International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

MACHINE LEARNING FOR DIAGNOSIS OF DIABETES

Kavya Mohandas [1], Pragati Patil [2], Nivedita Mhatre [3], Darshan Maskar [4], and Prof. Varunakshi Bhojane [5]

Department of Computer Engineering, Pillai College of Engineering, Navi Mumbai, India-410206

Abstract — *Diabetes is growing rapidly in India, where more than 65.1 million people were diagnosed with this disease, compared to 50.8 million in 2010. Rapid urbanization and lifestyle changes are major causes for the increase in its rapid growth. We wish to reduce the mortality rate of diabetic people by improving patient care, while also reducing the astronomical yearly cost of diabetic patient care. In order to accomplish this, we will create a model to predict whether a person will have diabetes or not based on user inputs. Therefore, the input to our various models would be patient data including details like glucose level, age, blood pressure, etc and we then use a variety of models (e.g. logistic regression, SVMs) to output a prediction as to whether the patient will be diabetic or not. It can help people make a preliminary judgment about Diabetes Mellitus and it can serve as a reference for the diagnosis.*

Keywords— Android Application, Diabetes, SVM (Support Vector Machine), Random Forest Algorithm, Machine Learning.

1. INTRODUCTION

The project is executed on the basis of the primary domain Machine Learning and secondary domain Android Application Development. The programming language used is Python and Java. All the coding of the algorithms will be done on Jupyter notebook to make the machine learning model. This is a classification problem so different classifiers are used on the dataset to classify them into positive (diabetes) and negative (diabetes). Data preprocessing is done to clean the data of any wrong, missing and invalid data if present. After training the model it will be combined with the android application to form an app. Diabetes (diabetes mellitus) - a metabolism disorder. Metabolism refers to the way chemical reactions are processed inside a human being. Major part of the food that we eat is broken down into finer granules of glucose. Glucose is a form of sugar in the blood - the primary source of fuel required for the functioning of our bodies. A person with diabetes has a condition in which the quantity of glucose in the blood [15] is too elevated (hyperglycemia). Because of any of these two cases that either enough insulin is not being produced in our body, or no insulin, or has cells that do not respond properly to the insulin the pancreas produces. This excess blood glucose eventually passes out of the body in urine [15]. So, even though the blood has plenty of glucose, the cells are not getting it for their essential energy and growth requirements.

2. Classification Algorithms

Classification algorithms are one which takes the input and predicts or classify the output based on that input. The main purpose of these algorithms is to classify to which class the given inputs belongs to. So to predict the outcome, the algorithm works on a training set containing a set of attributes and the respective outcome, usually called result or the attribute to be predicted. The algorithm analyses the dataset and discovers relationships between the attributes that would help in predicting the correct output. We have worked on multiple classification algorithms such as kNN, Random Forest, SVM, Decision tree, Gradient Boosting for PIMA Indian diabetes dataset and out of them, two models Random Forest and Gradient Boosting gave highest accuracy. SVM divides dataset into classes. And using a hyperplane placed dataset into a particular class. Hyperplane is the decision boundary where decide which data is placed in which class. So rather than logistic regression we used SVM for implementation.

2.1 RANDOM FOREST ALGORITHM

Random Forest is a supervised machine learning algorithm which can be used for both classification as well as regression. Random Forest starts by creating decision trees on the data points and then extracts prediction from each tree and at the end selects best prediction by means of voting. It is an ensemble method and it is much better than a single decision tree as it helps in reducing overfitting by averaging the results of each tree. Working of algorithm takes place in following steps:

Step 1- Initially start with the selection of random samples from a given dataset

Step 2- Then the algorithm proceeds by constructing a decision tree for every sample and generates the prediction result from every tree

Step 3: In this voting takes place for results returned by every individual tree and finally the best voted result is selected as the result for the particular input.

For random forest to perform well, prerequisites are as follows:

1: There needs to be some actual signal in our features so that models built using those features do better than random guessing.

2: The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

3. System Architecture

The system architecture (see Figure 2) describes the flow of how the project is going to work. The first step in the process is looking for the data required for the model building. Here the dataset used is Pima Indian diabetes dataset.. The next step in the process is

Diabetes Prediction System in which actual analysis and model building will take place which will help in prediction purpose. Here we first analyze the data, then find out correlations between the attributes of the data, then use the algorithm for model building and give new input for cross verification.

The next step is creating an android application using android studio, a server using ngrok, and then we will run the python script through which data would be passed to the model from the app. When the user will enter the details in the app, it would be processed and the predicted result will be displayed on the app whether the patient is having diabetes or not. The attributes given by the user are compared with the algorithm which we have used and the results are generated.

