International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Prognosis of Diabetes Mellitus using Data Mining and other Techniques

Gurwinder Singh[1], Mr. Siddharth Arora[2]

[1]Research Scholar, Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Mullana- Ambala, Haryana, India

[2]Assistant Professor, Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Mullana- Ambala, Haryana, India

Article Info
Volume 8, Issue 3
Page Number : 572-580
Publication Issue :
May-June-2021
Article History
Accepted : 15 June 2021
Published: 25 June 2021

ABSTRACT

Data Mining can be defined as a technology using which valuable knowledge can be fetched out from the massive volume of data. The big patterns can be explored and analyzed using statistical and Artificial Intelligence in big databases. The future trends can be predicted or hidden pattern can be discovered using data mining. Data mining techniques include classification, clustering, association rule, regression, outlier detection etc. The technology of data mining is gaining a lot of popularity in healthcare sector. Many researchers are implementing data mining techniques in the field of bioinformatics. Bioinformatics can be defined as a science of storing, fetching, arranging, interpreting and using information obtained from biological series and molecules. Prediction can be defined as a statement about future event on the basis of present situation. This work focusses on diabetic prediction with machine learning algorithms. The diabetic prediction has various steps. A voting-based classifier is devised in this research to predict diabetes. The performance for the diabetic prediction is optimized using proposed algorithm
Keywords: Diabetes prediction, Data Mining, Anaconda Navigator

I. INTRODUCTION

1.1 Data Mining

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, and reduce risks and more. In the last two decades, massive growth in the volume of data has been stored in databases. Also, and significant upsurge in the use of databases for commercial and technical purposes has been noticed. The successfulness of the interactive model for data storage as well as the evolution and growing of info extraction and handling techniques are the major reasons behind the explosion in the volume of data which is being stored in electronic manner. In the recent times, advanced techniques are being developed

to meet the demand requirements. Also, some efforts have been made for the development of software for data analysis. The companies on the other hand have realized the important of valuable data hidden within these massive amounts of data that was being considered as insignificant till now. The masses of stored data comprise information of a number of factors of different organizations in the offing to be extracted and adopted for supporting the business decision-making procedurewith higher competence. DMS employed the management of these databases in recent time that just permits the client for getting database information in explicit manner. In databases, the stored data represents merely a minor portion of the "mountain of information" existing in it.

1.2 Diabetes

The term "diabetes" is a disease that occurs when the blood glucose in the body, also called blood sugar, is too high. Blood glucose is the main source of energy and comes from the food we eat. According to doctors, diabetes occurs when a gland known as pancreas does not release a hormone called insulin in sufficient quantity. Insulin is a hormone that carries sugar from the bloodstream to various cells to be used as energy. Lack of insulin disrupts the body's natural ability to produce and use insulin accurately. As a result of this, high levels of glucose are released in urine. In the long-term, diabetes when not properly managed can lead to organ failure, cardiovascular diseases and disrupts other functions of the body

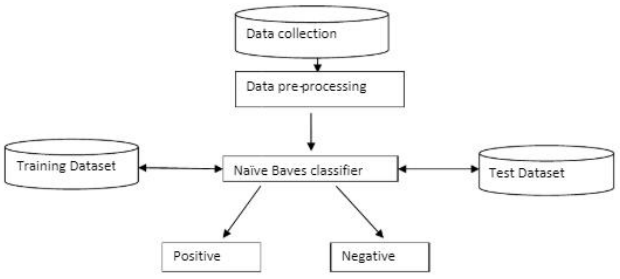

Fig 1.1 Categorisation of diabetic and Non diabetics

The following parameters were used for detecting and classifying the diabetes into positive and negative class, the parameters are: age, insulin, smoke cigarette, age first smoked, etc. Functional Requirement- A functional requirement describes system should do. The functional requirement also specifies the operations and activities that a system must be able to perform. Functional Requirements should include: Descriptions of data to be entered into the system, Descriptions of work-flows performed by the system, Descriptions of system reports or other outputs.

Some of the functional requirement of the proposed system includes:

i. The proposed system will provide a platform to analyze dataset for new patients.

ii. ii. The proposed system will measure dataset for accuracy

II. Literature Survey

Santosh Rani, et.al (2018) studied the data related to the heath that was generated in huge amount at several stages of health system [17]. This data was not processed easily and the analysis of this data was hard to extract because of its huge size. However, the data was processed using various approaches that based on the machine learning. The efficient data was obtained from the approach based on machine learning and this data was useful for the treatment of patients. This approach was also useful for predicting the disease's future. The contribution of past history of patients for various parameters had helped in the possibility of different health issues. In the continuous data, the data mining based on Association clustering and Time Series had utilized to build up the early warning system. When the existing parameters were analyzed, the disease had described using system based on prediction. The patient had saved from the disease at some extent with some level of care.

