International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-2, July 2020

Diabetic Prediction using Classification Method

Vimal Sen, Krishna Gupta

Abstract : Prediction analysis of diabetes mellitus is the main focus of this work. There are mainly three tasks involved in prediction analysis. These tasks are input dataset, feature extraction and classification. The earlier framework makes use of SVM and naïve bayes approaches for predicting this disease. This study implements voting classifier for prediction purpose. It is an ensemble approach. This classifier combines three classification models. These models are SVM, naïve bayes and decision tree. The implementation of available and new technique is carried out in python tool. These approaches give outcomes in terms of different performance parameters. In contrast to other classification models, proposed classification model performs better.

Keywords: Diabetic, SVM, Naïve Bayes, Feature Extraction.

I. INTRODUCTION

The technology using which the extraction of valuable knowledge is done from the rough data is known as data mining. The implementation of extraction process is necessary for getting the access of important information. Misnomer is the other name given to this practice. In current scenario, there is the availability of massive volume of data in the majority of areas [1]. It is a very complex and tedious process to analyze such enormous data. Generally, data occurs in rough format. This data is nothing but garbage in the absence of mining. Therefore, there is the need of an efficient data mining approach for extracting valuable information from this data [2]. Mining can be defined as a practice of rough material extraction. In today's technological era, huge knowledge denotes authority and victory. A lot of advanced tool such as processors, space probes etc have been developed. The massive growth of technology in the area of digital storage and processors has made the handling of massive volume of data easy. Ant sort of data can be stored using this technology [3]. All across the globe, enormous data is being gathered and accumulated in the large sized databases. With the passage of every year, this tendency is increasing. Within the research based applications and enterprises, the databases with Terabytes of data are very easy to be found. Depending upon the given input, a certain outcome is predicted by classification. A training set that includes various features and the relevant result is processed through the algorithm such that the possible outcomes can be predicted [4].

Manuscript received on May 25, 2020.
Revised Manuscript received on June 29, 2020.
Manuscript published on July 30, 2020.
***** Correspondence Author
Vimal Sen, M.Tech. Scholar **Yagyavalkya** Institute of Technology, Jaipur, Rajasthan, India vimal295sen@gmail.com
Krishna Gupta, Assistant Professor Yagyavalkya Institute of Technology, Jaipur, Rajasthan, India

This is otherwise known as prediction attribute or goal. For predicting the outcomes, the relationships among attributes are discovered by the algorithm. A prediction set includes the data set which is new and includes the similar sets of attributes [5].

The input is analyzed here and prediction is generated. The efficiency of an algorithm is defined by the prediction accuracy. For instance, the previously generated important information about patients is included in the training set of a medical database. A patient suffered from a heart disease or not is the prediction attribute. Cluster analysis or simply clustering can be defined as a method using which a set of data objects is partitioned into subsets [6]. Every subset is a cluster. Clustering can be described as a set of clusters obtained from a cluster analysis. It means that dissimilar clustering techniques may originate different clusters using the similar dataset. The metabolic disease due to which the blood sugar level of a person increases is called diabetes mellitus. There are the two main causes of this disease. The first cause is the insufficient generation of insulin by the pancreas while the other cause is non-response of cells in generated insulin. Therefore, the unavailability of adequate insulin within patient's body is the main reason of this disease. This disease can be divided in different categories. When pancreas does not generate sufficient insulin or if body is not responding properly to the generated insulin, diabetes can arise. TYPE 2 diabetics is very common in kids and teenagers over the past 20 years [7]. This type occurs because of the obesity in youngsters. According to a survey, around 90% diabetic patients are suffering from this type of diabetes. In this type, pancreas generally generates some insulin. However, either it's not sufficient or the body doesn't use it in efficient way. Resistance in insulin takes place when body cells don't react to insulin. It generally happens in fat, liver, and muscle cells. This diabetes type is quite milder than type 1 [8]. However, it can still cause severe health issues particularly in the small blood vessels in kidneys, nerves, and eyes. This diabetes type increases the chances of coronary disease and stroke. The term fuzzy mean refers to not so clear or vague things. In real time, anyone may face a condition where he is not able to decide about the statement being true or false. At that point, this approach provides extremely important flexibility for reasoning. The doubts of any condition may be taken into account as well [9]. Fuzzy logic algorithm assists to resolve an issue after taking into account all existing data. Afterward, it makes optimum possible decision making for the provided input. This algorithm copies the method of decision making in a human that considers all the probabilities amid digital values T and F.

II. LITERATURE REVIEW

Ioannis Kavakiotis, et.al (2017) studied Diabetes mellitus (DM) in order to detect which several techniques were developed [10]. Today, one of the greatest health challenges being faced is DM and for the prediction diagnosis and biomarker identification lots of research has been done.

