Diabetes Prediction Using Machine Learning

Ikra Bagwan [1], Nikita Mane[2], Aarti Pote[3], Sandip Chavan[4]
[1, 2, 3, 4] Dept of Computer Engineering
[4]Asst-Prof, Dept of Computer Engineering
[1, 2, 3, 4] Bharati VidyaPeeth C.O.E, Mumbai, India

Abstract- *Day by day from health industries large data is generating which is necessary to collect, process and store data in order to discover knowledge from it and utilize to take wise decisions.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 With the help of technology, it is necessary to build a system that store and analyze the diabetic data and predict possible risks accordingly. Predictive analysis is a method that integrates various data mining techniques, machine learning algorithms and statistics that use current and past data sets to gain insight and predict future risks. Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore three machine learning classification algorithms namely Decision Tree, Logistic Regression and Naive Bayes would be used inthis experiment to detect and predict diabetes at an early stage. Experiments will be performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms would be evaluated on various measures like Precision, Accuracy, F- Measure, and Recall. Accuracy will be measured over correctly and incorrectly classified instances.*

Keywords- Disease, Diabetes, Prediction., Naïve Bayes, Decision tree,Logistic Regression.

I. INTRODUCTION

Currently in a global world, there are so many chronic diseases are distributed throughout the world, both in the developing and developed country such serious disease are distributed. From those serious diseases, Diabetes is one of the chronic diseases in the world which cut human life at early age. Diabetes gets its name by health professionals' .At this time diabetes disease increases rapidly within the distance of light like Indian countries and some countries. It is not difficult to guess how much diabetes is very serious and chronic. There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person died that means the early presentation of diabetes in order to save human life. Eating is also one factor for diabetes diseases and also, exercise used for healthy even a person live with diabetes the patient can recover from the disease by doing exercise Diabetes diseases have the power or ability to damage different parts of the human being body, from those human body parts which are affected by diabetes are listed as follow:- human heart, human eye, human kidney, and human nerves . As it indicates it is easy to guess how much it is chronic and dangerous diseases that shorts human life Algorithms which are used in machine learning have various power in both classification and predicting. there is no single technique gives better performance and accuracy for all diseases, whereas one classifier provides or shows highest performance in a given dataset, another method or approach outdoes the others for The new proposed study follows the different machine learning techniques (MLTs) to predict diabetes at an early stage to save human life. Such algorithms are-Decision tree ,logistic regression and Naïve bayes .

II. LITERATURE REVIEW

The review on prior work gives many results on analysis of healthcare data which was carried out by different methods, techniques. Many researchers have developed and implemented various analysis and prediction models using different Machine Learning Algorithms and Hadoop techniques or combination of these techniques

Aiswarya Iyer et al. (2015), used classification technique to find out patterns from the diabetes data sets. They employed naive Bayes and Decision Tree algorithms by using Weka tool. Authors also compared performance of both algorithms against Pima Diabetes Data sets. Here experimental results showed effectiveness of each proposed classification model.

V. H. Bhat et al. (2009) proposed an approach of integration of regression,classification,genetic and neural network which deals with the missing values as well as outlier values inthe diabetic data set and replaced the missing values by the corresponding attribute domain.

For prediction they used classical neural network model and applied it on the preprocessed Dataset

Marius et al. have proposed this system that implements rather fast generating nearest neighbor and appropriate algorithm configuration. In this system, this system they have built up a framework that choses a fitting algorithm in view of the data bolstered which rather creates the fastest nearest neighbor. This algorithm is selected based on dimension of the data. For some PC vision issues, the most tedious segment comprises of nearest neighbor coordinating in highdimensional spaces. There are no known correct algorithms for tackling these high- dimensional issues that are speedier than straight pursuit. Rough algorithms are known to furnish expansive speedups with just minor misfortune in exactness,however numerous such algorithms have been distributed with just negligible direction on choosing an algorithm and its parameters for any given issue.

III. PROPOSED SYSTEM

The admin of the system will ask the patient to give provide the current records. Then the admin of the system will ask the patient his/her details needed for the prediction of diabetes The admin of the system will then choose three appropriate algorithms available. Thus, after using the system, the prediction will be done whether the patient isdiagnosed with diabetes or not. If the patient is found diabetic expert recommendations would be provided to the patient so that he/she can recover from diabetes. Diet plans will be provided to the patient. This system would be very much useful in the field ofhealthcare.

The Execution would be such: -

After coming on to the home page, next the loading and displaying page gets loaded, where the admin of the system must load and display the database of diabetes patients. After this the login page gets loaded, the user has to give details for following features that is responsible for diabetes With the help of the details available prediction is performed with the help of Bayesian , Logistic Regression and Decision Tree Algorithm and result will be generated on next page.

If found diabetic recommendation and diet plan would be provided to the patient.

The following features have been provided to help us predict whether a person is diabetic or not:

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skin fold thickness (mm)

Insulin: 2-Hour serum insulin (mu U/ml)

BMI: Body mass index (weight in kg/(height in m)2)

Diabetes Pedigree Function: Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)

Age: Age (years)

Outcome: In terms of Percentage

OVERVIEW-

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database

We’ll be using Python and some of its popular data science related packages.

