A REAL TIME CLOUD BASED MACHINE LEARNING SYSTEM WITH BIG DATA ANALYTICS FOR DIABETES DETECTION AND CLASSIFICATION

Shashikant[1], Anita S Harsoor[2]

[1]*M.tech, Dept. of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Karnataka, India*

[2]*Assistant professor, Dept. of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Karnataka, India*

Abstract

Class one diabetes is one of the major health related issue observed across the globe almost 20 percent of world's population is suffered by type2 Diabetes. Diabetes mellitus is a condition where body stops producing insulin which is responsible for Conversion of sugar into protein. Diabetes does not have any procure and is extremely difficult to identify at primary stage. Not only does it depends upon Sugar Content of body but also various other factors such as age, gender, geography, usage of particular medicine, so it is extremely difficult to diagnose. Therefore sugar content in the body keeps increasing which is harmful for various organs. In this work we present a novel category using Nave Bayesian classifier and KNN Classifier to classify a given set of diabetic parameters which includes Sugar observations such as Post Food, Fasting, and Average, and Gender, age of the patient's record into Normal or Diabetic. Results shows that the system can classify diabetic status with an accuracy of Bayesian classifier 65.25% and 62.5 %of KNN classifier. Our system run from the cloud as an independent kernel and provides an opportunity to classify a given data from anywhere in the World.

Keywords: *Bayesian classifier, KNN classifier, big data.*

----------***----------

1. INTRODUCTION

Diabetes mellitus in simple terms diabetes is a disease in which the blood glucose level of a person is high. This high blood sugar level causes symptoms such as frequent urination; increase in thirst level and increase hunger if diagnosis is not done at the right time. There are basically three main types of diabetes .They are Type1 that results if the insulin failed to be produced by the body, Type2 that results if the cells fail to use insulin properly and gestinational that is presence of high blood sugar level during pregnancy. The existing systems had no. Of drawbacks such as some needed the dataset which is of good quality [1], some needed to standardize the ontology [2], some were only concentrated for women and who were than 21 years [3]. So there is a need to develop a system that can predict diabetes diagnosis the probability of likelihood of getting a normal person suffered with the diabetes for normal person by considering the all factors into account. Neural network is a computational model that has usages in biological and artificial intelligence department for speech recognition, image analysis and diagnosis of diabetes. It reduces the amount of computation required. It helps in identifying diabetes based on earlier occurrences of same problem. The main motivation behind proposing this system is occurrence of diabetes and the fact that many people are unaware of how to diagnose and manage this disease. Mobile health technology could be an effective tool in sharing data, enhancing communication and improving glycemic control while enabling collaborative decision making in diabetes care. In this work, we propose a novel big data analytics based technique of tracking and monitoring the health improvements or detritions of a diabetic patient. The proposed system will ease the patient as they need to consult a doctor and also provides information's about their level of glucose in the early stages thereby decreasing the number of patients affected by Diabetes.

1.1 Proposed System

The proposed system uses the world diabetes forum data to extract the patterns and create a mechanism from the detecting the diabetic from the clinical observation. Provides an analytical overview of the diagnosis result .Finally can classify with the good accuracy level. Our system is primarily divided into two parts:

A. Detection of diabetic from the clinical observations.

B. Statistical and predictive analysis of the diabetic diagnosis process.

A. Detection: Detection is an initial phase which proceeds with the acquisition like post food, fasting, age, gender, average.

B. **Statistical and Predictive Analysis:** This phase involves the statistical and predictive analysis of the given set of records to the system.

1.2 Block Diagram of the Proposed System

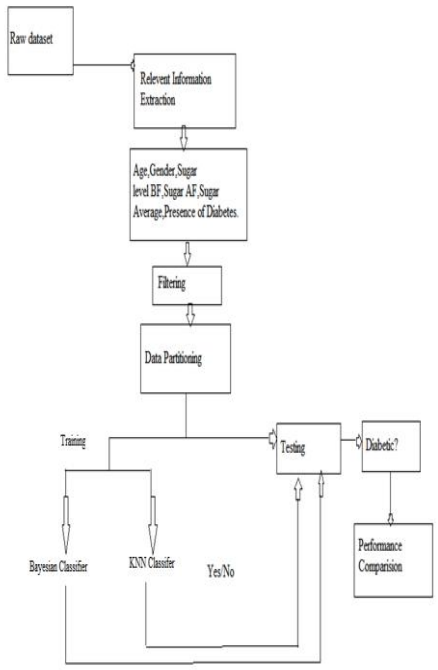

Fig-1: Block Diagram of the proposed system

Block Diagram of the proposed system involves Raw Dataset at the initial stage i.e., gender of the patient male or female, the age group 0 to 10 years or 10 to 20 years soon and it also has 3 features fasting sugar level, post food sugar level and average food sugar level.