There has been huge increment in the data being generated on daily bases with the growth in biotechnology over the years. Thus, to diagnose and treat DM several machine learning and data mining approaches were proposed by researchers. The datasets were generated here by collecting data from the clinics and other biological fields. Improvements were made in techniques to provide better diagnostic results.

Bayu Adhi Tama, et.al (2016) collected and analyzed the clinical record of patients suffering from T2DM. This work made use of KDD methods for extracting knowledge of T2DM patients from a famous Sumatran hospital [11]. An efficient data mining approach was used in this work for carrying out test. The tested outcomes revealed that this approach performed better than other existing approaches in disease detection. DT (decision tree) approach was used in this work for rule extraction. Doctors could use these rules for diagnosing T2DM disease. At last, the optimization of test was carried out. The future work would be focused on using more datasets for maximizing the disease detection efficiency.

Yu-Xuan Wang, et.al, (2017) analyzed various applications across different fields in which the machine learning based techniques are being used [12]. This work was focused on designing several components for the system since the previous researches were based on managing their properties which could be affected with time and usage. The main aim here was to find out a fresh, automatic method for optimizing framework in flexible manner without implementing complicated algorithms. The cache architecture was selected in this work for validating the given suggestions. Here, a decision maker managed to replace the cache content in automatic manner. Afterward, the decision maker replied on a data miner. It made the analysis of data gathered by system screen. A lot of tests were carried out for validating the efficiency of the used approach. This approach provided satisfactory results.

Zhiqiang Ge, et.al, (2017) presented a review in which the author studied the various industrial applications related to data mining. For data mining and analytics, 8 unsupervised learning algorithms were studied here. Further, this work had taken ten supervised learning algorithms into account for inspection [13]. The process business made use of both types of machine learning algorithms (i.e. supervised and unsupervised) in extensive manner. Around ninety to ninety five percent applications used these algorithms. In the recent years, the usage of semi-supervised methods has grown. The future work would be focused on improving this approach for more applications.

Jahin Majumdar, et.al, (2016) stated that Data Mining and ML were the two fields on which a lot of research works were carried out [14]. It was required to retrieve information from databases rapidly along with maintaining growing data volume. It could be very advantageous for carrying research in the area of data analysis and making improvements in industries and markets. In contrast to traditional data mining approaches, ML algorithms had more ability to improve these things through learning from complicated examples. This work was focused on several available schemes based on ML. It was possible to use these algorithms for improving data mining classification and pattern recognition. These algorithms were particularly used for selecting essential features. In this work, the comparison of various available approaches was carried out. After comparison, the selection of optimal approach was done among these approaches. Moreover, a heuristic method was recommended in this work for eliminating the shortcomings of different approaches in theoretical manner.

MS. Tejashrin. Giri, et.al (2014) studied diabetes disease. This disease was related to the metabolism of human body. In this disease, the blood sugar level of a person increases. There were the two main reasons of this disease. Firstly, the pancreas did not produce the required amount of insulin. Secondly, cells did not respond to insulin [15]. The information about this disease could play an important role in decision making by doctors. A lot of researches carried out in healthcare sector made use of data mining for analyzing massive volume of clinical data. The use of data mining method in this disease diagnosis provided high quality results. Hence, it was advantageous to use data mining approaches for testing purpose.

III. RESEARCH METHODOLOGY

The main focus of this work is to predict diabetes disorder which has four phases. The first phase is of pre-process, second phase is of clustering with back propagation algorithm and last phase of classification for final predications. It is a very challenging task to predict diabetes disease because of the availability of huge amount of features in a dataset. Voting classifier combines different classification models. Therefore, various classifiers are combined together for generating a voting classification model. The training and evaluation of all classifiers is done simultaneously in independent manner. In this classification model, the implementation of “hard” or “soft” voting is done. The prediction of last class label is done as the class label which is predicted generally by classifier in “hard” voting.

Different tasks carried out in predicting diabetes disorder by voting based approach are described below:

1. Data collection and pre-processing:- This is the first task. The collection of diabetes data is carried out from the UCI repository in this work. The gathered data is in rough format. Therefore, pre-processing approaches are implemented for transforming this data. This work makes use random sampler for transforming data.

2. Feature Extraction: - The next task carried out in diabetes prediction is called feature extraction. This task may create association among the features of the target set. The features causing maximal effect on the target set can be identified using this approach.

3. Model Building: - The third task is known as model building. In this task, the division of complete is carried out into training and test set. Generally, the size of training set will be bigger than the test set. In this task, the implementation of classification algorithm is done. The training and test set is given as input to the classifier. This work makes use of voting classifier. This classifier combines various other classifiers such as SVM, DT, NB for generating ultimate forecasting outcomes.

