learning** (2017)

[4] J.S. Skyler, G.L. Bakris, E. Bonifacio, T . Darsow, R.H. Eckel, L. Groop, *et al.* **Differentiation of diabetes by pathophysiology, natural history, and prognosis**
Diabetes, 66 (2017), pp. 241-255

[5] Z. Tao, A. Shi, J. Zhao **Epidemiological perspectives of diabetes** Cell Biochem Biophys, 73 (2015), pp. 181-185

[6] W.H. Organization **World health statistics 2016: monitoring health for the SDGs sustainable development goals** World Health Organization (2016)

[7] L. Cobos **Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat)** Endocr Pract, 24 (2018), pp. 43-44

[8] B. Dorcely, K. Katz, R. Jagannathan, S. S. Chiang, B. Oluwadare, I.J. Goldberg,

et al.
Novel biomarkers for prediabetes, diabetes, and associated complications Diabetes, Metab Syndrome Obes Targets Ther, 10 (2017), p. 345

[9]
P.P. Singh, S. Prasad, B. Das, U. Poddar , D.R. Choudhury **Classification of diabetic patient data using machine learning techniques** Ambient communications and computer systems, Springer (2018), pp. 427-436

[10] A. Negi, V. Jaiswal **A first attempt to develop a diabetes prediction method based on different global datasets** 2016 fourth international conference on parallel, distributed and grid computing, PDGC) (2016), pp. 237-241

[11]
N. Murat, E. Dünder, M.A. Cengiz, M.E . Onger **The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels** Biomed Res, 29 (2018), pp. 1370-1375.

[12] M.S. Radin **Pitfalls in hemoglobin A1c measurement: when results may be misleading** J Gen Intern Med, 29 (2014), pp. 388-394.

[13]. Zhou, Bin, et al."**Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants**." The Lancet 387.10027 (2016): 1513-1530.
[14] . Saru, S., and S. Subashree. "**Analysis and prediction of diabetes using machine learning.**" International Journal of Emerging Technology and Innovative Engineering 5.4 (2019). \
[15]. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz. "**COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS**." (2019).
[16]. Sneha, N., and Tarun Gangil. "**Analysis of diabetes mellitus for early prediction using optimal features selection.**" Journal of Big data 6.1 (2019).
[17]. Aada, A., and Sakshi Tiwari. "**Predicting diabetes in medical datasets using machine learning techniques.**" Int. J. Sci. Eng. Res 5.2 (2019).
[18]. Srivastava, Suyash, et al."**Prediction of Diabetes Using Artificial Neural Network Approach.**" Engineering Vibration, Communication and Information Processing. Springer, Singapore, 2019. 679-687
[19]. Kaur, Harleen, and Vinita Kumari. "**Predictive modelling and analytics for diabetes using a machine learning approach.**" Applied computing and informatics (2020).
[20]. Maniruzzaman, Md, et al."**Classification and prediction of diabetes disease using machine learning paradigm.**" Health Information Science and Systems 8.1 (2020).
[21]. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. **A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms**. SN COMPUT. SCI. 1, 101 (2020).
[22] Isha Vashi, Prof. Shailendra Mishra,"**A Comparative Study of Classification Algorithms for Disease Prediction in Health Care**", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.
[23] K. Priyadarshini1 , Dr.I.Lakshmi2 "**A Survey on Prediction of Diabetes Using Data Mining Technique**" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007

Certified Organization) Vol. 6, Special Issue 11, September 2017.
[24] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput, "**ASurvey on naive Bayes Algorithm for Diabetes Data Set Problems**", International journal for research in Applied Science & Engineering Technology (IJRASET),Volume 3 issue XII,December 2015.
[25] WEBSOURCE: https://www.researchgate.net/publication/273023827**_PREDICTION_OF_DIABETES_MELLITUS_USING_DATA_MINING_TECHNIQUES_A_REVIEW**

Health Monitoring System: A Review on Diabetes Prediction and Providing Medical Assistance to the Patients Using Big Data and Cloud IoT.

