INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

DIABETES WARINING SYSTEM USING MACHINE LEARNING

[1] A. LAKSHMAN, [2] S. SANTHOSH REDDY,[3] CH. SHIVA KUMAR, [4] A.HARISH KUMAR

[1.] *Assistant Professor Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.*

Email-: Lakshman.amgoth@gmail.com

[2,3,4.]*B.Tech StudentstDepartment of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Rangareddy (TS).India.*

Email:[2.]*Santhoshreddysoma4@gmail.com*,[3.]*Cheni.shiva1509@gmail.com*, [4.]*Harishallakonda3@gmail.com*

Abstract- Diabetes is taken into account together of the deadliest and chronic disease that causes a rise in glucose. Polygenic disease is that the kind wherever the exocrine gland doesn't manufacture hypoglycemic agent in line with International polygenic disease Federation 382 million individuals live with polygenic disease across the world. By 2035, this will be doubled as 592 million. Diabetes mellitus or just sickness may be a disease caused due to the rise of blood glucose level. Many difficulties might occur if the diabetes remains untreated and unidentified by the doctor. The complications are excretory organ injury, typically resulting in chemical analysis, eye damage that may end in visual impairment, or associate degree enhanced risk for cardiopathy or stroke. The tedious identifying methodology ends up in visiting of a patient to a diagnostic center and consulting the doctor for more treatment. Rise in machine learning approaches solves this essential draw back. The objective of this paper is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. Random Forest algorithms are often used for each classification and regression tasks and also it is a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently, and most importantly, instantly. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and

highest accuracy of 90%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

KEYWORDS: Diabetes, Machine learning Algorithms, Data mining

1.INTRODUCTION

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

2.LITERATURE SURVEY

A total of 255 high-quality data sources, published between 1990 and 2018 and representing 138 countries were identified. For countries without high quality in-country data, estimates were extrapolated from similar countries matched by economy, ethnicity, geography and language. Logistic regression was used to generate smoothed age-specific diabetes prevalence estimates (including previously undiagnosed diabetes) in adults aged 20-79 years.

Results: The global diabetes prevalence in 2019 is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. The prevalence is higher in urban (10.8%) than rural (7.2%) areas, and in high-income (10.4%) than low-income countries (4.0%).

One in two (50.1%) people living with diabetes do not know that they have diabetes. The global prevalence of impaired glucose tolerance is estimated to be 7.5% (374 million) in 2019 and projected to reach 8.0% (454 million) by 2030 and 8.6% (548 million) by 2045.

AUTHORS: J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes

Neural networks or connectionist models for parallel processing are not new. However, a resurgence of interest in the past half decade has occurred. In part, this is related to a better understanding of what are now referred to as hidden nodes. These algorithms are considered to be of marked value in pattern recognition problems. Because of that, we tested the ability of an early neural network model, ADAP, to forecast the onset of diabetes mellitus in a high-risk population of Pima Indians. The algorithm's performance was analyzed using standard measures for clinical tests: sensitivity, specificity, and a receiver operating characteristic curve. The crossover point for sensitivity and specificity is 0.76. We are currently further examining these methods by comparing the ADAP results with those obtained from logistic regression and linear perceptron models using precisely the same training and forecasting sets. A description of the algorithm is included.

3.EXISTING SYSTEM:

Mir and S. N. Dhage have used the WEKA tool for data analytics for diabetes disease prediction on Big Data of healthcare. They used the publicly available dataset from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART. Their approach starts with accessing the dataset, preprocess it in Weka tool and then did the 70:30 train and test split for applying different machine algorithms.

DISADVANTAGES OF EXISTING SYSTEM:

They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

The highest accuracy achieved with their experiment was 76.30%. They have also not practiced Cross-validation.

The existing diagnosis systems have some drawbacks, such as high computation time, and low prediction accuracy.

4. PROPOSED SYSTEM:

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old.

In this project we aim to develop a prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Random Forest Classifier.

ADVANTAGES OF PROPOSED SYSTEM:

Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The proposed system is to perform the diabetes disease prediction using machine learning algorithms for early care of an individual.

The performance of the proposed method in terms of accuracy is high as compared to other states of the art methods and we analyzed it statistically.

In this system we recommend that the proposed method can be used to effectively detect the diabetes disease and the system can be easily incorporated in healthcare.

