Prediction of diabetes using machine learning algorithms and comparison with other algorithms

Mr.C.Iyyappan

PhD Research Scholar

St.Peter's Institute of Higher Education and Research

Email:mail2iyyappan@gmail.com

Dr.R.Latha (Supervisor)

Professor & Head

Department of Computer Science and Applications

St.Peter's Institute of Higher Education and Research, Chennai-600054

Email:latharamavel@gmail.com

DOI: 10.31838/ecb/2023.12.si6.440

Abstract

Diabetes mellitus (DM) has numerous long-term effects on health. Diabetes may be caused by ageing, insufficient exercise, a sedentary lifestyle, a familial history of diabetes, high blood pressure, sadness or tension, poor diet, etc. Diabetes-related complications such as cardiovascular disease, diabetic neuropathy, diabetic retinopathy, diabetic nephropathy, and dementia are more prevalent in diabetics. Those between 25 and 74 years old are the most vulnerable. Diabetes complications may be severe if the condition is not properly diagnosed and managed. People are prevented from valuing their health by their frantic schedules. Right away following consuming sustenance, glucose is discharged into circulation. When blood glucose levels are excessively elevated, the pancreas secretes insulin. Without insulin, it is difficult for glucose to enter cells and be used as fuel. Diabetes can be detected early on, which aids in maintaining a wholesome lifestyle. Programmed based on machine learning will be successful because they can be taught and evaluated using massive amounts of data and are able to enhance themselves by generating future predictions. We present the results of training multiple algorithms on the data gathered for this article. SVM generated the most consistent results among the three evaluated algorithms.

Keywords: MachineLearning,Support Vector Machine,NaiveBayes,DecisionTree

Introduction

Diabetes mellitus is a metabolic disorder characterized by chronically elevated blood sugar levels due to insulin resistance or an inability to produce insulin. There is no evidence that any treatment strategy alleviates symptoms or delays the progression of this chronic metabolic disorder, which has far-reaching physical and mental health consequences for the patient. Diabetes is divided into two categories. In type 1 diabetes, the immune system attacks the insulin-producing cells of the pancreas, resulting in decreased manufacturing of insulin. There is a genetic component to the development of type 1 diabetes. While the body occurs insulin resistance and requires less insulin than usual, type 2 diabetes develops. This anomaly is associated with lack of exercise, poor diet, and excessive body obesity. Obesity, sedentary lifestyle, and genetics all contribute to the development of type 2 diabetes. Several hazards have been identified if diabetes is not controlled or detected early on. The prevalence of diseases such as type 1 diabetes in younger

age groups, the prevalence of overweight and obesity, and the occurrence of conditions like polycystic ovary syndrome and polycystic are all related to improper nutrition and unwholesome food.

During the literature review phase of our model's development, we discovered that traditional methodologies already existed. In order to create a model that is both unique and futuristic, a digital medical dataset containing measurements from fast food eaters, ordinary people who exercise, and their daily lifestyle is reviewed. This method yields more consistent results than the conventional approach, which relies on guesswork to arrive at a predetermined number. Consequently, they were unable to find evidence that their strategy was effective. Using machine learning techniques such as Random Forest, Decision Trees, and Naive Bayes, this problem is solved. And when applied to our digital dataset, Random Forest yielded the most accurate and efficient outcomes.

Low BMI, increased upper-body fat (adiposity), total body fat (%fat), and insulin resistance all increase the probability that an Indian will develop diabetes. The symptoms of diabetes include impaired vision, fatigue, increased appetite and thirst, vertigo, frequent urination, lethargic wound healing, recurring infections, and concentration difficulties. Hyperglycemia is characterised by excessively elevated blood sugar levels. Hyperglycemia occurs when the body does not produce enough of the hormone insulin. "sweet urine" translates literally as "urinating on sugar." Normal human urine contains no sugar. When blood glucose levels are high enough to cause urine leakage, sugar (or, more precisely, glucose) is present in the urine. Because the body cannot correctly metabolise glucose, it builds up in the circulation. Therefore, diabetes is a condition characterised by an impaired glucose metabolism.

Literature Review

KM Jyothi Rani has proposed an automated learning-based approach for diabetes prediction. In this work, they employed a dataset with 2,000 items while 9 characteristics, If you have type 2 diabetes, the score is 1. There were a total of five different AI methods used in the research. The Decision Tree method has a 98% teaching accuracy and a 99% testing accuracy.