The inclusion of a test on a feature is done by DT. An internal node represents it. The branch of DT classifier represents test result while leaf nodes represent classes. SVM is a classification model based on statistical learning. The amount of SVs (support vectors) is identified for representing the dataset. Merely, a segment of dataset is used for the training of classifier. Initially, the designing of SVM classifier has been carried out using binary classifier. NB classification assumes that classes are conditionally independent. It means that features in Naïve bayes classification model do not depend on each other. It is known as "Naive" as it makes the computation simple. This work makes use of three classification models independently. The input to voting classifier is given as the predictions generated by all classifiers. The ultimate outcomes are generated by applying soft voting based approach. This approach generates most accurate diabetes forecasting outcomes.

It is possible to compute Gini Index as:

$$Gini = \sum_{i \neq j} p(i)p(j) \quad \text{--- [1]}$$

Here, i and j are levels of target values. The variable p represents input dataset.

4. Evaluation: - This is the last task in diabetes prediction. In this task, the evaluation of outcomes generated by recommended paradigm is done by considering some metrics.

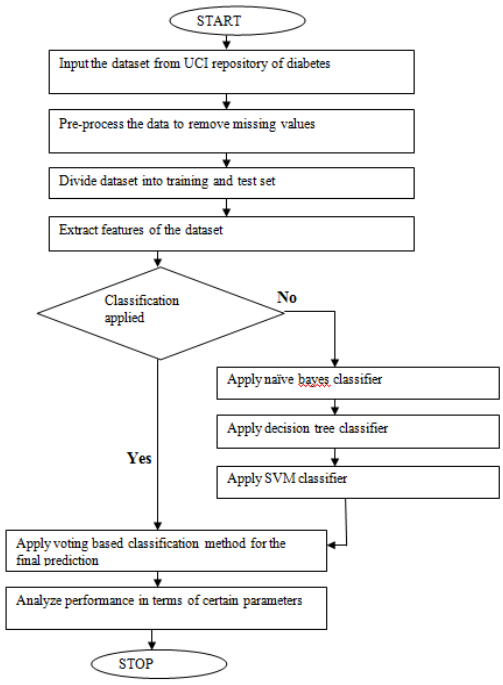

Figure 1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The dataset is taken from the UCI database. In this work, the implementation of two approaches i.e. available and recommended approaches is done. SVM is the existing classification algorithm while voting classifier is recommended approach. The aim of both of these approaches is same i.e. diabetes prediction. In this work, the testing of recommended approach is done for performance analysis.

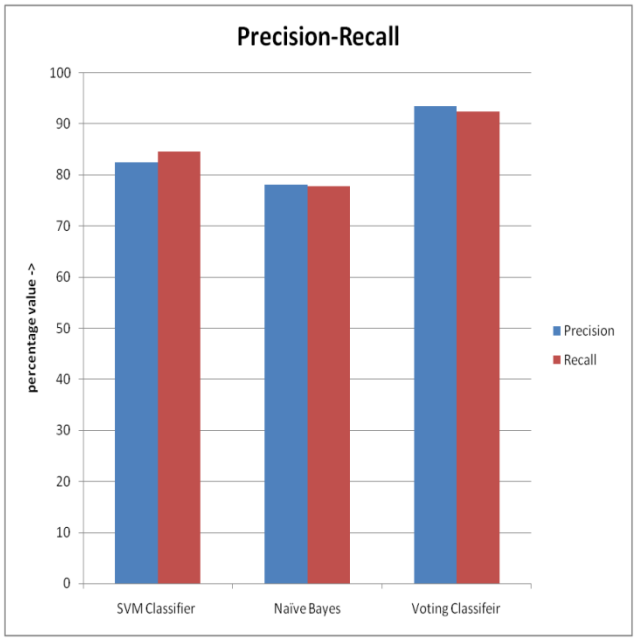

Fig 2: Precision-recall comparison

Figure 2 compares precision-recall values generated by three classifiers. These classifiers include SVM, NB and voting classifiers. Voting classifier is an ensemble classification model. In contrast to two other classifiers, this classifier generates optimal values of precision-recall.