Rukhsar Syed, et.al (2018) suggested an algorithm that received benefits from the tree-based partitioning [18]. Furthermore, this algorithm utilized the adaptive approach of SVM for the categorization. The preprocessing was carried out in sampling SMORT in this algorithm for pruning the data. The Weka tool had employed for the experiment purpose on the diabetic dataset. The comparison of the suggested algorithm had performed with RF based on tree, RT approach and

J48 approach. The experimental outcomes demonstrated that the suggested algorithm achieved the more effectiveness than conventional solution to process the diabetic data and the proficient categorization was found from that data.

Bhargavi Chatragadda, et.al (2018) focused on the implementation of data mining methodology to forecast diabetes [19]. For getting information from the stored information of dataset was the major target of this data mining methodology. The patterns were also analyzed using it. People faced numerous health related issues and some people were not even conscious about the symptoms of disease. The Diabetic Mellitus was the main health issue. There were many people who were suffered from the Diabetes. This disease was occurred in the people of young generation also. The predictive analysis was employed in HUE for forecasting the diseases. The Pima Indian Database was carried out to gather the dataset. The effective technique was obtained using this framework with SVM classification that proved useful for counting the people who had faced the diabetes.

Yang Guo, et.al (2012) analyzed that the diabetes was a chronic disorder and the main health challenge publicly throughout the world [20]. The data mining techniques were employed to aid people for predicting the diabetes. The Bayes Networks had recommended for envisage the patients who suffered with developing diabetes of type-2. The Pima Indians Diabetes Data Set was used to obtain report related to the patients whether they had developing Type-2 diabetes or not. In this study, Weka software was employed. The outcomes demonstrated that the recommended Bayes network was accurately and efficiently predicted the diabetes of type-2.

Purushottam, et.al (2015) observed that the progressive research area was prediction of diabetes disease in the field of healthcare [21]. The main reasons of diabetes disease were evaluated using number of data mining techniques but only some sets related to clinic risk factors were regarded for analysis. As a result, few efficient factors such as health condition in pre-diabetes were not taken into account during their analysis. Thus, such methods had not provided the suitable pattern and risk factors of diabetes with the exactness in their outcomes.

Ayman Alahmar, et.al (2018) proposed the stacked ensemble methodology that was carried out with deep learning and considered as the meta-learning algorithm [22]. The short as well as long LOS had foreseen for the patients of diabetes. The capability of stacked ensemble methodology was proved in this field by its outcomes. The superior performance for prediction had achieved in the results of algorithms of stacking multiple classification learning as compared to other basic learning algorithmic approaches. The sensible estimation on LOS for the diabetes patients had attained that proved useful to diminish the cost of healthcare and enhanced the contentment of patient who suffered from diabetes.

Geetha Guttikonda, et.al (2019) suggested the relevant data mining methods that were utilized for predicting the diabetes [23]. The extraction of preferred knowledge from the stored information of dataset and the analysis of the patterns of data was the major motive of data mining. The people were suffering from many issues related to health and some were not even conscious about the symptoms of these health problems. The Diabetes Mellitus was one among such health diseases. The young ones were also faced this disease. The prediction of these diseases was done employing the HUE for predictive analysis. The behavior of these diseases was relentless. The Pima Indian database was utilized to assemble the dataset. The persons who had diabetes were calculated using effectual technique that was achieved from the suggested framework together with SVM classification.