Sakshi Tyagi
Student
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
sakshi.2023mca1049@kiet.edu

Shivani Chauhan
Student
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
shivani.2023mca1008@kiet.edu

Mukul Dhama
Student
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
mukul.2023mca1019@kiet.edu

Abstract

We are conducting a scoping review of the academic literature on Big Data and diabetes care in order to review current Big Data applications in diabetes care and consider future potential. Data is being generated at an increasing rate in the healthcare industry, and this data has the potential to revolutionize the way diabetic care is delivered. Through data study, Diabetes care is beginning to be impacted by big data. Big Data's applicability in normal health treatment is yet in the future. Massive volumes of healthcare data are already being generated; the trick is to figure out how to use it to generate actionable insights. These objectives will necessitate a significant amount of development work.
Predictive analytics for diabetics are very important as it helps diabetics, their families, doctors, and medical researchers make decisions about diabetics based on a high amount of data. Diabetes is a condition that occurs when the body cannot use glucose normally. Glucose is the usual source of Body Cell Essence. Blood levels of glucose are caused by a hormone called insulin produced by the pancreas. There are basically two types of diabetes. Patients with type 1 diabetes and patients with type 2 diabetes. In patients with type 1 diabetes, the pancreas is persistently unable to produce insulin, and in patients with type 2 diabetes, the pancreas produces insulin and does not carry sufficient insulin or does not function. Today, type 1 diabetes is a widespread and prominent clinical problem. Current methods use methods other than snooping, and patient-tolerant data is typically sent to specialists using the IoT. Therefore, the proposed approach captures the blood glucose levels of diabetics, especially for robotized diabetics who check the structure in the ICU.

1. Introduction

Available clinical records show that type 1 diabetes is a notable clinical problem worldwide. There are about 2.6 million adults over the age of 18 suffering from diabetes, and the severity of diabetes should increase in Malaysia. Ketones are artificial substances that appear in the body when the muscle-to-fat ratio is used, as opposed to glucose due to urgency. This indicates that the body's cells cannot absorb enough sugar (glucose) from the blood, especially if the body's insulin is too low. Insulin is used by the body to use glucose, which focuses on what is important. In this sense, which screens ketone zookeepers one by one, helps control and screen the condition of diabetics with a vast number of ketones that remain wild in diabetics. Anyway, the two steps were considered intrusive, horrifying, and anomalous. CH3) 2CO is abstractly recognized by the method of diabetes biomarkers. CH3) 2CO stands for traditional stuff and knows the technology for Ketobetix and the fragrant scent of breathing in diabetics. The combination of breath CH3) 2CO is associated with glucose absorption and lipolysis. This is the method by which attachment to CH3) 2CO respiration is presented in the elevated modern form of diabetes and can be used to study the development of diabetes is called the easiest way and diabetes. Prevents rapid detection of sex ketoacidosis Type 1 diabetes. A method of checking ketone levels is performed using breath estimation. Atem presents a simple portable. Today, there are several chronic illnesses such as heart disease, stroke, cancer, chronic respiratory illness, and diabetes. It is a dangerous disease and has recently become one of the leading causes of death worldwide and requires careful monitoring to maintain patient health. The biggest challenge for diabetics is to raise or lower blood glucose within a specific interval, as diabetes is caused by insulin resistance and inadequate insulin production can lead to level up or level down of blood glucose. If these conditions are not met, some patients may need urgent medical care to avoid exacerbations [1] human management for checking diabetes levels. The strategy presented demonstrated advances in the hardware relationship with the Internet of Things (IoT) system to enhance patient evaluation and individual observation methods. The Arduino board will be used to study the sensor with breath detection capability. Breath observation levels are recorded on the system using distant correspondence. Data collection is linked to the site page. Ketone levels are assessed as the percentage of CH3) 2CO in exhaled breath accumulated when the patient breathes into a mouthpiece containing a gas sensor. This assessment is based on detecting the patient's blood glucose level by separating CH3) 2CO levels from respiration at and sending data to clinical authorities by the WIFI method via a message. These devices are a new way of continuous monitoring. They provide real-time information about blood glucose levels. This article introduces an intelligent diabetes monitoring system using the node MCU and machine learning algorithms. The MCU node is connected to a glucose meter and periodically records the blood glucose levels of diabetics. This collected data can be used by caregivers (patients, researchers, and doctors) to remotely monitor patients. As a result, patients and physicians alike need to process multiple records, interpret vast amounts of data, adjust insulin doses, and bring blood glucose levels as close to normal as possible

Diabetes is a life-threatening disease that has no cure. If you get this sickness once, it will be with you for the rest of your life. At the same time, having too much glucose in your blood might cause health problems. Kidney illness, heart disease, stroke, vision problems, dental problems, foot problems, and nerve damage are just a few examples. so that you can keep track of your diabetes and avoid complications

Kind 1 diabetes is the most well-known type of diabetes.

Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type 2 Diabetes Type

Type 1 diabetes occurs when the body is unable to manufacture insulin. It has a negative impact on children and young adults. It can also affect people of any age. People with this kind of diabetes must take insulin on a daily basis.
Type 2 diabetes occurs when the body is unable to manufacture or utilise insulin.
This kind of diabetes primarily affects people in their forties and fifties.
Diabetes During Pregnancy
This kind of diabetes primarily affects women. During pregnancy, this kind of diabetes develops. High blood sugar levels caused by gestational diabetes can harm your pregnancy and your baby's health.

A CHALLENGES

Diabetes can also lead to eyesight issues. It lowers blood glucose levels in the retina of diabetics who are older. It causes cataracts in diabetics in the future, and it causes bad vision quite readily. Patients' vision problems cause them a lot of trouble and interfere with their regular activities.
Hearing loss was induced by diabetes. Long-term high blood glucose levels can disrupt the delivery of blood and oxygen to the inner ear's tiny nerves and blood vessels, resulting in hearing loss. The nerves and blood arteries in the ear become damaged with time, limiting the person's ability to hear. It causes misunderstandings between people. As a result, every diabetic patient should assess their hearing ability.
Diabetic patients are physically weak and have a poor quality of life. As a result, diabetics must raise and maintain their daily activity levels. Controlling one's eating habits is crucial, as is physical activity. Diabetic patients must be encouraged to exercise on a daily basis. Brisk walking, bicycling, swimming, housework, and gardening are just a few examples.

2. Big Data and the Future of Diabetes

Diabetes affects 9% of the population in the United States, according to the Centres for Disease Control and Prevention. The demand for useful information on how to treat and manage the disease is greater than ever. A wide approach to data analysis can aid healthcare providers in gaining a better understanding of the condition, its prognoses, and its complications.
Great amounts of data are required for big data analysis, and persons with diabetes generate a large quantity of data just by going about their regular lives. Wearable exercise monitors, smart blood pressure cuffs, Bluetooth-enabled bathroom scales, and smart insulin pens all generate data.
For example, the FreeStyle Libre flash glucose monitoring system makes it simple for persons with diabetes to check their glucose levels in real-time.
Three basic types of big-data oriented solutions are presently accessible and progressively accepted from an IT standpoint. To begin with, cloud computing offers cost-effective ways to achieve great computational performance. **[2]** Second, parallel programming is

becoming increasingly simple and efficient: "MapReduce," a programming model that allows algorithms to be implemented in distributed contexts, is becoming a very popular and frequently used paradigm. **[3]** Finally, new database technologies, such as No-SQL databases, are now available to address both the scalability and variety issues. **[4**] Better healthcare is explained by various big data technology stacks and research into health care mixed with efficiency, cost savings, and other factors. The use of Hadoop in health care has become more essential as a means of processing data and doing large-scale data management tasks. Hadoop's cost effectiveness can be improved by employing analytics on coupled compute and storage **[5].**

3. Cloud computing with IoT architecture in prediction of diabetes

The Internet of Things (IoT) and mobile health care (m-healthcare) applications provide numerous dimensionalities and online services. These applications have offered a new platform for millions of individuals to benefit from health suggestions on a regular basis in order to live a healthy life. The numerous aspects of these healthcare online applications were reinforced after the introduction of IoT technology and related devices that are employed in the medical industry. IoT devices in the healthcare context generate a massive amount of big data. Cloud computing technology is used to manage massive amounts of data while also being simple to use. Cloud-based applications are playing an increasingly important role in today's fast-paced environment.

For safe storage and access, these medical apps make advantage of Cloud Computing technologies. We propose a novel Cloud and IoT-based Mobile Health Care application for monitoring and detecting critical diseases in order to provide better services to people through online healthcare applications. A new framework for the general public is being established here. In this study, a new systematic strategy for diabetes disorders and related medical data is developed using the UCI Repository dataset and medical sensors in order to forecast persons who are seriously impacted by diabetes. In addition, for detecting the disease and its severity, we present a new classification technique called Fuzzy Rule-based Neural Classifier. The studies were carried out using the standard UCI Repository dataset as well as real health records obtained from various hospitals. The experimental results suggest that the proposed approach beats existing disease prediction systems in terms of performance.

