Early Stage Diabetes Prediction Using Decision-tree Classification

Nithya Santhoshini Chillapalli
SRM University-AP,
Andhra Pradesh, India.
RollNumber: AP19110010502

Shreya Pandrangi
SRM University-AP,
Andhra Pradesh, India.
RollNumber: AP19110010433

Sanchita Rawat
SRM University-AP,
Andhra Pradesh, India.
RollNumber: AP19110010535

Abstract—**Diabetes is the most common disease worldwide. In India alone, around 11.8 % of the population is diabetic. Diabetes should not be ignored, if remained untreated it can be a major life threatening disease. It is one of those diseases where early detection can prove to be really helpful in controlling and reducing the symptoms. With correct prediction we can save more people from being severely diabetic. Furthermore, predicting the disease early leads to treatment of the patients before it becomes critical. To achieve this goal, in this project we will do early prediction of Diabetes in a human body or a patient with accuracy with the help of various Data Mining Techniques. Decision Tree classifier will be used in this model. These Data Mining techniques provide better results for prediction mining by constructing models from datasets collected from various patients. The implementation of this new system of predicting diabetes will also help to reduce the stressful process, doctors' face during prediction of Diabetes in a patient. The result of the experiment shows that the proposed system has a better prediction in terms of accuracy.**

Index Terms—**Diabetes, Data Mining algorithms, datasets, decision tree, accuracy**

I. Introduction

Diabetes is a metabolic disease which occurs when the blood glucose or the blood sugar levels are higher than normal. Blood glucose is the main source of energy that comes from the food we eat. Diabetes affects how the human body turns food into energy. In medical science it is described as a chronic disease that occurs either when the pancreas does not produce enough insulin, a hormone that regulates blood sugar. or when the body cannot effectively use the insulin it produces.The type 1 diabetes is a genetic disorder that often shows up early in life, and type 2 is largely diet and lifestyle related and develops over time. Insulin carries sugar from the bloodstream to various cells to be used as energy. Lack of insulin disrupts the body's natural ability to produce and use insulin accurately. As a result of this, high levels of glucose are released in urine. In the long term, diabetes when not properly managed can lead to life threatening illnesses like organ failure, cardiovascular diseases , damage to the nervous system and disrupting other functions of the body. The International Diabetes Federation Atlas Tenth edition in 2021 provides alarming statistics-

- Approximately 537 million adults (20-79 years) are living with diabetes.
- The total number of people living with diabetes is projected to rise to 643 million by 2030 and 783 million by 2045.
- 3 in 4 adults with diabetes live in low- and middle-income countries
- Almost 1 in 2 (240 million) adults living with diabetes are undiagnosed
- Diabetes caused 6.7 million deaths
- Diabetes caused at least USD 966 billion dollars in health expenditure – 9 % of total spending on adults
- More than 1.2 million children and adolescents (0-19 years) are living with type 1 diabetes
- 1 in 6 live births (21 million) are affected by diabetes during pregnancy
- 541 million adults are at increased risk of developing type 2 diabetes

This data clearly shows the seriousness of the problem and how important it is to have an accurate method for predicting diabetes early on.

Data mining is a relatively new concept used for extracting information from a large set of data. Mining means using available data and processing it in such a way that it is useful for decision-making. Using data mining algorithms we can process the dataset of information collected from various diabetic and non diabetic patients to process this data set and find pattern useful in predicting with great accuracy if a person will be diabetic or not.

II. Literature Survey

In [1], the author proposed a technique consisting of four modules:

- Diabetes dataset
- Pre-processing
- Classifier Algorithms
- Comparison Module

In the first step, the authors gathered the dataset using a direct poll circulated to 520 patients at an emergency clinic named Sylhet Diabetes Hospital in Bangladesh, who had been diagnosed to have diabetes or having diabetes-related side effects. It includes 16 attributes with 320 positive and 200 negative examples, with positive and negative markers used to decide if a patient is at risk of diabetes or not. In the second step, pre-processing techniques such as data cleaning and data transformation have been used to remove the irrelevant data, noisy data and to convert the nominal data to numerical data. In the third step, classifier algorithms are applied to the pre-processes data. Classifier algorithms are of two types: Machine learning and Deep Learning. In machine learning, the algorithms that have been used are:

- Support Vector Machine
- Decision Tree
- Logistic Regression
- XG Boast
- Random Forest
- K- Nearest Neighbors

In Deep learning, the algorithms that are used are:

- Artificial Neural Network
- Long short term memory
- Multi layer perceptron

In the final step, the author evaluated the accuracy for all the methods and concluded that the diabetes dataset was assessed utilizing nine distinct classification approaches, including XGB, RF, DT, KNN, SVM, ANN,MLP, and LSTM. According to the experiment performed by the author, they found XGBoost outperformed near 100.0% and was essentially better than other machine learning and deep learning approaches for distinguishing early stage diabetes.