III. Anaconda Navigator

To start up idle, log in to the server from an xterm and type IDLE. You will get a Python shell window, which is an ordinary Python interpreter except that it allows some limited editing capabilities. The real power of

idle comes from the use of the integrated editor. To get an editor window for a new file, just choose New Window from the File menu on the Python Shell window. If you want to work with an existing file instead, just choose Open from the File menu, and pick the file you want from the resulting dialog box. You can type text into the editor window, and cut and paste in a fashion that will probably be familiar to most computer users. You can have as many editor windows open as you want, and cut and paste between them. When you are done with your changes, select Save or Save as from the File menu of the editor window, and respond to the resulting dialog box as necessary. Once you have saved a file, you can run it by selecting Run module from the Run menu. You can actually use the integrated editor to edit just about any text file, but it has features that make it especially useful for Python files. For example, it colorizes Python key words, automatically indents in a sensible way, and provides popup advice windows that help you remember how various Python functions are used. As an exercise at this point, you should try creating and saving a short note (e.g. a letter of gratitude to your TA), and then try opening it up again in a new editor window. To exit from idle just choose Exit from the File menu of any window. An especially useful feature of the idle editor is that it allows you to execute the Python script you are working on without leaving the window. To do this, just choose Run Script from the Edit menu of the editor window. Then the script will run in the Python shell window. When the script is done running, you can type additional Python commands into the shell window, to check the values of various quantities and so forth. IDLE has various other powerful features, including debugging support. You can manage without these, but you should feel free to learn about and experiment with them as you go along. Once you have written a working Python script and saved it,say, as MyScript.py, you can run it from the command line by typing python MyScript.py. There is no need to start up IDLE just to run a script.

Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are optional. It has fewer syntactic exceptions and special cases than C or Pascal.

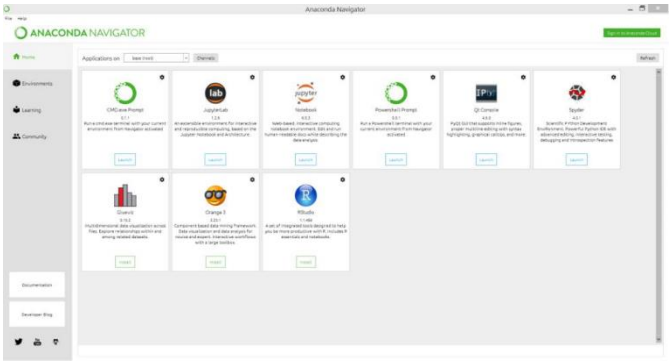

Fig-2: Home page of Anaconda Navigator

IV. COMPONENTS

The Jupyter Notebook combines three components:

• The notebook web application: An interactive web application for writing and running code interactively and authoring notebook documents.

• Kernels: Separate processes started by the notebook web application that runs users' code in a given language and returns output back to the notebook web application. The kernel also handles things like computations for interactive widgets, tab completion and introspection.

• Notebook documents: Self-contained documents that contain a representation of all content visible in the notebook web application, including inputs and outputs of the computations, narrative text, equations, images, and rich media representations of objects. Each notebook document has its own kernel.

V. ARCHITECTURAL DESIGN

System architecture is a conceptual model that defines the structure and behaviour of the system. It comprises

of the system components and the relationship describing how they work together to implement the overall system.

System design is the process of the defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements .Systems design could be seen as the application of systems theory to product development. Object- oriented analysis and methods are becoming the most widely used methods for computer systems design. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user. The UML has become the standard language in object oriented analysis and design.

5.1 Dataflow Diagram

A dataflow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated. DFDs can also be used for the visualization of data processing. A DFD shows what kind of information will be input to and output from the system, how the data will advance through the system, and where the data will be stored.

Fig-3: DFD Level Zero

VI. RESULTS AND TESTING

Software testing is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is Defect free. It involves execution of a software component or system component to evaluate one or more properties of interest.

Software testing also helps to identify errors, gaps or missing requirements in contrary to the actual requirements. It can be either done manually or using automated tools. Some prefer saying Software testing as a White Box and Black Box Testing.

In our system we have done manual testing as well as stress testing to check the breakpoint of the network. The manual testing was done using selenium software while stress testing was done manually with the help of hundreds of nodes that were rented from an online server.

The first Testing was done in the first module i.e. Data Pre-processing which is to ensure that the data set does not contain any missing value or unknown value. The original CSV file is taken as input and data cleansing is performed successfully.

The second and third testing is done in second module i.e. Feature Extraction to reduce the dimensionality of dataset .The pre-processed CSV file is taken and PCA and random forest are successfully applied separately to get the reduced feature dataset.

This chapter gives the outline of all testing methods that are carried out to get a bug free system. Quality can be achieved by testing the product using different techniques at different phases of the project development. The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components sub-assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Fig-4: Pre-processing of given data

6.1 Testing Process

After designing phase there is the coding phase. In this phase, every module identified and specified in the design document is independently Coded and Unit tested. Unit testing (or module testing) is the testing of different units or modules of a system. In this phase, the physical design of the system is converted into the logical programming language.

