Enhancing Diabetic Prediction using Machine Learning Algorithm

R Kranthi Kumar
Assistant Professor
Computer Science And Engineering
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad,India

Polisetty Sai Teja
Student
Computer Science And Engineering
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad,India

Selvaraj Rubin Kumar
Student
Computer Science And Engineering
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad,India

Reddy Sai Vikranth
Student
Computer Science And Engineering
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad,India

Nabilla Riyaz Ahmed
Student
Computer Science And Engineering
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad,India

***Index Terms*— Decision Tree Classifier, Flask Framework, Google Cloud, Pre-Processing.**

I. ABSTRACT

Diabetes is one of the most commonly found disease in the Indian population which if found in the initial stages or if the continues efforts were kept in the treatment, the disease could be cured. But because of the carelessness and lazy treatment of the disease it takes much cruel form of the disease. One of the main reason that diabetes takes its cruel form is because of its time consuming regular checkups based on which diet and exercise are dependent. The solution provided in this paper takes care of the problem by providing a platform where user can continuously have a check on individual's health state. In the solution Decision tree model is used to identify the end user's diabetes category, by providing application on various platform built using ML algorithm and python language. And so from now on instead of waiting/wasting for a long time for checkup can be avoided by making use of the application. This Solution considers all possible categories of diabetes and provides greater accuracy then the existing application.

II. INTRODUCTION

Diabetes is caused throughout the globe. This is caused due to more or less secretion of insulin rather than secretion in the correct amount as per the body requirement. The most common category of diabetes are Type 1 Diabetes and Type 2 Diabetes.

In this paper we are taking care of 6 categories of diabetes namely Type 1 Diabetes, Type 2 Diabetes, De novo Type 1 Diabetes, De novo Type 2 Diabetes, MODY and GDM. The last four categories are found very rarely.

According to surveys we found out that if proper diet and proper procedures to cure diabetes is taken care in the early stages then it is easily cured then compared to when it is on the later stage, As the body is habituated to different environment, and functionality of insulin secretion is changed by the liver due to long term improper functionality.

III. BACKGROUND

A. Dataset Development

As the Patients both Diabetic and non-Diabetic Patients are continuously visiting the hospitals regularly. To keep an eye on patient's health in order to provide with as much apt treatment as possible, [1]Hospitals make sure to maintain a continues reports about the patient. Based on these entries for

over more than two years, Reports of Diabetic Patients are considered and formed as a tuple of the Dataset consisting of 25 attributes and the attribute mentioning the category of Diabetes a particular patient belonging to the tuple, is suffering with. In this Dataset Initially we had the attributes as follows: [2]Admission Date of the Patient, Patient Id, Age ,Gender, Weight in Kg's, Height in cms, HIP, Blood Pressure (BP), Pulse, Index of Body Mass (BMI), Fasting Blood Glucose (FBS), PPBS, hemoglobin A1c (HBa1c), Thyrotropin (TSH) is an hormone secreted by a pituitary gland , Triiodothyronine (T3), iodothyronine tetra, or T4, T3 and T4 controls our body's metabolism, Triglyceride (TGL) it is The body's main form of fat stored and If more than 500 mg are found, dl may cause pancreatic inflammation, Total cholesterol (T.Chol), high lipoprotein density (HDL), low lipoprotein density (LDL), urea of the blood (B.Urea), Serum creatinine (S.Creat) : This is an important rental health [3]indicator because it is an easy to measure by - product of muscle metabolism that is excreted unchanged by the kidneys, Uric Acid, Category of Diabetes consisting of 6 different Values as mentioned above.

B. Tools and Libraries

For Model Creation, which includes data set processing such as using various techniques such as standardization, visualization and finally applying machine learning algorithms and choosing the most appropriate one, all this is done by using python libraries on the jupyter platform as part of the anaconda navigator.

Then, for the creation of web-app flask framework is used which is again making use of python environment, for the creation of mobile-app android studio is used and both are making use of model stored in the pickle file based on protocol 2.

C. And, finally, Google's provided deployment cloud environment is used, and so the scaling and handling of request / traffic is automatically taken care of.

D. Preprocessing

Dataset is initially freed with columns having values less 15% of filled values, since imputation based on mean, median, mode or imputation based on correlation will fail in these kind of cases. Then preprocessing here takes care of datatypes of the remaining columns in order to maintain compatibility among all the attributes for sake of inter and intra preprocessing.

[4]The next step is to impute values in each attributes based on its corresponding scientific nature but on the other hand to remove the redundancy correlation analysis is used.

[5] And the value of diabetes type is generalized into its original form, which was treated as separate due to differences in grammar and non - logy.

If value of correlation analysis is found to be greater than 0.5 between and two independent attributes then one of them is removed and if there is a correlation values less than -0.5 between any independent attribute and dependent attribute then the independent attribute is removed. And after correlation analysis.