In the block diagram the first diagram is raw data set which has been taken from physionet database. There will be extraction of that data set. After importing the dataset extract the relevant information. So there are many fields he has joined the hospital, how many times he attended by the doctor soon and finally there is a field weather he suffering from diabetic or not.

Key part is in two parts mainly gender of the patient, the age group of the patient and three sugar level fields weather that patient is diabetic or not. So these fields are being extracted from the entire data set. Then There will be partitioning of that data .Suppose there present 10,000 patients information, divide it into two parts 5000 and 5000 patients each. One part will be used as training and another part will be used as testing part.

In Training we are going to use two types of classification techniques:
KNN Classifier:
Bayesian Classifier:

1.3 KNN Classifier

As our system follows the pattern recognition technique, it is implemented as in pattern recognition, the K-nearest neighbors algorithm (K-NN) is a non-parametric method used for classification and regression. In both the cases, the input consists of the K-Closest training examples in the feature space. In k-NN Classification, the output is a class membership. An object is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common it's K nearest neighbors. If K=1, then the object is simply assigned to the class of that single nearest neighbor. A confusion matrix or a matching matrix is used is a tool to validate the accuracy of k-NN classification.

1.4 Bayesian Classifier

Bayesian Classifier is one of the techniques used in the machine learning for probabilistic classification. Naive Bayesian classifications are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. Maximum-likelihood training can be done by evaluating a closed form expression .which takes linear time, rather than by expensive iterative approximation.

Using Bayesian therom the conditional probability can be decomposed as (P (Ck) p (x|Ck))/ (P(x))

The above equation can be written as

Posterior=prior*likelihood/ Evidence

1.5 Algorithm

Step-1: Actor Data Formatter
Participant Raw Data
Raw Data->Data Formatter: Format Data
Activate Data Formatter
Step-2: Data Formatter -> Data Splitter: Formatted Data
Deactivate Data Formatter
Step-3: activate Data Splitter
Data Splitter -> Data Splitter: Split
Data Splitter -> Training: Trained Data
Data Splitter -> Testing: Test Data
Deactivate Data Splitter
Step-4: Testing -> Classification: Test data
Step-5: Training -> Classification: Training Data
Activate Classification #FFBBBB
Deactivate Testing
Alt Success
Step-6: Classification ->> Testing: Diabetic
Else failed
Classification ->> Testing: Not Diabetic
End
Testing --> Performance: Actual Result
Step-7: Classification ->Performance: Analyzed Result
Deactivate Classification

1.6 Diagnosis Phases

Following are the Phases of the Diagnosis:
1. Data Formatter
2. Data Splitter
3. Data Filter

4. Training
5. Testing

1. Phase 1: Data Formatter

In data formatter module we will be giving input to formatter. The data formatter extracts the numeric data, text data etc.

2. Phase 2: Data Splitter

These modules the data splitter will split the data into two parts. Testing and Training.

3. Phase 3: Data Filter

Every having data will be having anomalies or outliers. Here in this module we will be using low pass filtering to filter data.

4. Phase 4: Training

Training module will be extracting the features and storing it into the data base.

5. Phase 5: Testing

Testing module will be giving information to the classification system. The classification system includes neural network and support vector. The neural network and support vector machine which will give a output as a patient has diabetes or not

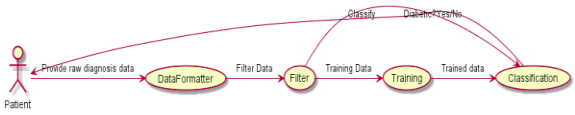

Fig-2: Use case diagram

In this use case diagram first patient will give the data which will get formatted, filtered then training and classifications. Classification takes the data from the training node.