4. Predictive analytics system for Diabetes data

Diabetes mellitus is one of the most common noncommunicable diseases today, with a significant impact on human life. It is currently regarded as one of the worst diseases on the planet. If diabetes is not addressed, it can lead to a variety of health problems.The healthcare business collects a massive amount of data, much of which is unstructured. The information must be organised into basic values. Medical intelligence systems will be created by applying computer analytics to the huge amounts of data collected in the healthcare system, which will aid medical prediction. This will result in a patient-centered healthcare system that lowers medical costs. In health care, predictive analytics is largely used to identify individuals who are in the early stages of diabetes, asthma, heart disease, and other serious life-threatening diseases. To predict type 2 diabetes, the suggested technique PDD employs data mining algorithms.

5. Data collection

The system receives the raw diabetes large data or data set as input. The unstructured voluminous input data can be obtained from a variety of Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems, and external sources (government sources, laboratories, pharmacies, insurance companies, and so on), in a variety of formats (flat files,.csv, tables, ASCII/text, and so on), and from a variety of locations **[6].**

6. Cloud-based patient profile analytics system

The computer or system is educated with clinical data obtained from healthcare organisations in order to forecast and analyse patient profiles. Sensors are attached to persons in rural regions to monitor their health on a monthly or weekly basis, and these sensors use IoT devices to detect their health **[7].** The IoT gadget, on the other hand, is linked to the service provider (hospitals, doctors, clinics, etc.). Basically, the cloud-based data structure has many modelling systems, thus the cloud structure components have saved clinical or medical profile data for each and every patient, which are then trained to the machine throughout the machine learning process **[8].** Patient profiles, which include patient health graphs, bio data, and other information.2

The research of a TB patient's body condition is more crucial in this criterion in order to examine severity measurements and supply medical instructions **[9].** Several strategies, such as feature selection and various segmentation processes, should be followed during the data analytic process. In addition, the analytic approach could comprise classification, regression, feature learning, and other prediction methods.

Cloud computing has become one of the most often used expressions to describe a type of on-demand computer service provided by companies such as Amazon, Google, and Microsoft. It is a virtual model on which computations are done and services are provided to consumers, hence the name "cloud." Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service are three main types of cloud services offered by cloud providers (SaaS). The notion of virtualization, which allows virtual machines (VM) to run on top of existing hardware to meet the needs and demands of users, lies at the heart of cloud computing **[10**]. In hospitals, cloud computing opens up new opportunities for exchanging information with patients about their health, treatment alternatives, and the treatment process. The usage of IT is influenced by continuous change in both supply and demand in the healthcare sector, and this is the primary driver for cloud computing adoption.Diabetes is a disease that affects a large number of individuals nowadays. Diabetes is caused by a shortage of insulin in the blood. Diabetes Mellitus **[11]** is the type of diabetes that is the most common.

7.Limitation and Strength

There is no doubt that big data has had a positive impact on the healthcare system. The main problem with unstructured and structured data in healthcare is a lack of data, which makes disease prediction and symptom analysis difficult. However, the data contained in the healthcare system is both uneven and balanced. Furthermore, healthcare patient profile data is maintained in legacy systems (often electronic medical record systems) that have limited interoperability. However, incorporating health information into big data is difficult, as there are numerous formats, metadata, schema, and standards that affect the data. Moral issues such as data privacy, discretion, and control of access to patients, as well as the commercialization of deidentified patients' data; and the ownership and control of patients' data are all factors that stymie the effective exchange of profile details between patients and healthcare providers. As a result, combining healthcare data from several sources becomes difficult. As a result, getting a detailed and complete image of a patient in a timely manner throughout care becomes an issue. Standards are agreed-upon

specifications that allow disparate systems, tools, equipment, and platforms to communicate with each other. Healthcare organisations, on the other hand, do not all follow the same pattern.Case reports, medications, disorders, and examinations, for example, have distinct titles and codes in different hospitals. Healthcare data should be sufficiently protected and secured, according to healthcare providers and big data analytics developers. When studying big data in healthcare, tools that assure the confidentiality, integrity, and availability of protected health information should be employed. Physical security, data encryption, user authentication, and application security should all be used to protect healthcare data. It's also a good idea to encourage the adoption of audit trail systems.

