International Journal of Research in Advent Technology (IJRAT) Special Issue
E-ISSN: 2321-9637
Available online at www.ijrat.org
National Conference on "Role of Information Technology in Social Innovations"
26th & 27th February 2019

Predictive analysis of Diabetic Patient Data Using Machine Learning and Big Data

Mrs. Ashwini Abhale[1] ,Shruti Gulhane[2],Sandhya Budhewar[3],Swanali Jathar[4],Harshada Sonwane[5]
Department of Information Technology
D.Y. Patil Collage of Engineering Akurdi ,pune

Abstract- Now days from health care industries large volume of data is generating. It is necessary to collect, store and process this data to discover knowledge from it and utilize it to take significant decisions. Diabetic Mellitus (DM) is from the Non-Communicable Diseases (NCD), and lots of people are suffering from it. Now days, for developing countries such as India, DM has become a big health issue. The DM is one of the critical diseases which has long term complications associated with it and also follows with various health problems. With the help of technology, it is necessary to build a system that store and analyse 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. In this paper, we use the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability

Keywords- Healthcare industry, Hadoop, MapReduce, Machine Learning, Predictive Analysis

I. INTRODUCTION

Computers have brought substantial improvements to technology that lead to the production of massive volumes of data. Healthcare industry contains very large and sensitive data. This data needs to be treated very carefully to get benefitted from it. There is need to develop some more accurate and efficient predictive models that helps in diagnosing a disease although it was revealed that diabetes mellitus is the diseases which becomes one of the global hazards. The problem about this project is it is not easy to do diagnosis whether it is positive or negative having diabetes. It is because of many reasons. Different people may be having different signs. So, it is not easily to assume that they have it or not. The sign of the diabetes is always thirsty, always hungry, weight become decrease, feel weak, have problem of sight, headaches, always do urination and so on. However, the real diagnosis is still needed to assign the real result. Hence there it is needed to analyses the already available huge diabetic data sets to discover some incredible facts which may help in producing some prediction model. The focus is to develop the prediction models by using certain machine learning algorithms. The machine learning is a sort of artificial intelligence that enables the computers to learn without being explicitly Machine learning emphases on the development of computer programs that can teach themselves to change and grow when disclosed to new or unseen data. Machine learning algorithms are mostly categorized as being supervised or unsupervised.

A. Types of Diabetes

Type 1 Diabetes is called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes. Type1 mostly occurs in young people who are below 30 years. This type can affect children or adults, but majority of these diabetes cases were in children.

Type 2 Diabetes is called non-insulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes. Risk factors for Type 2 diabetes includes older age, obesity, family history of diabetes, priorhistory of gestational diabetes, impaired glucose tolerance, physical inactivity, and race/ethnicity. Gestational Diabetes is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood glucose level.

Congenital Diabetes occurs in human due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, and high doses of glucocorticoids leads to steroid diabetes.

II. LITRATURE SURVEY

1. *Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus*

Data mining approach helps to diagnose patient's diseases. Diabetes Mellitus is a chronic disease to affect various organs of the human body. Early prediction can save human life and can take control over the diseases. This paper explores the early prediction of diabetes using various data mining techniques.

2. Predictive Methodology for Diabetic Data Analysis in Big Data

We use the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

3. Predicting Diabetes in Medical Datasets Using Machine Learning Techniques

Different machine learning techniques are useful for examining the data from diverse perspectives and synopsizing it into valuable information. The main goal is to determine new patterns and then to interpret these patterns to deliver significant and useful information for the users.

III. FUTURE SCOPE

In future work pattern matching will be employed by applying discovered patterns on testing data set to predict diabetic prevalent and risk levels associated with it.

A. PROPOSED SYSTEM

In view of the problem statement described in the introduction section, we propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different classifiers like Decision Trees, MLP and Naïve Bayes. The major focus is to increase the accuracy**.**

Advantage of proposed system:

1. Machine learning works independently and takes decision at its own.
2. Through HDFS massive volume data can be access with less time and efforts.
3. Reduce Processing Time.
4. Great accuracy with great performance.

B. SYSTEM SPECIFICATION

Hardware requirements:

• System	:	Pentium IV 2.4 GHz.
• Hard Disk	:	40 GB.
• Floppy Drive	:	1.44 Mb.
• Monitor	:	15VGA Colour.
• Mouse	:	Logitech.
• Ram	:	8gb.

Software Requirements:

• Operating system	:	Windows 10
• Coding language	:	JAVA
• Big Data	:	Hadoop
• IDE	:	Eclipse.
• Web server	:	Apache Tomcat 7.
• Front End	:	JSP, CSS etc.

