IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Survey On Diabetes Prediction Using Different Machine Learning Approaches

Dayananda R B, R Ramya, Vedavedya B H, Veera Sreenidhi, Venkatesh M N

Professor, Student, Student, Student, Student,
Department of Computer Science,
K S Institute of Technology, Bangalore, India

Abstract: Diabetes is an illness that is caused by a hormone called insulin in our body which when not functioning properly or not efficiently will raise the level of glucose in our body. There are many individuals who suffer from high blood sugar levels in their body which has a severe effect on other human organs. This condition is caused by an unhealthy lifestyle or food habits and lack of physical exercise. This will also lead to other complications such as coronary failure, blindness, urinary organ disease, etc. In such instances, the individual must visit a diagnostic center to collect their reports. Due to this, there is a loss of time and money in this process. As a solution for this problem, we can use Machine Learning methods to arrive at a solution to this issue, we have got advanced system mistreatment information processing that has got the potential to inform us whether the individual has got the disease or not. In addition to this, an early prediction of this disease will help in controlling it as well as be able to take precautions for its after-effects. Information withdrawal has the pliability to delete unseen data from a large dataset of diabetes information. The aim of our project is to predict illness in its early stages with better accuracy so that the risk factor is comparatively less. The model is based on methods such as Decision Tree, Random Forest, and Support Vector Machine algorithms.

Index Terms - Keywords: Diabetes Disease Prediction, Support Vector Machine, Machine Learning, Random Forest, Decision Tree

I. INTRODUCTION

Diabetes is one of the deadliest diseases in the world. It is not solely a malady however conjointly a creator of various sorts of diseases like heart failure, blindness etc. The conventional distinguishing method is that patients ought to visit a diagnostic center, consult their doctor, and rest for each day or additional to induce their reports. Moreover, whenever they need to induce their diagnosing report, they need to waste their cash vainly. There are of two different types of diabetes that can be classified into Type one polygenic disorder is that the kind wherever the exocrine gland doesn't manufacture hypoglycaemic agent. It had been erstwhile mentioned as endocrine dependent polygenic disorder or autoimmune disorder. Simple fraction of sufferers has this kind, individuals with this kind should acquire an artificial kind of endocrine they either receive it from an attempt or from associate degree endocrine pump. Diabetes Mellitus (DM) is formed public as a gaggle of metabolic disorders primarily caused by abnormal hypoglycaemic agent secretion and or action. Hypoglycaemic agent deficiency finally ends up in elevated blood glucose levels (hyperglycaemia) and impaired metabolism of carbohydrates fat and proteins. DM is one altogether the foremost common endocrine disorders moving quite two hundred million folks worldwide. The onset of polygenic disorder. The onset of polygenic disorder is calculable to rise, dramatically within the approaching year. In sort a pair of polygenic disorder the duct gland will create endocrine this way was antecedent named non-insulin dependent DM or non-insulin-dependent diabetes. However, it should not turn out enough. In different cases, the body doesn't use it properly. This can be called endocrine resistance folks with sort a pair of polygenic disorder may have to require polygenic disorder pills or endocrine. In inherited disease somebody usually suffers from high blood sugar Intensify thirst, intensify hunger and frequent evacuation of variety of the symptoms caused due to high glucose many complications occur if inherited disorder remains untreated.

Diabetes is divided into three main categories type 1, type 2 and gestational diabetes causes serious health concerns if it is not taken care properly. Type 1 diabetes was known insulin-dependent, childhood-onset or juvenile is characterized by body produce less insulin. People with type 1 diabetes require administration daily for insulin regulation of glucose in their blood. The person who has type 1 diabetes they cannot survive if they don’t have access to insulin. Type 1 diabetes currently is not preventable because the causes is still unknown. Symptoms like weight loss, excessive urination fatigue, vision changes, and thirst can indicate the type 1 diabetes. Type 2 diabetes also was known as non-insulin-dependent diabetes is due to improper food habits, overweight, not doing physical exercise. It is more common among diabetes patients.

Symptoms of type 2 diabetes may be similar to those on type 1 diabetes. Type 2 diabetes was common to adults but now some cases occur to the child who was diagnosed with type 2 diabetes. Gestational diabetes is occurred temporarily on a pregnant woman and disappears after giving birth. They are in the risk of some complication during pregnancy and delivering time. This can be prevented by proper exercise and weight before they become pregnant.

