Monitoring Diabetes Occurrence Probability Using Classification Technique With A UI

Jyotsna Rani Thota, Mahesh Kothuru, Shanmuk Srinivas A, S. N. V. Jitendra M

Abstract: Diabetes mellitus is responsible for not only high health-care costs, but has also become a major cause of death prevailing over the past few decades, accounting for about 2,23,000 deaths per year in India only due to it. Effective management and prediction of Diabetes can allow medical professionals to offer optimal treatment while lowering costs. Proper selections of machine learning algorithms are the parameters for implementation of such a decision support system. In this study, the different models were observed using various classifications, Decision tree techniques for each of the Data Reduction techniques to identify the best approach for diabetes prediction. In this paper, for detecting diabetes at an early stage over the instances of diabetic, non-diabetic and borderline patients we have chosen Gaussian Naive Bayes Algorithm in combination with Information Gain Attribute Evaluation, as we observed that this combination yields the best accuracy and performance.

Index Terms: Gaussian Naïve Bayes, Information Gain Attribute Evaluation, UI enhancement using Python, ROC graph, evaluation menu,

1. INTRODUCTION

Health Informatics (HI) is a multidisciplinary field that uses health information to improve health care. It is applied in many areas like dentistry, physical therapy, nursing, public health, clinical medicine, biomedical research, pharmacy, occupational therapy and alternative medicine, all of which are designed to improve the overall effectiveness of patient care delivery by ensuring that the data generated is of a high quality. Health informatics tools include computer systems, clinical guidelines, formal medical terms and communication system. It has been improving with the advent of data mining and other soft computing techniques. One such application is basically classifying the problem based on its parameters into either of the two classes, presence or absence of diabetes. Generally, diabetes is of two types – Type 1 and Type 2. Type-1 diabetes is caused mainly due to the decreased amount of insulin produced inside our body which results in high blood sugar levels. Type-2 diabetes is caused mainly due to obesity and when the body resists the effect of insulin. During this research work it is observed different models using multilayer perceptron (MLP), Bayesian networks (BNs), SMO classifier and Decision tree techniques for each of the Data Reduction techniques namely, Information Gain Attribute Evaluation, Principal Components Analysis and Correlation-based Feature Subset Evaluation to identify the best accurate diabetes predictive method. This research work focuses on identifying people's data who suffers from diabetes. For that, we developed a simple UI to monitor the diabetic state of a patient using Gaussian Naive Bayes Algorithm in combination with Information Gain Attribute Evaluation which yielded the best accuracy for the Diabetes dataset. The remaining research discussion in this paper organized as follows: Section-2 briefs Related Work of various techniques for prediction of diabetes, Section-3 describes the Methodology and brief discussion of Dataset used, Section-4 discusses Results and Section-5 determines Conclusion of research work.

2 RELATED WORK

In this study author [1] has includes the diabetes prediction and monitoring system which is designed using ID3 Classification algorithm such that the system will generate a message and bar chart to a patient's mail box by specifying sugar levels of patient. In [2] the research includes diagnosing heart disease by developing a system which is designed using Naive Bayes Classification such that a user will be answering the predefined questions which were implemented in a Web Based application by making health care professionals to take intelligent clinical decisions. After observation among set of researches, in [3] includes disease prediction using K-means machine learning techniques for a structured as well an unstructured data. Author in the study of [4] specifies diabetes prediction using Bayesian classification & decision tree algorithm in such a way that the performance measurements can be applicable to calculate accurate results during prediction process. In [5] the author firstly used naïve bayes for classification of a diabetes dataset and then author used a genetic algorithm for feature selection such that to retrieve 4 attributes among 8 attributes in the dataset and there by performed classification on selected attributes[6].

3 METHODOLOGIES PROPOSED

The procedure proposed is represented in figure-1 below in the form of a diagram. The figure shows the flow of the research conducted in constructing the model.

3.1 Model Diagram

3.2 Dataset Used

- *Jyotsna rani thota is currently working in in GITAM (deemed to be University), INDIA. E-mail: jyotsnarani28@gmail.com*
- *Mahesh k, Shanmuk Srinivas A,SNV Jitendra M is currently working in GITAM(deemed to be Univesity), INDIA. E-mail: mahesh.kothuru@gmail.com,shanmuk39@gmail.com, jitendra.mukkamala@gmail.com.*

