International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 7 ǁ July 2022 ǁ PP. 627-631

Diabetes Prediction Using Machine Learning

Neethu Stanly[1], Preethi Kuruvilla[2], S Nalanda[3], Sanju Sabu[4], Tony Jacob[5], Kavitha S[6]

1 - PG Student, Department of Computer Applications – Saintgits College of Engineering (Autonomous) Pathamuttom Kottayam Kerala, India
2 - PG Student, Department of Computer Applications – Saintgits College of Engineering (Autonomous) Pathamuttom Kottayam Kerala, India
3 - PG Student, Department of Computer Applications – Saintgits College of Engineering (Autonomous) Pathamuttom Kottayam Kerala, India
4 - PG Student, Department of Computer Applications – Saintgits College of Engineering (Autonomous) Pathamuttom Kottayam Kerala, India,
5 - PG Student, Department of Computer Applications – Saintgits College of Engineering (Autonomous) Pathamuttom Kottayam Kerala, India
6 - Assistant Professor, Department of Computer Applications – Saintgits College of Engineering(Autonomous) Pathamuttom Kottayam Kerala, India

Abstract - *Diabetes is a common, chronic disease. Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, decision tree and association rule mining for diabetes. Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy.*

Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository.

Date of Submission: 08-07-2022 Date of acceptance: 22-07-2022

I. INTRODUCTION

Diabetes remains one of the pressing health problems. This paper is devoted to solving the problem of classifying patients with diabetes and diagnosing this disease. To solve the problem, a machine learning model was built based on a decision tree method. Decision tree analysis is a predictive modeling tool that can be applied in many areas. Decision trees can be built using an algorithmic approach that can partition the dataset in different ways depending on conditions. The more data is allocated for training the model, the better the accuracy estimate weget.

Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not return correctly to insulin, or by both reasons. The day is now to prevent and diagnose diabetes in the early stages. According to the WHO (world health organization) report in Nov 14, 2016, in the world diabetes day DzEye on diabetesdz reported 422 million adults are with diabetes, 1.6 million deaths, as the report indicates it is not difficult to guess how much diabetes is very serious and chronic.

In 2014, 8.5% of adults whose ages are 18 and older than 18 had diabetes. In 2012 HBG (high blood glucose was the cause of 2.2 million people deaths Diabetes diseases damage different parts of the human body from those parts some of them are: eyes, kidney, heart, and nerves. Williams's textbook of endocrinology was predictable that in 2013 more than 382 million population in the world or all over the worldwerewithdiabetesorhaddiabetes.There are so many peoples are died every year by diabetes disease (DD) both in poor and rich countries in the world.

According to the centers for disease control and prevention (CDCP) they give information for the duration of 9 ensuing years that is between 2001 and 2009 type II diabetes increased 23% in the United States (US). There are different countries, organization, and different health sectors worry about this chronic disease

control and prevent before the person death. Most in the current time diabetes is grouped into two types of diabetes, type I and Type II diabetes. Type I diabetes this type of diabetes in heath language or in doctors' language this type of diabetes also called Insulin dependent diabetes illness. Here the human body does not produce enough insulin. 10 % of diabetes caused by this type of diabetes. Type II diabetes this type of diabetes. According to CDA (Canadian Diabetes Association) for 10 years, between 2010 and 2020, expected to increase from 2.5 million to 3.7 million. Therefore, the above-mentioned Diabetes diseases needs early prevention and diagnosis to safe human life from early death .By considering how much this disease is very series and leading one in the world. Algorithms which are used in machine learning have various power in both classification andpredicting.

Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million. Diabetes Mellitus (DM) is classified as Type-1 known as Insulin- Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non- Insulin Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. Machine learning algorithms are mostly categorized as being supervised or unsupervised. A supervised learning algorithm uses the past experience to make predictions on new or unseen data while unsupervised algorithms can draw inferences from datasets. The supervised learning is also called classification. This study uses classification technique to produce a more accurate predictive model as it is one of the most commonly applied machine learning technique that examines the training data and creates an inferred function, which can be used for mapping new or unseen examples. The major goal of the classification technique is to forecast the target class accurately for each case in the data. Classification Algorithms generally require that the classes be defined grounded on the data attribute values. They often define these classes by looking at the characteristics of data already known to belong to class. This process of finding useful information and patterns in data is also called Knowledge Discovery in Databases (KDD) which involves certain phases like Data selection, Transformation, Classification and Evaluation. Evaluation. Several real-world application for example medical diagnoses, fraud detection.

II. MACHINELEARNING

Tom Mitchell states machine learningas

–Acomputerprogramissaidtolearnfrom experience and from some tasks and some performance on, as measured by, improves with experience‖. Machine Learning is combination of correlations and relationships, most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest and etc. The name machine learning was coined in 1959 by Arthur Samuel. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allowresearchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

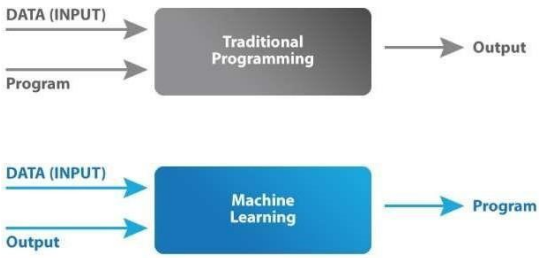

Fig. 1: Traditional Programming and Machine Learning

III. DECISION TREEALGORITHM

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). A decision tree is a graphical representation of all possible solutions to a decision based on certain conditions. On each step or node of a decision tree, used for classification, we try to form a condition on the features to separate all the labels or classes contained in the dataset to the fullest purity. Decision tree is one of the predictive modeling approaches used in statistics, data mining and machine learning. It is one of the most widely used and practical methods for supervised learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. Decision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one ofthepredictive modeling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, but the resulting classification tree can be an input for decisiontree.