The aim of the project is of detecting Diabetes at an earlier stage by using various machine learning algorithms and to improve the efficiency rate. Diabetes is an illness that affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. A person generally suffers from high sugar levels in blood which can have severe effects on other human organs. Insulin is an essential hormone produced by the pancreas that allows the cells to absorb glucose (blood sugar) from food supplies in order to provide them with the necessary energy. Some of the oms are intense thirst, intensify hunger, and frequent urination. In medicine, doctors and current research confirm that if his disease is discovered at an early stage, the chances of recovery will be greater. But the identifying process is tedious, visiting a diagnostic center and consulting a doctor, i.e., these tests take a lot of time and waste the budget of health care systems and people y year. But the rise in machine learning approaches gives a solution to this problem. The learning algorithms use recorded tests of former patients' information to prepare a model and then use this model with information of an unseen patient to predict if the patient has the desired disease or not.

II. LITERATURE SURVEY

H. Suriya Babu, T. TamilArasan, P. TheepakPrakash [5], in this paper, they have implemented a model using MapReduce and Hadoop techniques for analysing the dataset. The developed model will predict the type of diabetes and also the risk that comes with it. They have used the Naive Bayes algorithm and the decision tree algorithm. Since it is a Hadoop-based model, it is economical for society. The performance of both algorithms was compared, as was the efficacy of both methods. They have used a decision tree to detect the hidden patterns in the dataset. The accuracy is 96% for logistic regression. The best model was AdaBoost, with an accuracy of 98.8%.

Ayman Mir, Sudhir N. Dhage [2] **, t**hey utilised the WEKA tools to create their model in this study. It's a well-known machine learning and data mining toolbox for data-driven research. The version of WEKA used is WEKA 3.28. They used this toolkit because it effectively evaluates the performance of a model and allows for real-time data comparison. The Naive Bayes algorithm, the SVM algorithm, and the Random Forest algorithm were all employed. They have divided the dataset into a training set and a test set. The training time, testing time, and accuracy values of the algorithms are compared. The algorithm that had the highest accuracy was the Support Vector Machine, with a 0.7913 accuracy value.

Mitushi Soni, Dr. Sunita Varma [3], in this research, they experimented with classification and an ensemble of algorithms to predict diabetes. The aim of their project is to build a model to predict diabetes with better accuracy. The proposed approaches used in this paper are the SVM algorithm, k-nearest neighbour algorithm, RF algorithm, DT algorithm, logistic regression, and gradient boosting. They have removed all instances containing zero as it is not possible. Data pre-processing is done for the dataset as it is a very important process and gives us better accuracy and prediction of the model. They have used an ensemble of machine learning algorithms as it gives a higher accuracy compared to individual algorithms alone. They have got an accuracy of 77% using the ensemble technique.

Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad AbdulGhani, and Khalid Qaraqe [1], they investigated diabetes risk factors using a dataset from epidemiological population research. The average age ranges from 24 to 64 years. The goal of this study is to create a machine learning algorithm that can identify healthy people who are at high risk of getting type 2 diabetes. To determine the accuracy of their model, they have used recall, accuracy, and specificity. They used the Support Vector Machine algorithm for their model. According to their observation, the performance did not increase or improve by adding more features. The model's conclusion is that a high glucose level recorded at the two-hour point during the OGTT may significantly suggest the possibility of acquiring type-2 diabetes.

K. VijiyaKumar, B. Lavanya, I. Nirmala, S. Sofia Caroline [7], in this paper, they have collected their dataset from the database. In pre-processing techniques, they have used data cleaning, integration, and transformation. They have selected random forest for their model as it gives higher accuracy compared to various machine learning algorithms. Data cleaning is a process that detects and removes corrupted or inaccurate data from a dataset. Data reduction is the conversion of numerical or alphabetical digital data into an ordered or simplified form of data. The goal of this study is to create a system that can accurately predict diabetes in an individual at an early stage. The accuracy of the random forest algorithm is more than 90%.

H. Suriya Babu, T. TamilArasan, P. TheepakPrakash [6], in this paper, the proposed system focuses on using the AdaBoost algorithm for its model. For this algorithm, they have tried various base classifiers such as decision trees, support vector machines, and naive Bayes. Their system is employed in four phases. They have used a global dataset for training and a local dataset for testing their model. The performance parameters that they have used are sensitivity, error rate, and specificity. They have used the Weka interface for accuracy verification. Among the four classifiers used, the Decision Stump shows the highest accuracy of 80.72%. The outcome of other algorithms is that the decision tree is 77.6%, the accuracy of the Support Vector Machine is 76.987%, and the accuracy of the Nave Bayes algorithm is 79.687%. The lowest error rate is also achieved by the Decision Stump algorithm with 19.27%. The DT method has an error rate of 22.39 percent. The SVM algorithm has an error rate of 20.31 percent, and the Naive Bayes algorithm has an error rate of 20.31 percent.