In [2], the author performed four steps:

- Dataset collection
- Data Pre-processing
- Association Rule Mining
- Modeling

The author collected the dataset directly from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains 9 class attributes, 768 records describing female patients (of which there were 500 negative examples (65.1%) and 268 positive examples (34.9%)). After collecting the dataset the author implemented pre-processing such as data cleaning to fill the missing values, data reduction (dimensionality reduction) to reduce the representation the given dataset,etc. Though the dataset contains less data compared with the original data, it gives the same result.For smoothing of data, binning method has been implemented. After the data has been pre-processed, association rule mining is performed to determine the frequent patterns and items present in the dataset. In this paper, three algorithms were used to predict the early stage of diabetes:

- Artificial Neural Network
- Random Forest
- K-means Clustering

At the end, the results will be generated on the basis of accuracy and AUROC curve. The accuracy of each model will be predicted with the help of confusion matrix.

In [3], the algorithms that are utilized to predict the diabetes are Naive Bayes, J48 Decision Trees, Logistic Regression and Random forest algorithm. Naïve Bayes is a probabilistic classifier which means it predicts based on the probability of an object. J48 algorithm is a decision tree and it belongs to supervised learning algorithms. It is one of the most important classifiers.The algorithm is easy and simple to implement. Utilizing the decision tree, the dataset can be broken down into smaller subsets and at the same time an associated decision tree is incrementally developed. Logistic regression is a statistical model that in its essential structure utilizes a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is assessing the parameters of a logistic model. Random forest is a Supervised Machine Learning Algorithm that is utilized most part in Classification and Regression issues. It uses bagging method to train the dataset. It builds decision trees on different samples and takes their majority vote in favour of classification and normal in case of regression. After implementing all the algorithms, the performance of each algorithm is evaluated. By observing the results, the best algorithm will be proposed to predict the early stage of diabetes. According to this paper, Random Forest algorithm had performed with the best accuracy in percentage split evaluation test and it is the best approach for diabetic risk prediction.

III. Proposed model

The approach for this method is separated into 6 modules:

- Dataset Collection
- Pre-processing
- Training and Testing
- Decision tree classification
- Performance evaluation
- Decision tree visualization

A. *DATASET COLLECTION*

The early stage diabetes data set is taken from UCI machine learning repository.The dataset has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

1) ATTRIBUTE INFORMATION:

- Age 1.20-65
- Sex 1. Male, 2.Female
- Polyuria 1.Yes, 2.No.
- Polydipsia 1.Yes, 2.No.
- sudden weight loss 1.Yes, 2.No.
- weakness 1.Yes, 2.No.
- Polyphagia 1.Yes, 2.No.
- Genital thrush 1.Yes, 2.No.
- visual blurring 1.Yes, 2.No.
- Itching 1.Yes, 2.No.
- Irritability 1.Yes, 2.No.
- delayed healing 1.Yes, 2.No.
- partial paresis 1.Yes, 2.No.
- muscle stiffness 1.Yes, 2.No.
- Alopecia 1.Yes, 2.No.
- Obesity 1.Yes, 2.No.
- Class 1.Positive, 2.Negative.

B. PRE-PROCESSING

In this module, data cleaning and data transformation will be performed. In the Data cleaning step, the dataset is checked for missing values. If missing values are present in the dataset, then the missing values will be either removed or replaced by using any of the methods like replacing with mean/median, replacing with global constant, removing the missing values, filling values manually,etc. to improve the quality of data analysis. This dataset does not consist of any missing values. Decision tree does not support categorical values. So, the columns which contain categorical values will be transformed to numerical values. For example,the column “polyuria” which contains values as ”Yes” and ”No” have been transformed to numerical values 1 and 0.