Decision tree are a non-parametric supervised learning method is used for both classification and regression method. Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this. Decision tree model has a tree structure, which can describe the process of classification instances based on features. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification". Each element of the domainofthe classificationiscalleda class. A decision tree or a classification tree is a tree in which each internal (non- leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes, signifying that the data set has been classified by the tree into either a specific class, or into a particular probability distribution(which, if the decision tree is well-constructed, is skewed towards certain subsets of classes). A tree is built by splitting the source set, constituting the root node of the tree, into subsets— which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees fromdata.

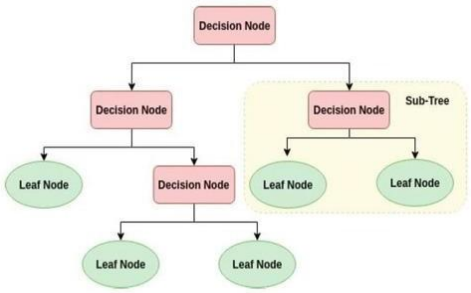

Fig. 2: A DECISION TREE MODEL

- Decision Tree algorithm belongs to the family of supervised learning algorithms.
- Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problemstoo.

- The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from priordata.
- A decision tree is a tree where each node represents a feature(attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).

IV. EXISTING SYSTEM ANDITS PROBLEMS

Diagnosis of diabetes is considered a challenging problem for quantitative research. Some parameters like A1c ,fructosamine , white blood cell count, fibrinogen and hematological indices were shown to be ineffective due to some limitations. Different research studies used these parameters for the diagnosis of diabetes. A few treatments have thought to raise A1C including chronic ingestion of liquor, salicylates and narcotics. Ingestion of vitamin C may elevate A1c when estimated by electrophoresis but levels may appear to diminish when estimated by chromatography. Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension. A family history of diabetes has not been associated withBMI and insulin. However, an increased BMI is not always associated with abdominal obesity. 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 differingattributes.

V. PROPOSEDSYSTEM

The proposed system is designed based on the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the Diabetes incidents at a particular age, with higher accuracy using Decision tree. An important challenge in the fight against diabetes is the classification of patients and the diagnosis of the disease. To solve this problem, it is advisable to use a machine learning apparatus.

In modern science, several models have been implemented that make it possible to diagnose diabetes according to specified parameters. Researchers are passionate to try different types of classifiers and build new models with an effort to enhance the accuracy of diabetes prediction. In this paper, the same vision was followed to reach high prediction accuracy. Decision tree analysis is a predictive modelling tool that can be applied in many areas. Decision trees can be built using an algorithmic approach that can partition the dataset in different ways depending on conditions. The aim of the research is to build a model that allows classifying a person's condition in relation to the incidence of diabetes using machine learningmethods.

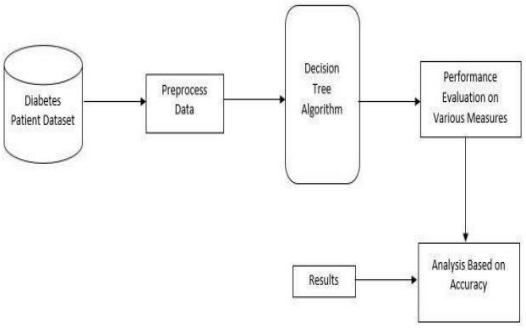

Fig : Proposed System for the model

VI. EXPERIMENTS AND RESULTS

All the Machine Learning (ML) classifiers that have been used in the last six years were reviewed regarding their frequency of use and accuracy. ML classifiers with oneor zero frequency have been implemented on the PID dataset to set recommendationsregarding their usage. The obtainedaccuracy by these ML techniques was68%–74%. For the ML algorithms, thehighest accuracy achieved by researcherswas 80% by using decision tree algorithm. As a future work, the non-usedclassifiers can be applied to other datasetsin a combined model to enhance further theaccuracy of predicting the Diabetesdisease.

VII. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage.

Overall, it can be said that decision tree analysis is a predictive modelling tool that can be applied in many areas. Decision trees can be built using an algorithmic approach that can partition the dataset in different ways depending on conditions. After analysing the constructed diagnostic model, the following advantages can be identified: fast learning process; generation of rules in areas where it is difficult for an expert to formalize his knowledge; intuitive classification model; high prediction accuracy, comparable to other methods of data analysis (statistics, neural networks); construction of nonparametric models. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

REFERENCES

[1]. Clinton Sheppard, "Tree-based Machine Learning Algorithms: DecisionTrees, Random Forests and Boosting",kindle Edition.
[2]. Rodrigo C. Barros, Andre C.P.L.Fde Carvalho, Alex A. Freitas, "AutomaticDesignof Decision-Tree InductionAlgorithms", Springer.
[3]. Chris Smith, "Decision Tree andRandom Forests:A visual introduction forbeginners", kindle.
[4]. Classification and Prediction ofDiabetes Disease using Decision TreeMethod.
[5]. Vizhi K, Dash A. DiabetesPrediction Using Machine Learning.International Journal of Advanced Scienceand Technology. 2020;29(06):2842–2852.
[6]. V. V. Ramalingam, AyantanDandapath, M Karthik Raja "Heart DiseasePrediction using Machine LearningTechniques: a survey", InternationalJournal of Engineering andTechnology.