3.3 Proposed Algorithm

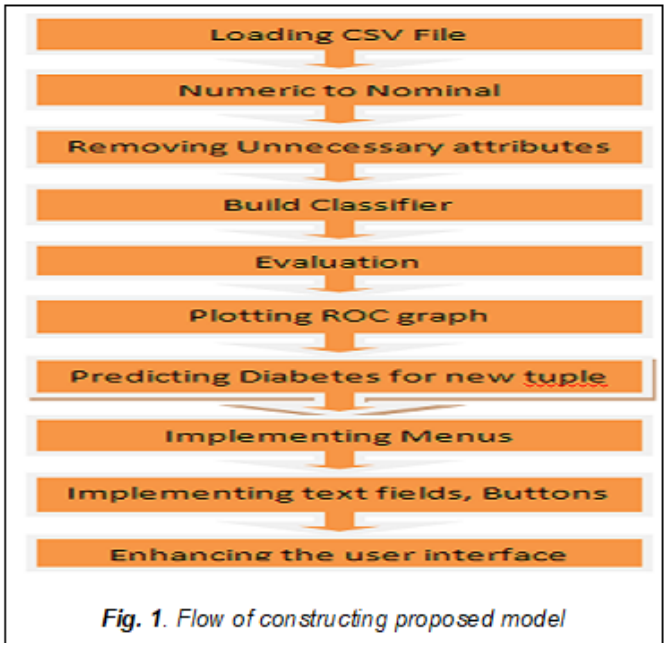

Fig. 1. Flow of constructing proposed model

Step1- Start the Process.
Step2- Collect all patients Dataset as an initial step.
Step3- Load the dataset into a CSV File.
Step4- Preprocess the collected dataset into a valid form.
Step5-Under the step of Preprocess, dataset should be removed with unnecessary attributes.
Step6-Build or Implement Classification algorithms like Gaussian Naive Bayes Algorithm in combination with Information Gain Attribute Evaluation using WEKA.
Step7-Develop a GUI using Python in such a way that, user can enter values with respect to the proposed dataset attributes.
Step8- Run data in the CSV file to predict the accurate Prediction on diabetes for those patients.

TABLE I. INPUT DATASET USED IN MODEL

Feature	Type
Pregnancy count	numeric
Glucose level	numeric
Blood pressure (mmHg)	numeric
Thickness of skin (mm)	numeric
2-hour serum insulin (mu U/ml)	numeric
BMI (kg/m)2	numeric
Diabetes pedigree function	numeric
Age	numeric
Class label (positive -1, negative -0)	numeric

Step9-Added new tuple through GUI to the CSV file executed will also generates message on diabetic prediction to that particular patient.
Step10-Finish the Process.

3.4 Description of proposed Algorithms

Analyzed various advanced machine learning techniques such as Voted perceptron, Bayesian networks (BNs), SMO classifier and Decision tree techniques for each of the Data Reduction techniques namely, Information Gain Attribute Evaluation, Principal Components Analysis and Correlation-based Feature Subset Evaluation effectively, to diagnose Diabetes. Proposed model using Gaussian Naive Bayes Algorithm in combination with Information Gain Attribute Evaluation as we observed that this combination yielded the best accuracy [7].

3.4.1 Tools and Technologies

3.4.1.1 Used Data

Various types of files (CSV file, database file, etc.,) can be loaded into WEKA. In our case we loaded data from a CSV file of 9KB. It was downloaded from https://www.kaggle.com/uciml/pima-indians-diabetes-database website.It consists of 768 tuples each with 8 attributes. Out of 768 tuples, 500 tuples are diagnosed negative and the rest are diagnosed positive for Diabetes. Included few tuples through developed GUI[8].

3.4.1.2 Filters

Here all of the WEKA's filters are available in which there are two types:

1. Supervised
   Here we go to Attribute Selection and select the Data Reduction algorithm we want to perform such as Information Gain Attribute Evaluation, Principal Components Analysis and Correlation- based Feature Subset Evaluation [9].

2. **Unsupervised**
   Here conversions such as numeric to nominal can be done depending on what classifier we want to use.For example, if we want to use Naive Bayes Classifier we have to convert from numeric to nominal because Naive Bayes doesn't support numeric attributes [10].

3.4.1.3 NAIVE BAYES CLASSIFIER:

Naive Bayes is a classification algorithm for binary and multi-class classification problems, as the calculations of probabilities of every hypothesis was simplified to make their calculation tractable. Instead of calculating every attribute value P(x1, x2, x3|h), these were assumed as conditionally independent of each other given the target value and calculated as P (x1|h) * P(x2|h) and soon. Naive Bayes model is represented using probabilities.A learned Naive Bayes model includes, Class Probabilities: The probabilities of each class in the training dataset. Conditional Probabilities: The conditional probabilities of each input value given each class value[11].