A strategy for predicting diabetes using data mining techniques has been presented by BalaBisandu. Classification and prediction were carried out using the Naive Bayes Classifier, and the accuracy was close to 95%. The article relies on five characteristics for making its predictions about diabetes.

A large data processing system based on machine learning techniques to detect diabetes has been proposed by B. Suvarnamukhi. The rapid advancement of technology has allowed for the storing of data in the form of electronic health records (EHR) and its processing via big data, enabling prediction of three types of diabetes using Lsm and comparison with different algorithms.

Machine learning techniques were suggested by MitushSoni to improve diabetes prediction. The dataset used in this article includes 500 cases with negative results (no diabetes) and 268 cases with positive outcomes (diabetes). They employed six different machine learning algorithms to make their predictions, and the random forest approach had the highest success rate (77%)

N. Sneha1 and TarunGangil developed an algorithm for analysing and initial detection of diabetes mellitus by selecting the most pertinent features. The data set contains 2500 documents, 15 characteristics, and 768 test cases. Assistance vectoring achieves the maximum efficacy (77%) of the four employed methods.

Abdullah A. Aljumah and M.G. Ahmad developed a technique for detecting diabetes in both young and elderly persons using regression-based data mining. The Ministry of

Health in Saudi Arabia compiled a study detailing the variables that increase a person's chance of developing a noncommunicable disease. utilising data mining research, the information was used to make predictions about the efficacy of various therapies in both younger and older age groups.

Proposed System:

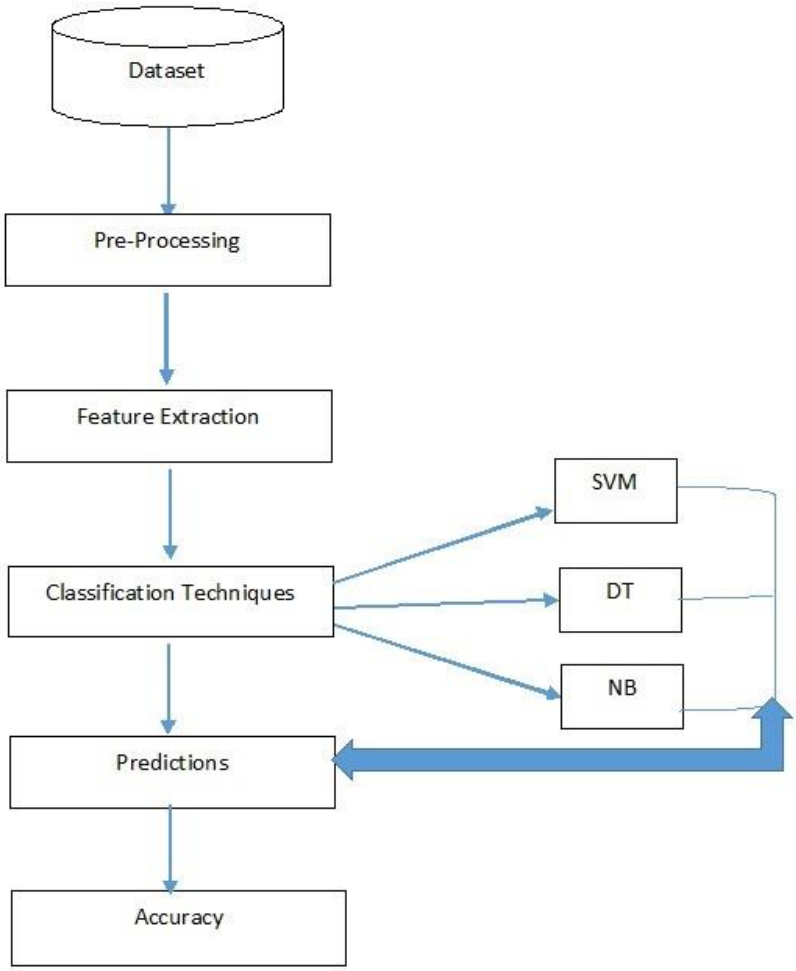


Fig 1: Flow of process

Data set description

The data is gathered from kaggle database which is name of diabetics and it contains more than 50,000 entries of data and 18 columns.