C. SYSTEM ARCHITECTURE

In our approach, a sequential pattern is used to identify Sentence Element Inference. A sequential pattern is called an indicative extraction pattern (IEP) if it can be used to identify what user want to express and extract structured analysis with perfected information from distributed storage media. We plan to design Auto-Disease Inference based on prediction algorithm. The prediction process starts with a single IEP. We extract a set of initial tokens from the given unstructured statements. For each token and their pairs all questions containing the pair are retrieved from a question collection and regarded as structured questions. From the given the given unstructured statements all possible sequential patterns are generated and evaluated by measuring their reliability score based on Pattern Evaluation.

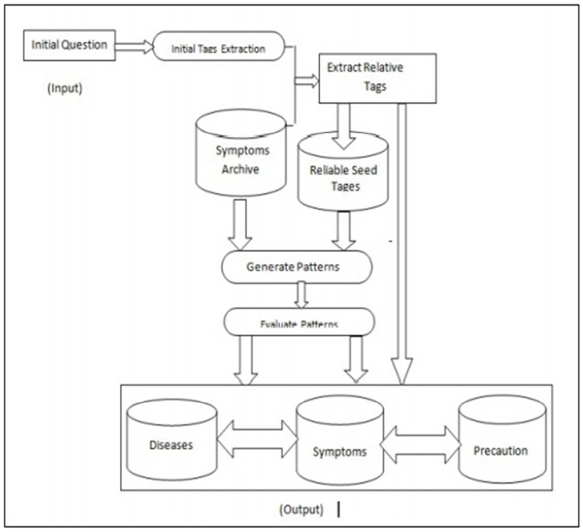


Figure: System Architecture of Proposed System

D. ALGORITHMS

1. Two Thresholds Two Divisors

Step 1: Start

Step 2: Calculate Data length denoted by DLen.

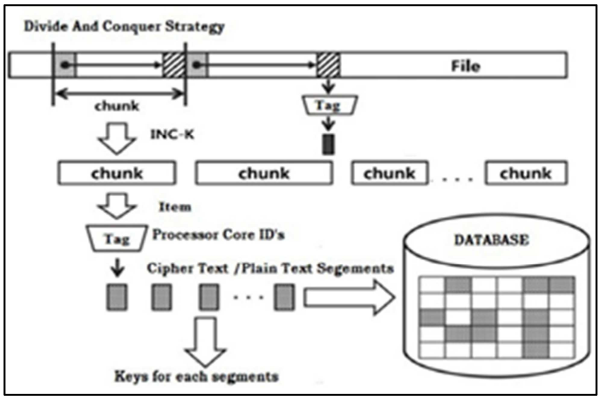
Step 3: Read The total count of thread as no of core of system denoted by NCore.

Step 4: Calculate block size denoted as BSize.

BSize=DLen%NCore==0?DLen/Ncore:DLen/NCore ;

Step 5: SetUp the range of each Chunk start index and (Sx) and end Index (Ex)

Step 6: Stop.



2. MLP (Multilayer Perceptron)Algorithm: -

Step 1. Start

Step 2. Initialize weights at random, choose a learning rate

Step 3. Until network is trained

Step 4. Do forward pass through net (with fixed weights) to produce output(s)

Step 5. Inputs applied

Step 6. Multiplied by weights

Step 7. Summed

Step 8. Squashed by sigmoid activation function

Step 9. Output passed to each neuron in next layer

Step 10. Repeat above until network output(s) produced.

E.PROCEDURE OF PROPOSED SYSTEM

Step 1: At first user will register to the system with his/her basic information then password will be auto generated and it will be provided to user’s email account.

Step2: User will login to the system by entering username and password.

Step 3: After successful login User will be asked to enter the symptoms.

Step 4:Related symptoms will be suggested to the user.

Step 5: If user have any of the suggested symptoms,then user can select from it.

Step 6: If user don’t have any of the suggested symptoms, then user can click on “I have none of the above symptoms”
Step 7: suspected disease will be predicted and it will be shown on the screen.

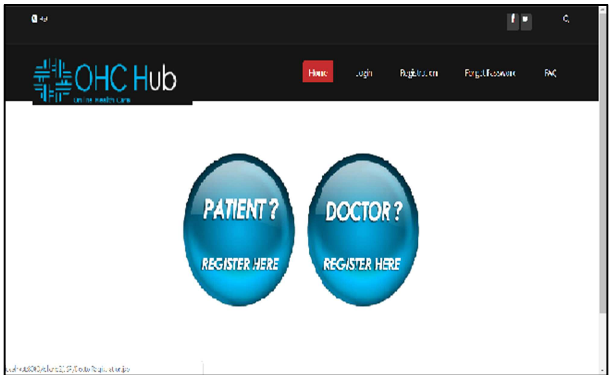
Step 8: Doctor will be suggested according to the particular disease.