4 RESULTS AND DISCUSSION

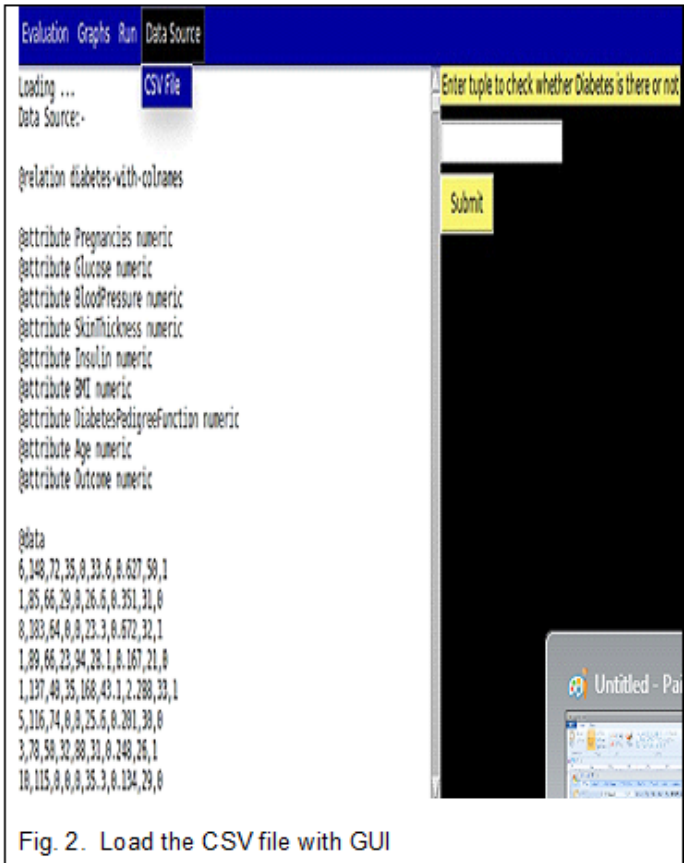

Fig. 2. Load the CSV file with GUI

In the above given results Figure 2 represents the screen which is with GUI developed with python, in right side of the screen there is a textbox which can take diabetes attribute values based up on the specified attributes and in left side of the screen we can find a link to retrieve data source from a CSV file from which diabetes prediction can be done to the overall available data base [12]. Figure 3 clearly represents the diabetes prediction of a person with respect to the given values. Table 2 is a summary of overall correct and incorrect classified instances and also with the overall error rate predictions [13], [14], [15].