Data Preprocessing

This model handles missing values from the data base such as characters will come for numerical data, irrelevant dataset.

Cases where no value was offered are discarded. The technique of reducing the number of significant qualities by deleting superfluous occurrences is known as characteristics subset selection.

Data partitioning is the process of

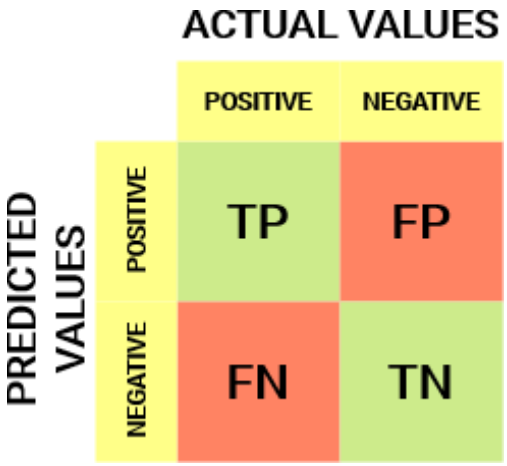
establishing training and testing data after removing irrelevant samples. Once data has been partitioned, it may be utilized to train an efficient algorithm on the training set and evaluated on the test set.After preparing the data, we'll divide it into training and testing halves, giving 80% of the data to training and 20% to testing. This information will then be taught using machine learning classification methods.

.

Random Forest, Decision Trees, and Naive Bayes are some examples of these algorithms. We'll use these methods to train the data, and then look at how well they did on test data to see how well they did.

Several different success markers can be used to measure how good the guesses are.

How Effective an Algorithm Is Described Creating a Matrix from a Messy Matrix



Where,

TP – True Positive,FP – False Positive

FN – False Negative,TN – True Negative

Accuracy - It is the ratio of number of corrected predictions to by total number of observations. The model is best if it has high accuracy