6.2 Testing Objectives

The coding is done in java before starting of the coding, we have tried to follow some coding standards and Guidelines.

The coding standards are: -

- Naming standards for the Classes and variables etc.
- Screen design standards.
- Validation and checks that need to be implemented. The Guidelines are: -
- Code should be well documented.
- Coding style should be simple.
- Length of function should be short.

6.3 Levels of Testing

Unit Testing- In this, the programs that made up the system were tested. This is also called as program testing. This level of testing focuses on the modules, independently of one another. The purpose of unit testing is to determine the correct working of the individual modules. For unit testing, we first adopted the code testing strategy, which examined the logic of program. During the development process itself all the syntax errors etc. got rooted out. For this we developed test case that results in executing every instruction in the program or module i.e. every path through program was tested. (Test cases are data chosen at random to check every possible branch after all the loops.).Unit testing involves a precise definition of test cases, testing criteria, and management of test cases.

User Input- In User interface the data entry is done through GUI and tested. Each element is tested for valid range and invalid range of data.

Error Handling- In this system we have tried to handle all the errors that are occurred while running the GUI forms. The common errors we saw are reading the empty record and displaying a compiler message, etc.

System Testing- Once we are satisfied that all the modules work well in themselves and there are no problems, we do in to how the system will work or perform once all the modules are put together. The main objective is to find discrepancies between the system and its original objective, current specifications, and system documentation. Analysts try to find molds that have been designed with different specifications, which could cause incompatibility. At this stage the system is used experimentally to ensure that all the requirements of the user are fulfilled. At this point of the testing takes place at different levels so as to ensure that the system is free from failure. Testing is vital to success of the system. System testing makes a logical assumption that whether all parts of the system are correct. Initially the system was given to the user for entry validation was provided at each and every stage, So that the user is not allowed to enter unrelated data. The training is given to user about how to make an entry. While implementing the system it was observed that the user was initially resisting the change, however the system being the need of the hour and user friendly, the fear was overcome. Entering live data of the past months records was little tedious, prior to the actual day to day transaction. The best test made on the system was whether it produces the correct outputs. All the outputs were checked out and were found to be correct. Feedback sessions were conducted and the suggested changes given by the user were made before the acceptance test. Finally the system is being accepted and made to run with live data. System tests are designed to validate a fully developed system with a view to assuring that it meets its requirements. There are three main kinds of system testing:

- Alpha Testing.
- Beta Testing.
- Acceptance Testing.

Alpha Testing: This refers to the system testing that is carried out by the test team with the organization.

Beta Testing: This refers to the system testing that is performed by a select group of friendly customers.

Acceptance Testing: This refers to the system testing that is performed by the customer to determine whether or not to accept the delivery of the system.

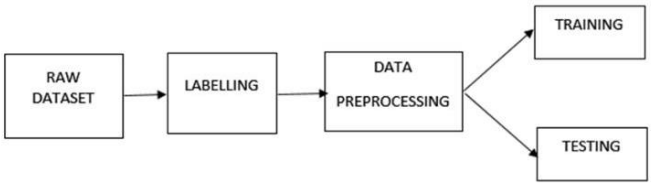

Fig-5: Testing of the given datasets

VII. FUTURE SCOPE AND CONCLUSIONS

7.1 Future Scope

The proposed system can be developed in many different directions which have vast scope for improvements in the system.

These include:

1. Increase the accuracy of the algorithms.
2. Improvising the algorithms to add more efficiency of the system and enhance its working.
3. Working on some more attributes so to tackle diabetes even more.
4. To make it as a complete healthcare diagnosis system to be used in hospitals.

Future work should be done on improving the accuracy of the prediction by increasing the level of training data. Its performance can be further improved by identifying and incorporating various other parameters and increasing size of training.