M D. Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, Mahmodul Hasan [4], in this paper, the machine learning models were trained and tested using publicly available datasets. They took the datasets from Pima datasets, which contain 768 data points with nine different features. The outcome of the model is either zero or one. One indicates that the individual has tested negative for diabetes and one indicates that the individual has tested positive for diabetes. They employed the decision tree method, the Nave Bayes algorithm, Logical Regression, the SVM algorithm, the Random Forest (RF) algorithm, the k-nearest neighbour algorithm, and the AdaBoost algorithm, among other machine learning techniques.

Dr. Kayal Vizhi, Aman Dash [9], in this paper, the project model can evaluate only a specific or particular parameter, not taking into consideration the remaining parameters. This paper needs more modification. They took the datasets from Pima datasets, which contain 768 data points with nine different features. The outcome of the model is either zero or one. One indicates that the individual has tested negative for diabetes and one indicates that the individual has tested positive for diabetes. Their project is totally dependent on an IoT-based framework and cloud computing. They have given high preference to the individual's privacy and safety. They have used a logical regression algorithm, gradient boost, and feature selection. They have concluded by saying that logical regression gives better accuracy when compared with other algorithms, with an accuracy of 78%.

Jobeda Jamal Khanam, Simon Y. Foo [10], the decision tree method, the Naive Bayes algorithm, the Logical Regression, the SVM algorithm, the RF algorithm, the KNN algorithm, and the AdaBoost algorithm were all employed in this research. They took the datasets from Pima datasets, which contain 768 data points with nine different features. The outcome of the model is either zero or one. One indicates that the individual has tested negative for diabetes and one indicates that the individual has tested positive for diabetes. Using pre-processing techniques, they were able to eliminate three of the nine parameters. The Weka tool is used to check the accuracy of the models. All the model accuracy was above 70%. The highest accuracy was from the ANN algorithm, with 88.57%.

Anjali C, Veena Vijayan V [8], the goal of their study, as described in this paper, is to diagnose diabetes at an early stage by utilising various machine learning techniques and to increase efficiency. In the proposed system, they have applied a combination of PCA and k-means algorithms. The dataset used is from the public domain. They have built multiple models using k-means, ANN, and support vector machine algorithms. A pre-processed dataset will be used to train the models. They constructed a confusion matrix and put the user's input data to the test. Fresh data can be tagged with K-cluster centroids. For every new evaluation, the centroid has to be recalculated. They have used Panda data to load CSV data. They have concluded that the k-means algorithm gives better accuracy compared to other algorithms.

III. DATASET

Here is the description of the dataset that has been used as an input to classifiers implemented using various algorithms. The name of the dataset that has been considered is Pima Indians Diabetes Database which is collected from National Institute of Diabetes and Digestive and Kidney Diseases. The total No. of Instances are 768 and the size is 37 KB. The total no. of attributes is 9 including the target class attribute. The name of two target classes is tested positive and tested negative. The no. of instances for tested positive are 268 and the no. of instances for tested negative are 500.

IV. METHODOLOGY

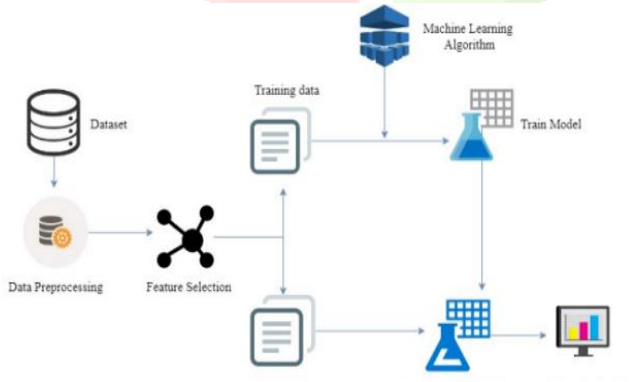

Fig 4.1 Working of the model

Data Pre-Processing

It is a technique used to convert raw data into an understandable dataset. We do data pre-processing such as (i) replacing missing values and (ii) normalisation of values because this dataset may include null values or missing values. The KNN imputer method will be used to replace numerical and nominal values. Remove all the instances that have zero. A value of zero is not feasible. As a result, this instance isn't valid anymore.