8. Diabetes data analytics tools and techniques

A variety of techniques and approaches are utilised to assess patient data and explore the symptoms that accompany it. IoT with diabetetic systems is a hot topic in big data analytics right now, and AI is also being used in prediction of diabetes for machine learning purposes. . A description of the proposed structures and the used algorithms on these paintings are given. In their paintings, the authors supplied a smart structure for the surveillance of diabetic ailment that displays the fitness of diabetic sufferers through sensors incorporated into smartphones [12]. In some other paintings, the prediction of diabetes sorts the use of evaluation algorithms and Hadoop map-lessen, prediction of complications, and the prediction of the kind of remedy had been investigated [13], at the same time as in [14], the authors proposed a machine which could carry out predictions for Mellitus that's a form of diabetes the use of Hadoop/MapReduce. A new machine for the prediction of glucose awareness has been proposed in. Where the facts generated via way of means of the Continuous Glucose Monitoring may be analysed via way of means of the glucoSim software program the use of the Kalman Filter (KF) to reducing noise. Many researchers have used different data mining techniques to develop and implement different analytical and predictive models. In [15], the author found a pattern in the Diabetes dataset using the classification method by Naive Bayes and the decision tree algorithm using the Weka tool. In [16], the author uses naive Bayes and decision trees in a model classification technique to examine the hidden patterns in the diabetes dataset. The authors of [17] predicted patient diabetes using the C4.5 decision tree algorithm, Neural network algorithm, Kmeans clustering algorithm, and visualizations.

8.1. Support Vector Machine

SVM stands for supervised knowledge representation model. Furthermore, it is the associated algorithm model that is used to study patient data for regression and classifications.

A Support Vector Machine creates a hyperplane or group of hyperplanes in high layered space, maps all of the models in a guide, and divides the examples by an identifiable hole about as broad as could be expected, with each side presenting one class. We should adjust the regularisation boundary C, which determines the model's complexity, in this strategy. Greater C indicates that the misclassification will be punished more severely, implying that the model will be more likely to overfit. The SVM classifier is used to divide the data into a particular number of categories. It is quite tough to use AI and data mining to each and every exploration investigation in order to deconstruct diabetes. We'll deconstruct multiple approaches and apply them to the dataset. We will make every effort to achieve the best possible result. The present enhancement focuses on increasing characterisation

precision while reducing execution time.

8.2. Decision Tree

The supervised learning approach, which is used to solve problems with arrangements. A decision tree is a mechanism that iteratively divides a given dataset into at least two examples. The technique's goal is to predict the class worth of the objective variable. The decision tree will assist in separating the informative collection and constructing the choice model in order to predict the obscure class marks. A decision tree can be built for both parallel and continuous factors. The root hub is ideally observed by a decision tree based on the most significant entropy value. As a result, choice tree has the advantage of selecting the most stable theory from the preparation dataset. A dataset including a few credits and occasions esteems is a contribution to the Decision tree, and the result is the choice model. Choosing the parting characteristic, parts, halting rules, and pruning, as well as preparing tests, quality and quantity, and the request for parts, are all issues that must be addressed while developing a choice model.

8.3. Naive Bayes (NB)

Prior research has shown that Induction algorithms based on Naive-Bayes have shown surprising accuracy in many classification tasks even when the assumption of conditional independence on which they are founded is broken. The majority of studies, however, were conducted on small databases. We demonstrate this in certain larger databases, Naive-Bayes accuracy does not scale up as well as decision trees. Following that, we suggest a new algorithm, NBTree, which creates a hybrid of decision-tree and Naive Bayes classifiers: As with ordinary decision trees, the nodes of the decision tree include univariate splits, whereas leaves contain Naive-Bayesian classifiers. The approach combines the interpretability of Naive-Bayes and decision trees with the performance of classifiers that outperform both components, especially when applied to large databases.

9.Solution foreseen

Prior to implementing algorithms in patient risk prediction, it is necessary to examine previously existing algorithms, approaches, and events. The use of Cloud IoT in the health-care system is an effective technique of predicting and assessing high risk. The suggested health-care disease prediction system combines traditional health informatics with cloud IoT platform-based big data analytics. The disease in patients can be sensed by the patients themselves, and the information about the diseases is sent to the hospital's doctors, who study the data and then assess the patients based on their illnesses.

10. Structures includes in Diabetes analytics

Several models in the domains of healthcare informatics and big data analytics are described in this study. Table3.1 summarises the classification of these platforms, including all tools and approaches utilised in the comparison.