2. RESULTS AND DISCUSSION

2.1 Age Wise Probability Prediction

As Diabetes mellitus is the symptom occurring among all the ages, Fig-3 Shows the age wise Probability prediction when given the data of all the ages based on the knowledge based data. It involves all the age groups from 0 to 100 with the Diabetic percentage.

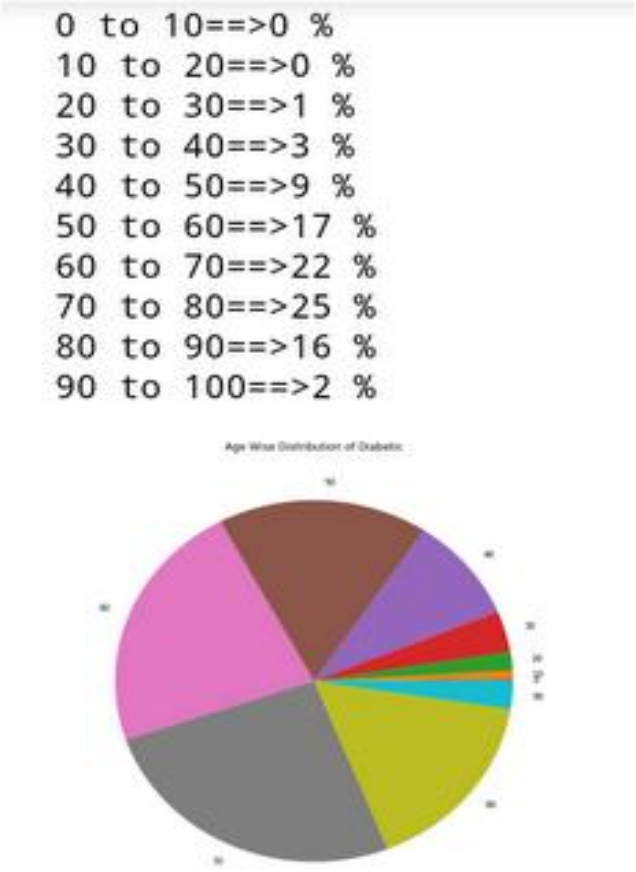

Fig-3: Age wise Probability Prediction

2.2 Accuracy Analysis of the Bayesian Classifier

The Fig-4 Shows that the accuracy analysis of the detection of a person who accessed our system Where in it have various parameters like TP, TN, FP, and FN.

1. True Positive: Is a case where actual the person is Normal & Detected as Normal.
2. True Negative: Is a case where actual the person is Abnormal & Detected as Abnormal.
3. False Positive: Is a case where actual the person is Abnormal & Detected as Normal.
4. False Negative: Is a case where actual the person is Normal & Detected as Abnormal.

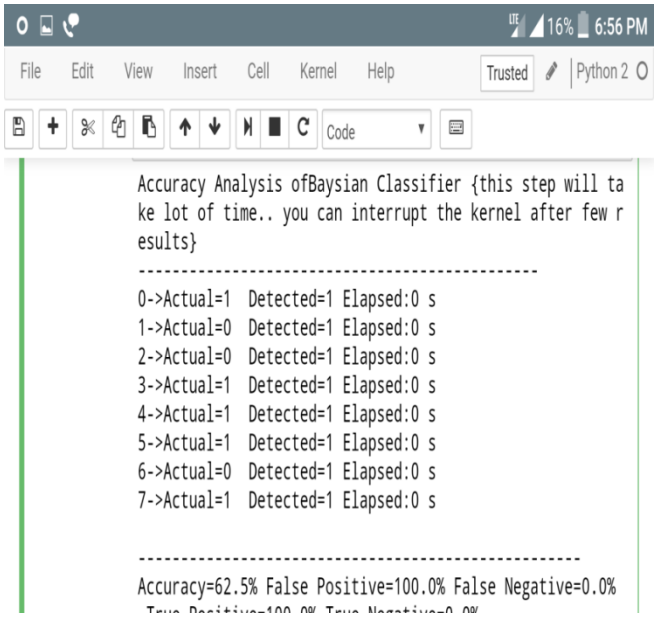

Fig-4: Accuracy analysis of the Bayesian Classifier

2.3 Diagnosis Result of a Particular Patient

The Fig-5 shows the result of a person who accessed our system and check the Diabetes status. It involves the Test class and Detected result of the system.