5. MODULES:

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Diabetes dataset taken from

kaggleLink:

https://www.kaggle.com/c/diabetes-classification

Dataset: In this data set we are taken 9 columns in the dataset, which are described below.

ISSN: 2057-5688

Pregnancies: Number of times pregnant

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

BloodPressure: Diastolic blood pressure (mm Hg)

SkinThicknes : Triceps skin fold thickness (mm)

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

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

DiabetesPedigreeFunction: Diabetes pedigree function

Age: Age (years)

Outcome: Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Data Preparation:

We will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

Model Selection:

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. Solets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.

Here we imported train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset.
The *random_state* parameter seeds random number generator that helps to split the dataset.
The function returns four datasets. Labelled them as *train_x, train_y, test_x, test_y.* If we see shape of this datasets we can see the split of dataset.
We used decision tree classifier, with multiple ensemble methods. Finally I train the model by passing *train_x, train_y* to the *fit* method.

The Decision Tree Algorithm:

The “Decision Tree Algorithm” may sound daunting, but it is simply the math that determines how the tree is built (“simply”…we’ll get into it!). The algorithm

ISSN: 2057-5688

currently implemented in sklearn is called "CART" (Classification and Regression Trees), which works for only numerical features, but works with both numerical and categorical targets (regression and classification). At each node, it determines the feature and split threshold of that feature which will yield the "largest information gain" for the model. This "information gain" is measured based on the splitting criteria specified by the user.

If you are so mathematically inclined, here is the formula for information gain, taken from the sklearn documentation (where G = information gain, n = number of data points on left/right side of the threshold, Nm = total number of data points, H = the chosen splitting criteria function, Q = the data at node m, Θ = the feature and threshold being evaluated)

IMAGE CAPTURING:

In this step used a webcam to acquire the RGB image (frame by frame) and based on only bare hand without glove:

Pre-processing:

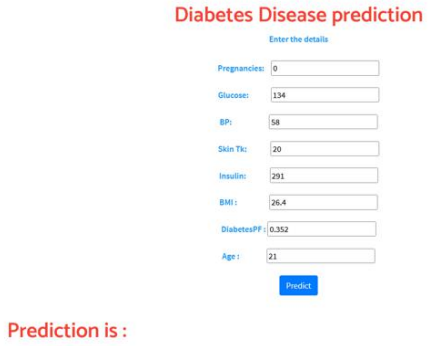
In this step in order to minimize the computation time we took only the important area instead of the whole frame from the video stream and this is called Region Of Interest (ROI). In image processing prefers to convert the color images into a grayscale images to increase the processing and after complete the processing can restore the images to its original color space, therefore, we convert region of interest into a grayscale image. Then blurring the (ROI) by Gaussian blur to reduce the objects that have high frequency but not the target. Notice that in this step the algorithm will fail if there is any vibration for the camera.

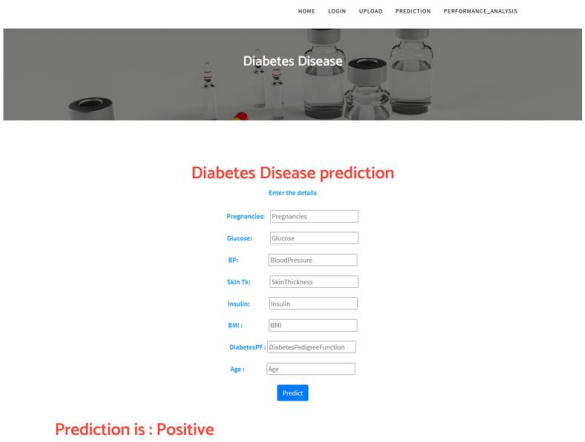
Hand Landmark Detection:

Media Pipe is a framework mainly used for building audio, video, or any time series data. With the help of the MediaPipe framework, we can build very impressive pipelines for different media processing functions.

6.RESULT

ISSN: 2057-5688





7. CONCLUSION

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 90% using the Random Forest Classifier. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well.

8. REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R.Williams, “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition,” Diabetes Research and Clinical Practice, vol. 157, p. 107843, 2019.

[2] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.

[3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

[4] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forcast the onset of diabetes mellitus," Proceedings - Annual Symposium on Computer Applications in Medical Care, vol. 10, 11 1988.

[5] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

[6] Wes McKinney, "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, St´efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.

[7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

1.