Author	**Source and Year**	**Disease Diagnoses method**	**Outcome**
Z. Mian *et al* [1]	**Source:** Continuous Glucose Monitoring: Review of an Innovation in Diabetes Management. **Year:** 2019	Continuous glucose monitoring and sensor-enabled pump technology are used.	This technology eliminates the need for frequent blood glucose monitoring, which is often inconvenient for patients, and instead offers them a more convenient option.
Lin *et al* [2]	**Source:** Enabling large-scale biomedical analysis in the cloud. **Year:** 2013	Explains the data-intensive computing system and lists available cloud-based bioinformatics resources.	To make a large amount of variety of data understandable and usable for biomedical research, we need to make it easier.
Mohammed *et al* [3]	**Source:** Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. **Year:** 2014	The Hadoop platform and the MapReduce programming framework's potential applications.	To process large amounts of clinical data in sectors connected to medical health informatics
Lee *et al* [4]	**Source:** Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage **Year:** 2013	The feasibility of three database technologies - NoSQL, XML-enabled, and native XML - for structured clinical data is evaluated.	The greatest choice for query performance is a NoSQL database, although XML databases are better in terms of scalability,

			flexibility, and extensibility, all of which are necessary to deal with the features of clinical data.
D. Peter Augustine [5]	**Source:** Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services. **Year:**2014	Analyze and demonstrate the benefits of Big Data Analytics and Hadoop in healthcare applications.	The use of Big Data Analytics and Hadoop illustrates the significance of these technologies in providing healthcare services to everyone at the lowest possible cost.
Wullianallur Raghupathi *et al* [6]	**Source:** Big data analytics in healthcare: promise and potential. **Year:**2014	To describe big data analytics' promise and potential in healthcare.	For healthcare academics and practitioners, it provides a wide understanding of big data analytics.
Kalet *et al* [8]	**Source:** Quality assurance tasks and tools: The many roles of machine learning. **Year:**2019	Describe some study topics and some of the particular obstacles each one faces.	Improvements in planning time, plan quality, advanced dosimetric QA, predictive machine maintenance, higher safety checks, and advancements are all important for new adaptive planning-driven QA paradigms.

Pande, Tripti, *et al* [9]	**Source:** Prevalence of diabetes mellitus amongst hospitalized tuberculosis patients at an Indian tertiary care center: A descriptive analysis. **Year:**2018	Calculate the prevalence of Diabetes Mellitus in adulthood.	Age, type of TB, and undernutrition were all found to be significant predictors of TB-DM co-prevalence.
Ahmad M. Manasrah [10]	**Source:** A Variable Service Broker Routing Policy for data center selection in cloud analyst. **Year:**2017	Variable Service Broker Routing Policy is used to reduce the processing and response time of customer requests while staying within a reasonable cost range.	The proposed policy alters the old policy's sorting and selection equations.
A. Rghioui *et al* [11]	**Source:** A Smart Glucose Monitoring System for Diabetic Patients. **Year:** 2020	A compact portable device capable of detecting blood glucose levels and body temperature in diabetics.	Diabetes disease surveillance would allow doctors to remotely monitor their patients' health using sensors included in smartphones and smart portable devices.
N. M. S. Kumar *et al* [12]	**Source:** Predictive Methodology for Diabetic Data Analysis in Big Data. **Year:** 2015	Hadoop/predictive MapReduce's analytical method Reduce the environment to forecast the types of diabetes that are common and the complications that come with it.	This approach enables patients to be cured and cared for in a more efficient manner, with improved outcomes like as affordability and accessibility.
Ahmed, H.B *et al* [13]	**Source:** Effects of External Factors in CGM Sensor Glucose Concentration Prediction. **Year:**2016	To develop a blood glucose prediction system for usage in conjunction with a continuous glucose monitoring (CGM) device.	This technology eliminates the need for frequent blood glucose monitoring,

			which is often inconvenient for patients, and instead offers them a more convenient option.
Dr Saravana Kumar N M *et al* [14]	**Source:** Predictive Methodology for Diabetic Data Analysis in Big Data. **Year:** 2015	Hadoop/predictive MapReduce's analytical method Reduce the environment to forecast the types of diabetes that are common and the complications that come with it.	This approach enables patients to be cured and cared for in a more efficient manner, with improved outcomes like as affordability and accessibility.
Dost Muhammad Khan1 *et al* [15]	**Source:** An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm. **Year:** 2011	To create a more efficient data mining approach employing intelligent agents, we combine the K-means clustering algorithm with the Decision tree (ID3) algorithm.	Data mining algorithms are used to uncover hidden patterns and connections across variables in large datasets.
Dost Muhammad Khan1 *et al* [16]	**Source:** An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm. **Year:**2011	Integrating K-means clustering algorithm with Decision tree (ID3) algorithm.	To develop a more efficient data mining technique that makes use of an intelligent agent called Learning Intelligent Agent (LIAgent), which can conduct classification, grouping, and interpretation tasks on datasets.