Finally, preprocessed dataset is allowed to go through the applicable machine learning models.

E. Machine Learning Models

Here we considered five algorithms for building up of our model, algorithms are as following: [6] Logistic regression, k closest neighbors, Vector Machine Support, Naïve Bayes and decision tree algorithm. In K nearest neighbor, value of k is looped from 1 to 50 but as per the nature of dataset, we found k nearest algorithm is working to its finest behavior at k value = 25. Accuracy provided by each algorithm as follows:

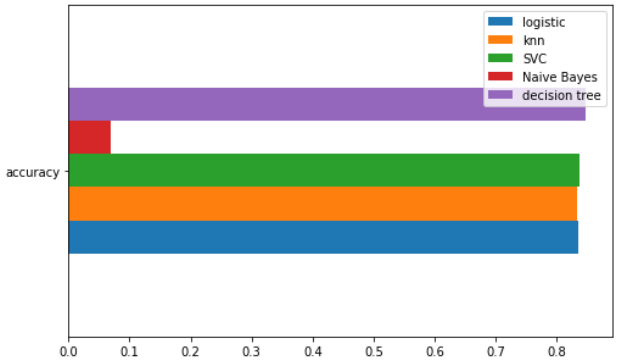

Fig 1: Accuracy Provided by each Algorithm

Accuracy by logistic regression: 82.67%
Accuracy by KNN: 84.54%
Accuracy by SVC: 84.45%
Accuracy by decision tree: 85.35%

Chosen Model:

So, [7]we found out that decision tree is providing us with highest accuracy on implementing decision tree based on Gini index with minimum sample split as 30.

IV. METHODOLOGY

Gini Index:

Decision tree model which is finally choose to build a model is internally making use of the following algorithms:

ID3, Gini Index, Chi - Square and ultimately Variance Reduction.

[8] This ID3 algorithm uses a top - down, greedy search to build decision tree within the space of possible branches without backtracking. ID3 uses Entropy and Information Gain internally.

Entropy play its role in calculating The sample's homogeneity. If the sample is completely homogeneous, the entropy is zero and if the sample is divided equally, it will have one entropy.

Entropy has to be calculated in two ways: one on the dependent variable as follows: (on target variable)

E(S) = sum i=1 to c (-(pi)*log2 (pi))

And other between independent and dependent variables: (branch wise)

E (T, X) = sum c belongs X P(c) E(c)

Information Gain is based on the decrease in entropy following the splitting of a data set on an attribute. Building a decision tree is about finding an attribute that returns I, e's highest gain of information. The branches are homogeneous. Our Information Gain is simply difference of entropy as mentioned below:

Gain(X, Y) = Entropy(X) – Entropy(X, Y)

As the decision node, we will select the attribute with the greatest gain of information. Further division of the dataset and the same process is repeated on the basis of the optimal behavior required or until pure branches are obtained.

Gini Index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

It is suitable for our model as out output is of type categorical value. It performs binary splits and calling it in recursion could allow us to apply it on multi-categorical values as well. Higher Gini's value is higher than homogeneity. Classification and regression tree creates binary splits using the Gini method.

Based on the formula sum of the square probability for success and failure (p^2 + q^2), sub node calculation is done. Now, calculate Gini for split using Gini's weighted score of each split node.

Minimum sample leafs: As splits increases correspondingly number of leaf nodes will also increase due to their direct proportional relation with the number of splits , so choosing the optimal value for the max number of leaf nodes after which further division of decision tree should be stopped is very important, as choosing less number of leaf nodes will lead to elimination of too many sub model cases and will lead to decrement of accuracy and increase of the number of leaf nodes will consume unnecessary amount of load in training and providing nearly negligible amount of improvement of accuracy which is not required, hence for this model we have choose number of leaf nodes as 30.[9][10]

V. UNITS

The model which is being used here is considering the attribute Age in terms of years and its value is restricted from the age of 2years to the age of 100years,

The next attribute that is Gender has a drop down box and is consisting of only two values Male and Female, followed by the attribute weight which is considered here in terms of KG's and is restricted between 10KG to 200KG,then coming to height it is considered in terms of centimeters (cm's). Talking about Body Mass Index (BMI) it is considered in terms of kg/m^2 and is restricted in the range of 10kg/m^2 to 50kg/m^2. Now, next attribute i.e. fasting blood glucose (FBS) is considered in terms of millimoles per liter or millimolar (Mm) and in the application FBS input to the model is restricted in the range of 100Mm to 300Mm.