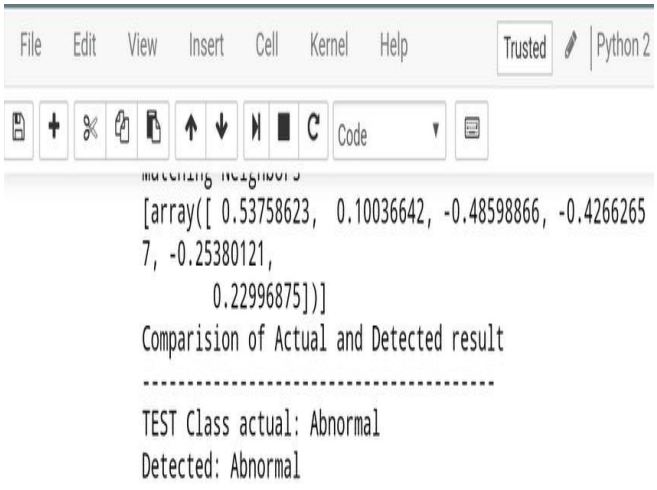

Fig-5: Diagnosis Result of a particular patient

2.4 Diagnosis Results of Various Persons

The Fig-6 shows the Performance of Diabetic detection status of various people. X-axis represents the number of persons Y-axis represents the percentage of detection status. This is the final step in the proposed framework. The expected report generated shows the efficiency of the approaches used and also their accuracy of Detection rate for the Diagnosis of diabetes.

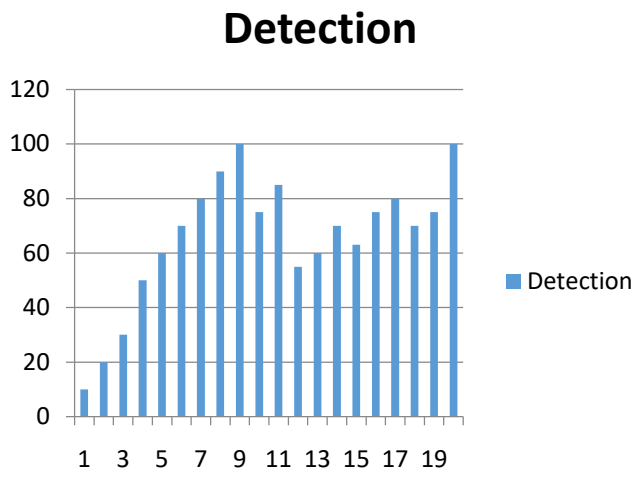

Fig-6: Performance of Diagnosis Results of Various persons.

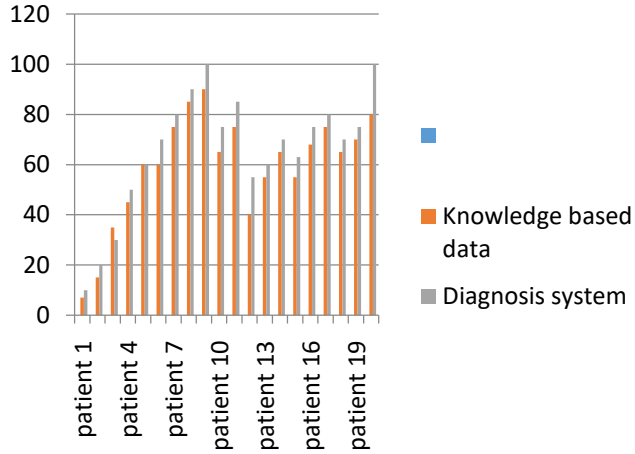

Fig-7: Knowledge based data VS Diagnosis system performance.

The Fig-7 shows the performance graph of knowledge based data with our Diagnosis system. X-axis represents the number of patients Y-axis represents the percentage of detection.