Deepti Sisodia *et al* [17]	**Source:** Prediction of Diabetes Using Classification Algorithm. **Year:**2018	The goal is to create a model that can accurately predict the likelihood of diabetes in people.	The designed system can be used to predict or diagnose diabetes using machine learning classification methods.

11.Conclusion

This research looked at machine learning classification techniques for better diabetes illness prediction. The accuracy of this literature analysis in the SVM classification technique was the greatest. Different measures are used to compute the various performance values of categorization algorithms. Pima Indians Diabetes Dataset can be used to train and test data. The categorization algorithm reached the highest level of testing precision. This research gathered a variety of categorization techniques and combined them to improve accuracy, specificity, and sensitivity.

Using elastic net regression, the challenges that are studied in the previous works for enhancing accuracy for diabetes prediction and diagnosis will be worked out further. Elastic net regression is a hybrid of LASSO and Ridged Regression approaches that allows data in category, numerical, and picture form to be fed into the regression. The AdaBoost classifier was found to be the best model, with an accuracy of 98.8%. With two separate datasets, we compared the accuracies of machine learning algorithms. When compared to the existing dataset, it is obvious that the model enhances the accuracy and precision of diabetes prediction. This research could be expanded to see how likely non-diabetic persons are to develop diabetes in the coming years.

References:

[1] Z. Mian, K. L.Hermayer, A. Jenkins, "Continuous Glucose Monitoring: Review of an Innovation in Diabetes Management", The American Journal of the Medical Sciences Vol. 358, pp: 332-339, Issue 5, November 2019

[2] Lin, YC, Yu, CS, Lin, YJ. Enabling large-scale biomedical analysis in the cloud. Biomed Res Int. 2013;2013:185679.

[3]

Mohammed, EA, Far, BH, Naugler, C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. 2014;7:22.

[4]

Lee, KK, Tang, WC, Choi, KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed. 2013;110(1):99-109.

[5]. D. Peter Augustine, "Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services", International Journal of Computer Applications, vol 89(16), pp 44-50, 2014.

[6]. Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2(3) pp. 2-10, 2014.

[7] Liu, Ruiqing, et al. "Profile of Consecutive Fecal Calprotectin Levels in the Perioperative Period and Its Predictive Capacity for Early Endoscopic Recurrence in Crohn's Disease." Diseases of the Colon & Rectum 62.3 (2019): 318-326.

[8] Kalet, Alan M., Samuel MH Luk, and Mark H. Phillips. "Quality assurance tasks and tools: The many roles of machine learning." Medical physics (2019).

[9] Pande, Tripti, et al. "Prevalence of diabetes mellitus amongst hospitalized tuberculosis patients at an Indian tertiary care center: A descriptive analysis." PloS one 13.7 (2018): e0200838.

[10] Ahmad M. Manasrah, "A Variable Service Broker Routing Policy for data center selection in cloud analyst", Journal of King Saud University – Computer and Information Sciences 29, 365–377, 2017.

[11] P. Suresh Kumar, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Vol 7, June 2017.

[12] A. Rghioui, J. Lloret, M. Harrane, A. Oumnad. "A Smart Glucose Monitoring System For Diabetic Patient". Electronics.9 (4), 678, 2020.

[13] N. M. S. Kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," ProcediaComput. Sci., vol. 50, pp. 203–208, Jan. 2015

[14] Ahmed, H.B.; Serener, A. Effects of External Factors in CGM Sensor Glucose Concentration Prediction. Procedia Compute. Sci. 2016, 102, 623–629

[15] Dr Saravana Kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015

[16] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm", Journal Of Computing, Volume 3, Issue 12, December 2011

[17] Deepti Sisodia, Dilip Singh Sisodia," Prediction of Diabetes Using Classification Algorithm", www.elsevier.com/locate/procedia, Procedia computer science 132(2018)

1578-1585.