VIII. CONCLUSION

Data-mining methods have been extensively utilized for predicting blood sugar levels. The data-mining methods do not need strong model suppositions for making prediction models for blood sugar levels. Data mining has the ability to get subtle underlying patterns and associations in experiential data. Therefore, data mining provides efficiently predicts the sugar level within blood. In general, different studies have used data-mining methods for predicting blood sugar blood levels with and without fasting. However, some studies have tried to use data-mining approaches for predicting or classifying the postprandial blood sugar as regular or irregular. In addition, available researches on blood sugar levels in diabetic patients rely on a constant glucose screening system. In this paper, diabetic is predicted in various steps. The algorithm of PCA is employed for the feature reduction. The k-means approach performs clustering of like and unlike data. In the last, the voting classifier method is implemented for the diabetic and non-diabetic prediction. In results, new method shows better accuracy, precision and recall values as compared to existing methods

IX. REFERENCES

[1]. Desmond BalaBisandu, Dorcas DachollomDatiri, Eva Onokpas, Godwin Thomas, Musa Maaji Haruna, Aminu Aliyu, Jerry Zachariah Yakubu, “Diabetes Prediction Using Data Mining Techniques”, 2019, International Journal of Research and Innovation in Applied Science (IJRIAS) | Volume IV, Issue VI

[2]. L.H.S De Silva, NandanaPathirage and T.M.K.K Jinasena, “Diabetic Prediction System Using Data Mining”, Proceedings in Computing, 9th International Research Conference-KDU, Sri Lanka

[3]. Priya B. Patel, Parth P. Shah, Himanshu D. Patel, “Analyze Data Mining Algorithms For Prediction Of Diabetes”, 2017, International Journal of Engineering Development and Research, Volume 5, Issue 3

[4]. Mr. R. Sengamuthu, Mrs. R. Abirami, Mr. D. Karthik, “VARIOUS DATA MINING TECHNIQUES ANALYSIS TO PREDICT DIABETES MELLITUS”, 2018, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 05

[5]. B. Suvarnamukhi, M. Seshashayee, “Big Data Processing System for Diabetes Prediction using Machine Learning Technique”, 2019, International Journal of Innovative Technology

and Exploring Engineering (IJITEE), Volume-8 Issue-12

[6]. Amina Azrar, Muhammad Awais, Yasir Ali, Khurram Zaheer, "Data Mining Models Comparison for Diabetes Prediction", 2018, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 8

[7]. Murat Koklu and YauzUnal, "Analysis of a D. population of Diabetic patients Databases with Classifiers", 2013, International Journal of medical, Health,Pharmaceutical and Biomedical Engineering", vol.7 No.8

[8]. P. Radha, Dr. B. Srinivasan, "Predicting Diabetes by consequencing the various Data mining Classification Techniques", 2014, International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, pp. 334-339

[9]. Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", 2014, ISRASE eXplore digital library

[10]. Veena vijayan, Aswathy Ravikumar, "Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", 2014, International Journal of Computer Applications (0975-8887) vol. 95-No.17

[11]. K. R Lakshmi, S.Premkumar, " Utilization of Data mining Techniques for prediction of Diabetes Disease survivability", International Journal of Scientific & Engineering Research, vol.4 Issue 6, June 2013

[12]. Amira Hassan Abed, Mona Nasr, "Diabetes Disease Detection through Data Mining Techniques", 2019, Int. J. Advanced Networking and Applications Volume: 11 Issue: 01

[13]. Uswa Ali Zia, Dr. Naeem Khan, "Predicting Diabetes in Medical Datasets Using Machine Learning Techniques", 2017, International Journal of Scientific & Engineering Research Volume 8, Issue 5

[14]. D. Jeevanandhini, E. Gokul Raj, V. Dinesh Kumar, N. Sasipriyaa, "Prediction of Type2 Diabetes Mellitus Based on Data Mining", 2018, International Journal of Engineering Research & Technology (IJERT)

[15]. K.Priyadarshini, Dr.I.Lakshmi, "A Survey on Prediction of Diabetes Using Data Mining Technique", 2017, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Special Issue 11

[16]. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes disease prediction using data mining", 2017, International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[17]. Santosh Rani, Sandeep Kautish, "Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction", 2018, Second International Conference on Intelligent Computing and Control Systems (ICICCS)

[18]. Rukhsar Syed, Rajeev Kumar Gupta, NikhleshPathik, "An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction", 2018, International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)

[19]. Bhargavi Chatragadda, SupriyaKattula, Geetha Guthikonda, "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data", 2018, 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)

[20]. Yang Guo, Guohua Bai, Yan Hu, "Using Bayes Network for Prediction of Type-2 diabetes", 2012, International Conference for Internet Technology and Secured Transactions

Cite this article as :

Gurwinder Singh, Mr. Siddharth Arora, "Prognosis of Diabetes Mellitus using Data Mining and other Techniques", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 3, pp. 572-580, May-June 2021. Journal URL : https://ijsrset.com/IJSRSET229129