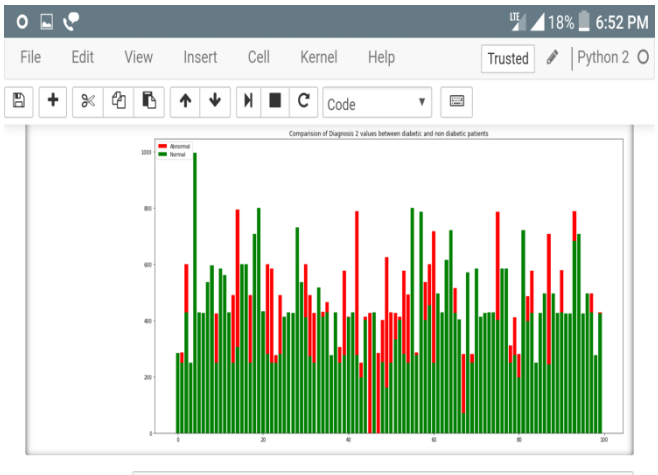

Fig-8: Diagnosis2 output by knowledge based data

3. CONCLUSION AND FUTURE WORKS

3.1 Conclusion

This paper presents a mixture of Bayesian Classifier and KNN Classifier, approaches for detection of diabetes. The proposed system helps in diagnosing diabetes mellitus. With the literature survey performed and analysis carried over, our approaches would serve as a better method for diabetes diagnosis. To conclude, we have proposed a unique Machine learning based system to classify a given records intro Normal or Diabetic. Our System relies on the real time cloud as an independent kernel so by providing an opportunity to diagnose Diabetes from anywhere in the world. The system will ease the patients undergoing medical tests for diagnosing this disease without consulting a doctor thus helping the patients to take precautionary measures well in advance.

3.2 Future Works

Performance of our system can also be enhanced using Re-enforcement of Deep learning and Neural Networks. Our system can also be converted into a custom API which can be invoked by a wide range of applications like a Website or a mobile application.

REFERENCES

[1]. R.Vijayamadheswaran, Dr.M.Arthanariand M.sivskumar, "Detection of diabetic retinopathy using radial basis function", International Journal of Innovative Technology & Creative Engineering, vol 1,pp 40-47,2011.
[2]. LipingZheng,Guangyao Li and Junqing LI, "The Design and Implementation of oral Disease Aided Diagnosis System", Jornal of next Generation Information Technology,Volume 2,pp 38-44,2011.
[3]. Asma A. AlJarullah, "Decision Tree Discovery for the Diagnosis of Type II Diabetes", International Conference on Innovations in Information Technology, pp 303-307,2011.
[4]. Peters AL, Legorreta AP, Ossorio RC, et al.: Quality of outpatient care Provided to diabetic patients. A health maintenance organization experience.Diabetes Care 1996;19:601–606.
[5]. Davis RM, Hitch AD, Salaam MM, et al.: TeleHealth improves diabetes self management in an underserved comm.
[6]. Turner J, Larsen M, Tarassenko L, et al.: Implementation of telehealth support for patients with type 2 diabetes using insulin treatment: an exploratory study. Inform Prim Care 2009;17:47–53.unity: diabetesTeleCare Diabetes .
[7]. Peters AL, Legorreta AP, Ossorio RC, et al.: Quality of outpatient care Provided to diabetic patients. A health maintenance organization experience. Diabetes Care 1996; 19:601–606.
[8]. http://www.mailnor.net/blog/?p=855
[9]. http://www.slideshare.net/Bigdata/preditctive-analytics-with-data
[10]. http://en.wikipedia.org/wiki/python kernel
[11]. http://en.wikipedia.org/wiki/Bayesian Classifier
[12]. http://en.wikipedia.org/wiki/KNN Classifier
[13]. World Health Organization: Global Status Non communicableDiseases.www.who.int/nmh/publications/ncd_report_fl_en.pdf (accessed February2015).
[14]. Inzucchi SE, Bergenstal RM, Buse JB, et al.: Management of Hyperglycemia in type 2 diabetes: a patient-centered approach: positionStatement of the American Diabetes Association(ADA) and the European .
[15]. Joslin Diabetes Center: Joslin Clinical Guideline for AdultswithDiabetes.https://www.joslin.org/docs/Adult_guideline-update_thru_10-23-14_2.pdf (accessed December 15, 2014).

BIOGRAPHIES

Mr.Shashikant has completed B.E.(CSE) from P.D.A College of Engineering and now pursuing M.Tech (CSE) from P.D.A College of Engineering, Kalaburagi, Karnataka. His academic interest area is Data Science, Big Data

Dr. Anita S. Harsoor Asst.Prof.in Department of Computer science and Engineering, P.D.A College of Engineering kalaburagi, Karnataka, India. She has obtained Ph.D. from Gulbarga University Gulbarga. Her research interest is in Digital Image Processing. She has published research papers in reputed journals and presented papers in conferences.