Modules:

1. Registration
2. Login
3. Prediction system
4. Hadoop
5. Graph generation

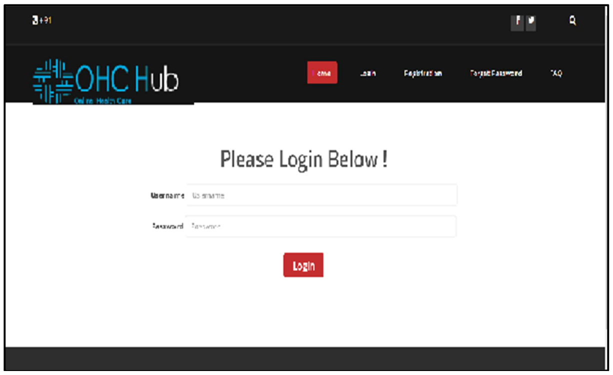
1. Registration

The user will register to the system with normal information. At the time of registration, password will be auto generated and it will be provided to user’s mail.



2.Login

For login to the system, user will enter the Username and password, if entered details are correct then the system will redirect him to home page otherwise it will show an error message.



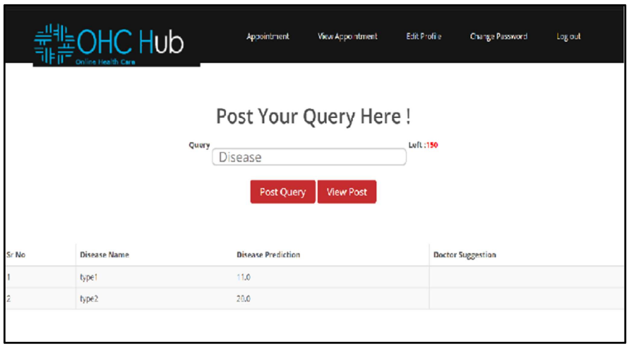
Please Login Below !

Username Username

Password Password

3.Prediction System

- <u>The Disease Prediction system</u>**:**
- It will predict the risk level of diabetes
- It will predict the type of diabetes i.e. TYPE 1,TYPE 2, TYPE 3.
- System will give the highest percentage of correct prediction.



Post Your Query Here !

Query Disease Left :150

Sr No	Disease Name	Disease Prediction	Doctor Suggestion
1	type1	11.0	
2	type2	20.0	

4.Hadoop

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

5.Graph generation

Graph will be generated for predicted disease.

IV. CONCLUSION

Big Data Analytics in Hadoop's implementation provides systematic way for achieving better outcomes like availability and affordability of healthcare service to all population. Non-Communicable Diseases like diabetes, is one of a major health hazard in India. By transforming various health records of diabetic patients to useful analyzed result, this analysis will make the patient understand the complications to occur. The goal of this research deals with the study of diabetic treatment in healthcare industry using big data analytics. The design of predictive analysis system of diabetic treatment may give enhanced data and analytics yield the greatest results in healthcare. By employing location aware healthcare service, anyone from rural area can get proper treatment at low cost. Treatment can be offered when it is identified in advance.

ACKNOWLEDGMENT

We have taken efforts in this project; however, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We are highly indebted to Mrs, Ashwini Abhale for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project. We would like to express our gratitude towards our parents & our Head of I.T. Department Dr.PreetiPatil for their kind co-operation and encouragement which helped us in completion of this project. Furthermore, I would also like to acknowledge with much appreciation the crucial role of the staff of DYPCOE Akurdi, who gave the permission to use all required equipment and the necessary materials to complete my project stage I. We are also deeply grateful to the Principal of DYPCOE, Dr.VijayWadhaiand my parents for their financial and logistical support and for providing necessary guidance concerning project's implementation.

REFERENCES

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

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

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

[4] 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.

[5] Rajnik L. Vaishnav, Dr. K. M. Patel, "Analysis of Various Techniques to Handling Missing Value in Data set," International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 2, 2015

[6] Wei Dai, Wei Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm," International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.49-60

[7] Machine Learning tutorials and examples https://www.toptal.com/machine-learning/machinelearningtheory- an-introductory-primer

[8] Gauri D.Kalyankar, Shivananda R Poojara, N V Dharwadkar,"Weblog Analysis Using Hadoop," National Research Symposium on Computing - RSC 2016, ISBN: 978-81931456-1-8, Dec. 19-20, 2016.

[9]S.Salian and G. Harisekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," International Journal of Science and Research, vol. 4, April 2015.

[10] K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from bigdata using R,"International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.