Postprandial blood sugar (PPBS) is also measured in terms of milimolar and in this application it is restricted to 100Mm to 300Mm units. Hemoglobin A1c (HbA1c) is calculated in Mm percentage, if it is between 4% to 5.6% then the range is normal but if it is found between 5.7% to 6.4% then there is high chance of getting diabetes and if it is greater than 6.4 then the patient is definitely suffering from Diabetes, In the application it is restricted in the range from 5% to 15%. Thyroid-stimulating hormone (TSH) is measured in terms of milli-international units per liter and in the application it is allowed to enter values in the range of 0% to 10%.Finally, Serum Creatinine (S.Creat) is measured in terms of milligrams per deciliter (mg/dL) and in the application it is restricted in the range of 0mg/dL to 10mg/dL.

VI. PROPOSED MODEL

Here we are going to make use of a decision tree which is having a minimum sample split of 30 and minimum sample leafs as 30,

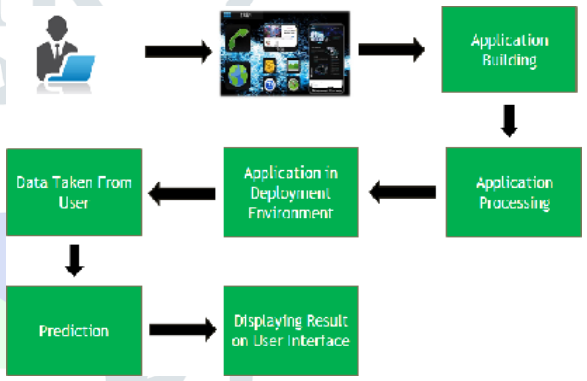

Fig 2: System Architecture

There are two splitting methods available for the classification of decision tree namely "best" and "random," the method used to create our model is "best." In our model we have not restricted branch building to any specific depth level, so our leaves are of type "pure," i.e. each leaf branch will consist of tuples of the same type. And for the formation of leaf nodes, we are considering samples of equal weight. And whenever decisions are to be made every time the number of features that are considered is the maximum number of categories that our model predicts which are six categories. For the random number generation i.e. seed we are making use of random function provided by the numpy mathematical library. And if the impurity value of the node is found to be greater than zero at the point of instance, it will be divided into one more level. This minimum weighted impurity is calculated as follows:

Nt / N * (impurity – NtR/(Nt*rightImpurity) – NtL / Nt * leftImpurity)

Number of samples is denoted here by N, and number of samples given by Nt in the node currently focused, and number of samples given by NtL and Ntr respectively in the left child and right child. And everything is the weighted sum.

VII. Result

Based on the attributes used in the Dataset and type of environment in which it is trained and later on used to predict, It is surely providing comparatively more accuracy that is with 85.35 percentage and is used to predict six category of diabetes then compared to existing ones which are providing a maximum accuracy of 82 percentage, the models which are predicting which accuracy greater than 85.35 percentage are prediction only two categories of diabetes.

VIII. Conclusion

From now Onwards after the initial checkup with doctor, i.e. the end-user or patient is now provided with values of all the attributes required for the individual to get ready for regular checkups from individuals place by making use of web-app or mobile-app which are compatible with all operating systems which internally works based on decision tree algorithm and provides the predicted result on the very same interface that the user is making use of.

IX. References

[1] Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis and Albert-Laszlo Barabasi, "Predicting Individual Disease Risk Based on Medical History".

[2] Darcy A. Davis , Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, Albert-Laszlo Barabasi,"Predicting Individual Disease Risk Based on Medical History" in proc. With CIKM 2018.

[3] Dipanwita Dasgupta and Nitesh V. Chawla, "MedCare: Leveraging Medication Similarity for Disease Prediction".

[4] Wanapol nsuwan, Ureerat Suksawatchon, Kakkarin Suksawatchon, "Improving Missing Values Imputation in Collaborative Filtering With User-Preference Genre and Singular Value Decomposition" in proc. Of International Conference on Knowledge and Smart Technology (KST).

[5] Karsten Steinhaeuser , Nitesh V. Chawla,"A Network-Based Approach to Understanding and Predicting Diseases"

[6] Zhilbert Tafa, Nerxhivane Pervetica Bertran Karahoda, "An Intelligent System for Diabetes Prediction in proc. Of 4th Meditterranean Conference on Embedded Computing" (MECO - 2015).

[7] Wenqian Chen, Shuyu Chen, Hancui Zhang and Tianshu Wu, "A hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree" in proc. Of National Natural Science Foundation of China and Research Fund for the Doctoral program of Higher Education of China.

[8] Veena Vijayan V and Anjali C, "Prediction and Diagnosis of Mellitus – A Machine Learning Approach" in proc. Of RAICS.

[9] Feng Xue-yuan and Li Peng, Computer Aided Diagnosis Based on "K-means Collaborative Filtering Algorithm" in proc. Of International Journal of Hybrid Information Technology.

[10] Hao Ma, Irwin King and Michael R. Lyu, "Effective Missing Data Prediction for Collaborative Filtering" in proc. Of SIGIR 2007