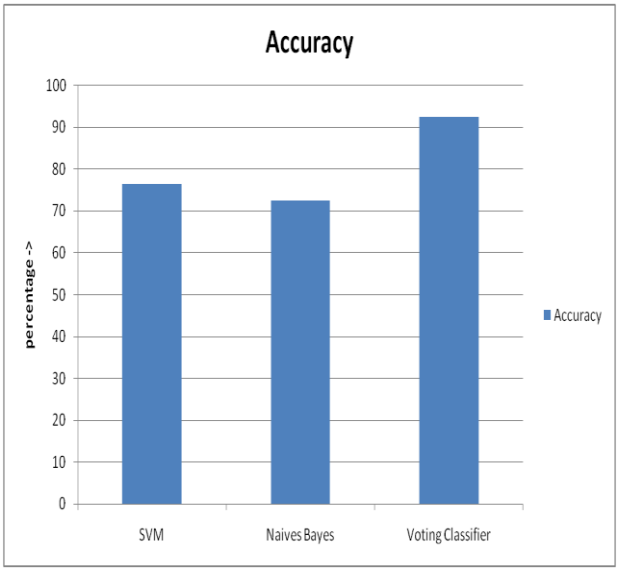

Fig 3: Accuracy comparison

Figure 3 shows the comparison of three classifiers in terms of accuracy. These classifiers include SVM, NB and voting classifiers. In contrast to two other classifiers, voting classifier generates optimal value of accuracy.

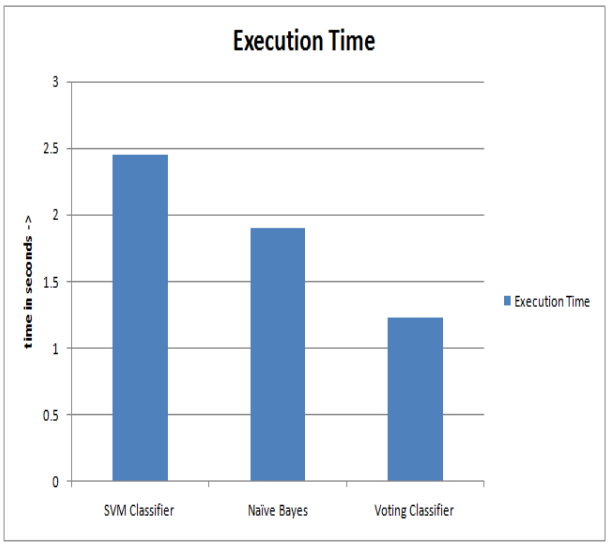

Fig 4: Execution Time

Figure 4 shows the comparison of three classifiers in terms of execution time. These classifiers include SVM, NB and voting classifiers. In contrast to two other classifiers, voting classifier takes minimum execution time.

Accuracy of Voting Classifier is 0.8398268398268398

	precision	recall	f1-score	support
0	0.86	0.91	0.88	151
1	0.80	0.71	0.75	80
avg / total	0.84	0.84	0.84	231

Fig 5: Classification report

Figure 5, the classification results of certain parameters (e.g. precision, recall and f-measure) for the class into 0 and 1.

V. CONCLUSION

Diabetes mellitus is a long-term disease. A lot of people all over the world have lost their lives because of this deadly disease. A survey carried out globally has realized that around 285 million people on this earth have this disease. In current scenario, there is no such technique that can minimize or prevent its consequences wholly. T2DM is a very common type of this disease. It is a very complex task to predict this diabetes type as no existing approach can predict its consequences completely. Most of the research works make use of data mining technology for getting efficient prediction results. This technology also assists in discovering information from data. This work is reached to the conclusion that predicting diabetes disease is a complex problem. This disease can be predicted using various classification algorithms such as SVM, NB. A voting classifier was constructed in this study for the prediction analysis. The classifier combines SVM, naïve bayes and decision tree. This classifier shows better performance than other existing classifiers in terms of different performance metrics.

REFERENCES

1. Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.
2. Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, vol. 7, 2012, pp-56-62.
3. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, vol. 5, 2012, pp. 959-963
4. Ankur Goyala, Vivek Kumar Sharmaa, "Improving the MANET Routing algorithm by GC-Efficient Neighbor Selection Algorithm", International Conference on Advancements in Computing & Management (ICACM-2019),April 13-14, 2019, Jagannath University, Jaipur, India, SSRN: 3446673
5. Ankur Goyal, Dr. Vivek Sharma, "Study of Position Based Greedy Routing algorithm with Interference in the MANET", 2nd International Conference on Emerging Trends in Engineering & Applied Science (ICETEAS' 19) Volume: 5 Issue: 1, ISSN: 2454-4248 04-06, January 2019.
6. Ankur Goyal, Vivek Kumar Sharma, "Modifying the MANET Routing algorithm by GBR CNR-Efficient Neighbor Selection Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019.
7. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, 2017, pp-215-227
8. Akhilesh Kumar Yadav, Divya Tomar and Sonali Agarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.
9. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", vol. 8, 2014, pp. 142-147.
10. Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15 (2017) 104–116
11. Bayu Adhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.
12. Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, "Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization", Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.
13. Zhiqiang Ge, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.
14. Jahin Majumdar, Anwesha Mal, Shruti Gupta, "Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.
15. MS.Tejashri n. Giri, prof. S.r.todamal, "data mining approach for diagnosing type 2 diabetes", international journal of science, engineering and technology, vol. 2 issue 8, 2014.