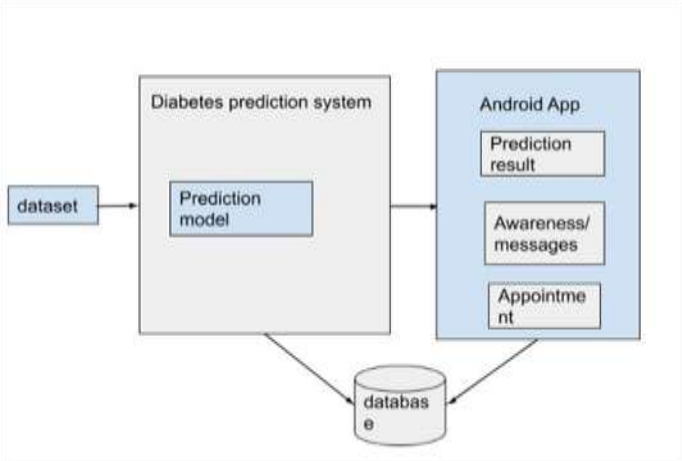

Figure 2: System Architecture

4. Dataset

The Dataset we have used for our project is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. In particular all patients here are females at least 21 years old of PIMA Indian heritage. The dataset consists of 768 rows(instances) with 9 features(columns). Every instance is characterized in data set by 8 attributes.

All attributes are numerical values. Attributes are:

1: Pregnancies-No of times pregnant.

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

3: Blood Pressure-(mm Hg).

4: Skin Thickness-Triceps skinfold thickness (mm).

5: Insulin- 2Hour serum insulin.

6: BMI (Body Mass Index)

7: Diabetes Pedigree Function-function which scores likelihood of diabetes based on family history.

8: Age.

9: Outcome-0 if non diabetic and 1 if diabetic.

ACKNOWLEDGEMENT

We have put in a lot of effort and dedicatedly worked towards completing this report. However, this would not have been possible were it not for the constant motivation and encouragement that we received from the teaching as well as non-teaching faculty of our college.

We would like to express our heartfelt gratitude to our Principal Sir **Prof. Dr. Sandeep Joshi** for letting us explore and providing a magnificent platform for the students.

We are thankful to The Department of Computer Engineering and our Head Of Department **Prof. Dr. Sharvari Govilkar** for providing us with the necessary resources and information.

We extend our thanks to our guide **Prof. Varunakshi Bhojane** for the constant supervision, guidance and motivation during the development of our project without which our project would not have been at par.

REFERENCES

[1] E.Knorr.E and R.Ng, “Algorithms forming distance -based outliers in large datasets”, in proceedings of the 1998 International Conference on Very Large Data Bases (VldB 98), pp. 392-403 New York, 1998.

[2] Han, J Kamber, M: Data Mining; Concepts and Techniques, Morgan Kaufmann Publisher (2000).

[3] S.C.Liao & M.Embrenchts, “Data Mining techniques applied to medical information”, Med.Inform, 2000, pp.81 102.

[4] Kaur H, Wasan SK,” Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science,2(2):194-200,2006

[5] F.C. Li, F.L. Chen, G.E. Wang, Comparison of Feature Selection based on the SVM Classification, IEEE ,2008.

[6] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for heart-related disease diagnosis ." Expert Systems with Applications 35.1 2008.

[7] Fawzi Elias Bekri, A. Govardhan, EMA-QPSO based Feature Selection by LS-SVM for Diabetes Diagnosis, International Journal of Engineering and Advanced Technology (IJEAT), 2012.

[8]Henry Han1,2 and Xiaoqian Jiang3 "Overcome Support Vector Machine Diagnosis Overfitting,2014"

[9] Khyati K. Gandhi,Prof. Nilesh B.Prajapati Diabetes prediction using feature selection and classification, 2014.

[10] Sanakal, Ravi, and T. Jayakumari. "Prognosis of diabetes using data mining methods "Int. J. Comput. Trends Technol.(IJCTT) 11.2 ,2014.

[11] Szakacs-Simon, P. Dept. of Autom., "Transilvania" Univ., Brasov, Romania Moraru, S.A. ; Perniu, L.Android application developed to extend health monitoring device range and real-time patient tracking International Journal of Advanced Research in Computer Science and Software Engineering,2015.

[12] V.Priya, A.Monika, P.Kavitha "Android Application to Predict and Suggest Measures for Diabetes Using DM Techniques", Rajalakshmi Engineering College, Chennai,2015.

[13] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in. IEEE, 2015.

[14] Sakshi Gujral Department of Computer Science & Engineering IGDTUW, Delhi, India Early Diabetes Detection using Machine, 2017.

[15] Deepti Sisodia , Dilip Singh Sisodia National Institute of Technology Prediction of Diabetes using Classification Algorithms, 2018.