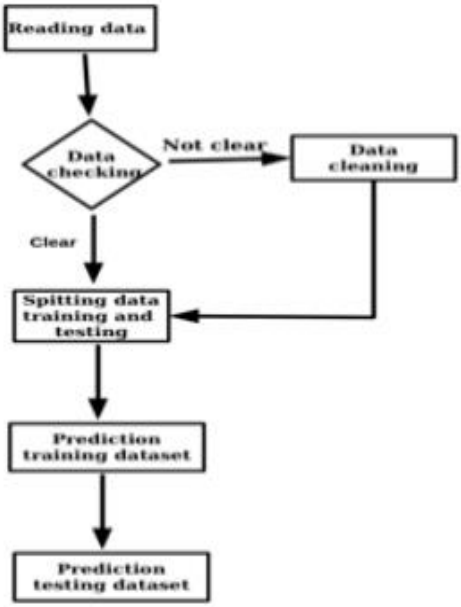

Fig 4.2 Pre-processing the data

Splitting of Dataset

A percent split option is available for training and testing. In 768 cases, 75% of the time is spent on teaching and 25% on testing. After the data has been cleaned, it is normalised for training and testing the model. The training data set is used to train the algorithm, whereas the test data set is placed aside once the data is separated. The training model will be created using logic and procedures as well as the values of the feature in the training data.

Feature selection

We create feature subsets by deleting non-essential qualities, a technique known as "feature subset selection," which reduces data dimensionality and accelerates labour. Feature selection methods can reduce the number of attributes, which can avoid redundant features. There are many methods for selecting features. In this model, we use the recursive feature elimination method. It's a feature selection algorithm that's simple to set up and use, and it's good at picking the right mix of features to get the best results.

Model Training

This is the most crucial step, which includes the development of a diabetes prediction model. When the data is ready, we use machine learning algorithms to forecast diabetes. The major purpose is to assess the performance of various techniques and verify their accuracy using machine learning algorithms. The methods are the most prominent supervised learning algorithms for classification and regression issues are SVM, RF, and DT. The pre-processed dataset will be used to train these models.

Based on the training set, create a classifier model for the stated machine learning technique. On the basis of the test set, evaluate the classifier model for the aforementioned machine learning algorithm.

Deployment of model

Creating a Machine Learning model is not enough until we make it available to general use or to a specific client. Streamlit is a popular open-source framework used for model deployment by machine learning. Streamlit lets to create apps for machine learning projects using simple python scripts.

V. CONCLUSION AND FUTURE WORK

This survey helps to propose a model that helps in diabetes prediction as explained. We can detect diabetes in the early stages to get treated with a minimum cost and minimum risk. The machine learning model helps in the detection based on clinical data model forms a computer-assisted diagnostics to help physicians and radiologists in supporting their diagnostic decisions and helpful to reduce high costs of diagnosis. And machine learning model will help the patients to start the medication early and it will help to diagnose more patients within a less time period. Proper feature selection methods help to reduce the number of features needed for the prediction algorithms and practically it reduces the number of medical tests to be taken.

Future research should analyze different supervised and unsupervised machine learning techniques and feature selection techniques with additional performance metrics for better diabetes prediction.

Figures and Tables

Table 1 The following table describes the 9 attributes of the diabetes dataset briefly

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

V. ACKNOWLEDGMENT

We would like to express our deep gratitude to Mr. Dr. Dayananda B R for his valuable and constructive suggestions during the planning and development of this project. We would also like to thank all the professors, staff, and management of KSIT for their continuous support and encouragement.

REFERENCES

[1] Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad AbdulGhani, and Khalid Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning", IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)-2019.
[2] Ayman Mir, Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)-2018.
[3] Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT)-2020.
[4] MD. KAMRUL HASAN, MD. ASHRAFUL ALAM, DOLA DAS, EKLAS HOSSAIN, MAHMUDUL HASAN, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", IEEE-2020.
[5] H. Suriya Babu, T. TamilArasan, P. TheepakPrakash, "PREDICTION OF DIABETES MELLITUS USING MACHINE LEARNING ALGORITHMS", International Journal of Advanced Engineering Science and Information Technology (IJAESIT)-2021.
[6] Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema, Diabetes Mellitus Prediction using Classification Techniques, -International Journal of Innovative Technology and Exploring Engineering (IJITEE)-2020.
[7] K. VijiyaKumar, B. Lavanya, I. Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes", Proceeding of International Conference on Systems Computation Automation and Networking-2019.
[8] Anjali C, Veena Vijayan V, "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", IEEE Recent Advances in Intelligent Computational Systems (RAICS)-2015.
[9] Dr. Kayal Vizhi, Aman Dash, "Diabetes Prediction Using Machine Learning", International Journal of Advanced Science and Technology-2020.
[10] Jobeda Jamal Khanam, Simon Y. Foo, "A comparison of machine learning algorithms for diabetes prediction", The Korean Institute of Communications and Information Sciences (KICS)-2021.