C. DECISION TREE CLASSIFICATION

Decision tree classification algorithm is a supervised machine learning algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain or minimum Entropy. The steps in decision tree classification algorithm are as follows:

1) Calculate entropy for dataset.
2) For each attribute/feature.
    a) Calculate entropy for all its categorical values.
    b) Calculate information gain for the feature.
3) Find the feature with maximum information gain.
4) Repeat it until we get the desired tree.

D. TRAINING AND TESTING

In this module, the dataset is split into two categories i.e. training data and testing data. Training data is used to create a data analysis model whose accuracy is calculated by testing the model using the testing data. Seventy percent of the dataset is considered as training data and the remaining thirty percent of the dataset is considered as testing data. Hence, this dataset has been split into 721 columns of training data and 309 columns of testing data.

E. PERFORMANCE EVALUATION

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. Accuracy of the classification model is calculated using metrics from sklearn.

F. DECISION TREE VISUALIZATION

In this step, the generated decision tree will be visualized using pydotplus package.

IV. Results

The obtained decision tree model, as shown in Figure 1, can now be used to predict the likelihood of early stage diabetes based on attributes such as age,gender,muscle stiffness, delayed healing, visual blurring and itching, with an accuracy of 89.6%.

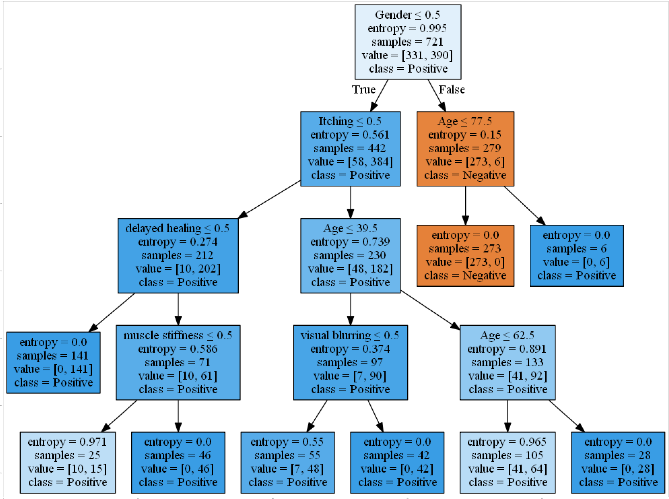

Fig. 1. Decision tree

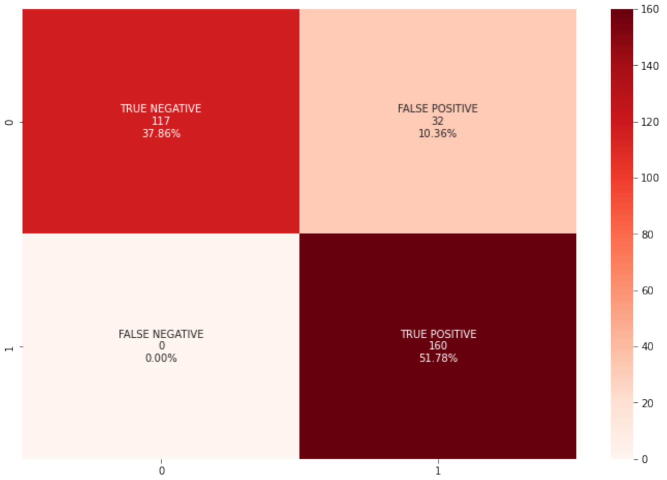

Fig. 2. Confusion matrix

V. Conclusion

Machine learning and data mining techniques are efficient in disease prediction. The ability to predict disease in early stages significantly reduces the risk factor and is vital for appropriate treatment especially in diseases like diabetes where early diagnosis can even reduce the effects. In this paper, diabetes is predicted using decision tree classification algorithm on the diabetes dataset taken from UCI machine learning repository. The model is further tested using a test dataset. Results are generated based on accuracy. The accuracy of the classification model is 89.6%. The results of our implementation show that the model correctly predicted 51.7% samples as positive. The limitation of this paper is that dataset is assumed to be structured, dealing with unstructured dataset might be used in the future. Other attributes including physical inactivity, family history of diabetes, drinking and smoking habit, are also planned to be considered in the future for diagnosis of diabetes.

References

[1] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach," pp. 654–659, 2021.

[2] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[3] M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, 2020, pp. 113–125.