First of all, we will import pandas to read our data from a CSV file and manipulate it for further use.

We will also use numpy to convert out data into a format suitable to feed our classification model.

We’ll use seaborn and matplotlib for visualizations.

We will then import Decision tree, Naïve Bayes ,Logistic Regression algorithm from sklearn.

This algorithms will help us build our classification model.

IV. ALGORITHMS

A. Naive Bayes:

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other.

It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose.

It works well for the data with imbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem.

Using Bayes theorem posterior probability P(C|X) can be calculated from P(C),P(X) and P(X|C)

Therefore, $P(C|X) = (P(X|C)\ P(C))/P(X)$

Where,

P(C|X) = target class's posterior probability
P(X|C) = predictor class's probability.
P(C) = class C's probability being true.

P(X) = predictor's prior probability.

B. Decision Tree Classifier:

Decision Tree algorithm belongs to the family of supervised learning algorithm. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation

Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

Decision Tree Algorithm Pseudo code-

- Place the best attribute of the dataset at the root of the tree.
- Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for anattribute.
- Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of thetree.
- In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the nextnode.

We continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.

Assumptions while creating Decision Tree

The below are the some of the assumptions we make while using Decision tree:

- At the beginning, the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building themodel.
- Records are distributed recursively on the basis of attributevalues.
- Order to placing attributes as root or internal node of the tree is done by using some statisticalapproach.

. The popular attribute selection measures:

- Informationgain
- Giniindex

Gini Index:

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower gini index should be preferred.
Overfitting:

Overfitting is a practical problem while building a decision tree model. The model is having an issue of overfitting is considered when the algorithm continues to go deeper and deeper in the to reduce the training set error but results with an increased test set error i.e, Accuracy of prediction for our model goes down. It generally happens when it builds many branches due to outliers and irregularities in data.

Two approaches which we can use to avoid overfitting are:

- Pre-Pruning

- Post-Pruning

Pre-Pruning:

In pre-pruning, it stops the tree construction bit early. It is preferred not to split a node if its goodness measure is below a threshold value. But it’s difficult to choose an appropriate

.

Post-Pruning:

In post-pruning first, it goes deeper and deeper in the tree to build a complete tree. If the tree shows the overfitting problem then pruning is done as a post-pruning step. We use a cross-validation data to check the effect of our pruning. Using cross-validation data, it tests whether expanding a node will make an improvement or not.

C .Logistic Regression:

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression.

The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application is about to predict the likelihood of a homeowner defaulting on amortgage.

Conditional random fields, an extension of logistic regression to sequential data, are used in natural language process.

V. EXPECTED RESULT

The goal of our project is to know whether patient is diabetic or not, patient will be diagnosed and it will be depending on the attributes that we are going to take, such as age, pregnancy, Glucose, body mass index (bmi), blood pressure ,skinthickness ,insulin ,pedigree function ,outcome i.e. the factors which are majorly responsible for diabetes. So, to reduce the correctly know whether the patient is diabetic or not, we are developing a system which will be a prediction system for the diabetes patients. Another best thing about the system is it is will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that we are going touse added the recommendations we are going to provide based on the diabetic levels of the patients. Also, the prediction of the disease will be done with the help of Bayesian ,Decision Tree ,Logistic Regression Algorithms

VI. CONCLUSIONS

By our in-depth analysis of literature survey, we acknowledged that the prediction done earlier did not use a large dataset. A large dataset ensures better prediction.

Also what it lacks is recommendation system. When we predict we will give some recommendation to the patient on how to control or prevent diabetes in case of minor signs of diabetes.The recommendations would be such, that when followed it will help the patient. Thus we will build up a system which will anticipate diabetic patient with the assistance of the Knowledge base which we have of dataset of around 2000 diabetes patients and furthermore to give suggestions on the premise of the nearness of levels of diabetes patients. Prediction will be done with the help of algorithms Naïve Bayes ,Decision Tree and Logistic Regression will compare which algorithm gives better accuracy on the basis of their performance factors. This system which will be developed can be used in HealthCare Industry for Medical Check of diabetes patients.

VII. FUTURE SCOPE

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes:

1. Increase the accuracy of thealgorithms.
2. Improvising the algorithms to add more efficiency of the system and enhance itsworking.

3. Working on some more attributes so to tackle diabetes evenmore.
4. healthcare diagnosis system to beused inhospitals.

REFERENCES

[1] Dr Saravanakumar , Eswari, Sampath, Lavanya "Predictive Methodology for Diabetic Data Analysis in Big Data," ELSEVIER, ISBCC 2015.

[2] V. H. Bhat, P. G. Rao, P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques,Architecture," Springer-Verlag Berlin Heidelberg, pp. 328- 335, 2009.

[3] Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly "Diagnosis of Diabetes Using Classification Mining Techniques," IJDKP Vol.5, No.1, January 2015.

[4] Sabibullah M, Shanmugasundaram V, Raja Priya K, "Diabetes Patient's Risk through Soft Computin Model,"International Journal of Emerging Trends Technology in Computer Science, vol 2(6), 2013.

[5] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.