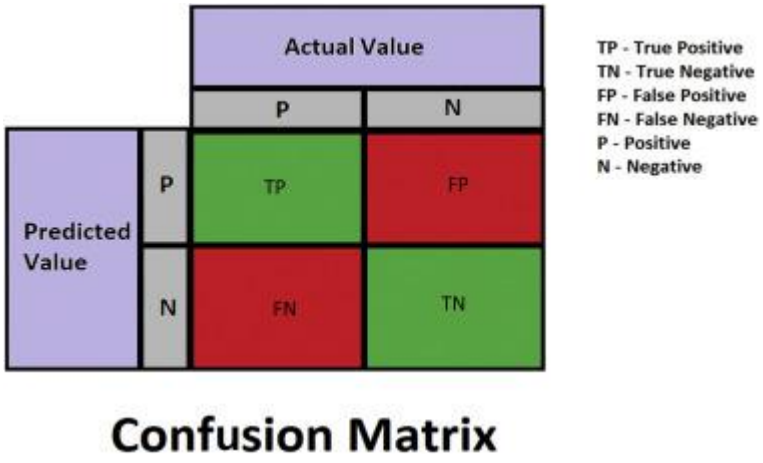
Accuracy =TP+TN/TP+TN+FP+FN

Precision =TP/TP+FP

Recall =TP/TP+FN

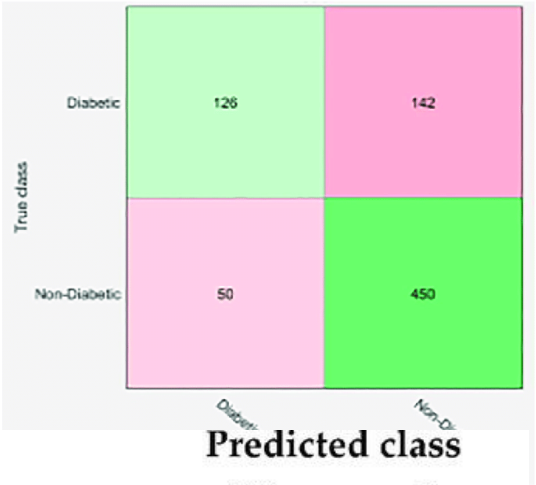
Confusion Matrix

SVM

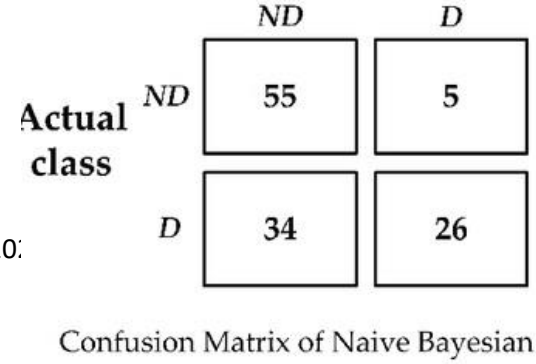


Confusion Matrix

Decision Tree



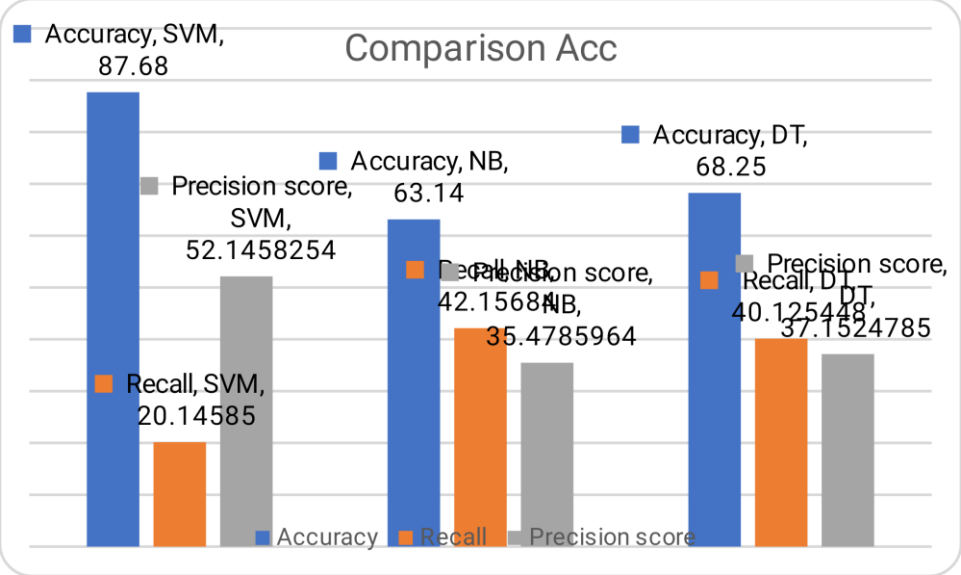
Naïve Bayes



Confusion Matrix of Naive Bayesian

Comparison Algorithms with Accuracy

Algorithms	Accuracy	Recall	Precision score
SVM	87.68	20.14585	52.1458254
NB	63.14	42.15684	35.4785964
DT	68.25	40.12545	37.1524785



Conclusion:

Many individuals are unaware of their own status with regard to this. This study seeks to investigate the concept of machine learning approaches for early diabetes prediction, as well as how to apply a variety of supervised machine learning algorithms to the data set for maximum accuracy. Support Vector Machine

(SVM) will outperform Naive Bayes (NB) and Decision Tree (DT) algorithms when applied to datasets including information on diabetes patients.

Reference:

1. M. Ramakrishna Murthy et,al(2022),Diabetes Prediction Using Machine Learning Algorithms,International Research Journal of Modernization in Engineering Technology and Science.
2. KM Jyoti Rani, "Diabetes Prediction Using Machine Learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, volume. 6, pp. 294-305, 2020.
3. Raja Krishnamoorthi, Shubham Joshi, and Hatim Z. Almarzouki, "A Novel Diabetes Healthcare Disease Prediction Framework using Machine Learning Techniques," Journal of Healthcare Engineering, pp. 1- 10 2022.
4. Desmond BalaBisandu, Godwin Thomas "Diabetes Prediction using Data mining Techniques," International journal of research and Innovation in Applied Sciences, volume 4, pp. 103-111, 2019.
5. B.Suvarnamukhi , M. Seshashayee, "Big Data Processing System for Diabetes Prediction using Machine Learning Techniques," in International Journal of Innovative Technology and Exploring Engineering, volume 8, pp. 4478–4483, 2019.
6. MitushiSoni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology, Volume 9, pp. 921-925, 2020.
7. N.Sneha , TarunGangil, "Analysis of diabetes meelitus for early prediction using optimal features selection", Journal of Big data, pp. 1-19, 2019