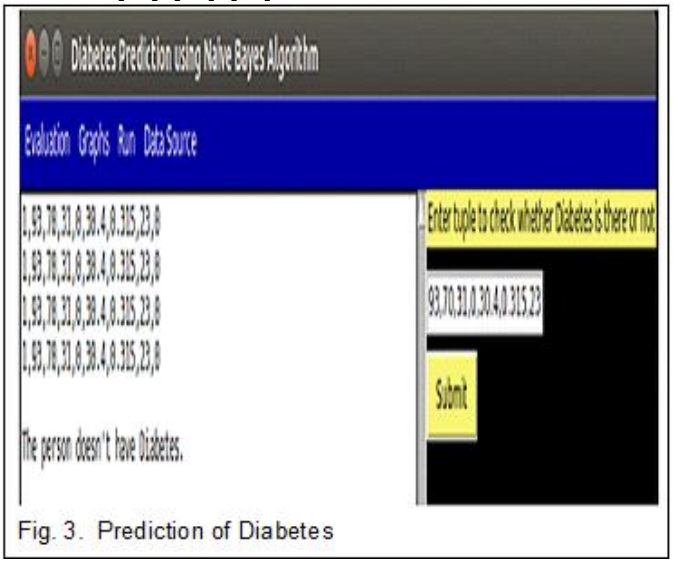

Fig. 3. Prediction of Diabetes

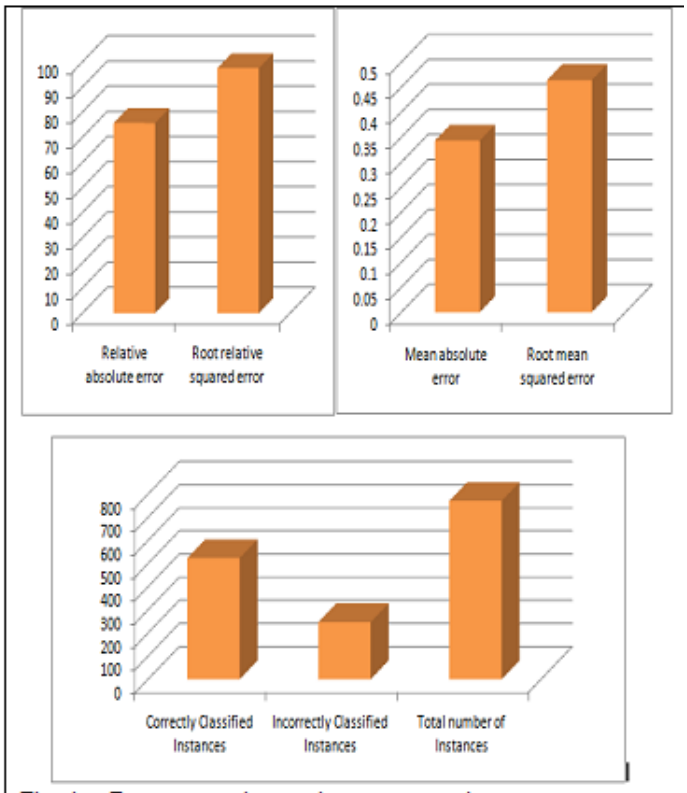

Fig. 4. Representation on Instances and error rate

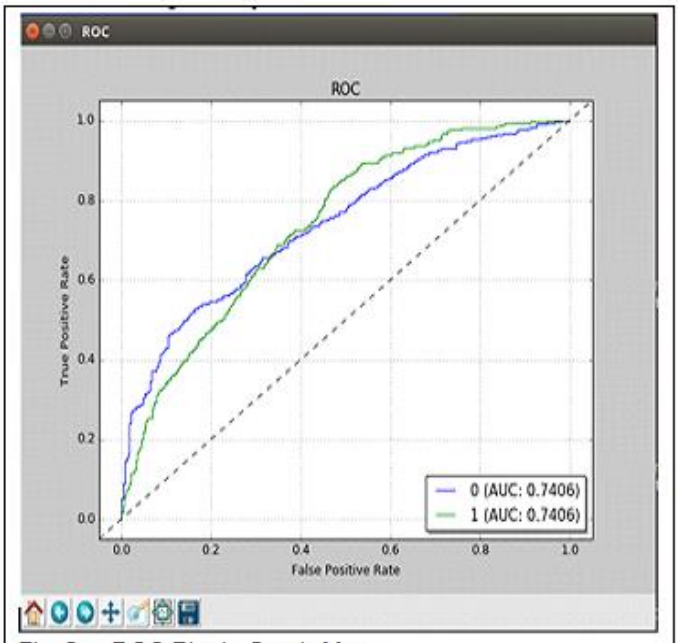

Fig. 5. ROC Plot in Graph Menu

5 CONCLUSION

APPLYING MACHINE LEARNING TECHNIQUES

TABLE I. VISUALIZING RESULTS

Name	Value
Correctly Classified Instances	67.9637%
Incorrectly Classified Instances	32.0363%
Kappa statistic	0.2979
Mean absolute error	0.3432
Root mean squared error	0.4642
Relative absolute error	75.6393
Root relative squared error	97.4847
Total Number of Instances	771

In the medical field has helped to make the systems smarter and reduce the manual effort and errors. This also helps the patients to keep a tab on their health on a regular basis and notify them when there is a change in their conditions and also specify the respective precautionary measures, hence helping them to stay healthy. The existing systems have a lot of disadvantages such as a lot of manual work, no intelligent algorithms for faster computations, no way to effectively monitor on a regular basis at homes, no graphical interfaces for better visualization of the output and longer diagnosing period. So to overcome these problems, we analyzed many algorithms in terms of accuracy and speed and proposed a system to improve identification and prediction of Diabetes. Proposed a classification system using Gaussian Nai Bayes algorithm in combination with Information Gain Attribute that which can predict whether a person has Diabetes or not with an accuracy of 67.96% on a complete dataset. Since the advancement of machine learning and its adaptation in health care industry, advancements in electronic medical records have been remarkable. It automates the existing manual procedures and thereby saving time of patients at the hospital due to reduced computational time. Also as it is system performed, it is error prone. Further advancements can be done to this system such that this can be implemented in most of the hospitals.

ACKNOWLEDGMENT

The authors wish to thank Department of CSE, GIT, GITAM University, and Visakhapatnam for providing infrastructure for project.

REFERENCE

[1] S.,R., Priyanka shetty, Sujatha Joshi., 2016. A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique, I.J. Information Technology and Computer Science. DOI: 10.5815.

[2] S. S. Amiripalli, V. Bobba, and S. P. Potharaju, “A Novel Trimet Graph Optimization (TGO) Topology for Wireless Networks,” in Cognitive Informatics and Soft Computing, vol. 768, P. K. Mallick, V. E. Balas, A. K. Bhoi, and A. F. Zobaa, Eds. Singapore: Springer Singapore, 2019, pp. 75–82.

[3] S. S. Amiripalli, V. Bobba, “Research on network design and analysis of TGO topology”. International Journal of Networking and Virtual Organisations, 19(1), pp. 72-86. 2018.

[4] Shadab Adam Pattekari., Asma Parveen., 2012. Prediction system for heart disease using naïve bayes International Journal of Advanced Computer and Mathematical Sciences. pp 290-294.

[5] Akash C., Jamgade., S., D., Zade., 2019. Disease Prediction Using Machine Learning. International Research Journal of Engineering and Technology (IRJET). Pp 6937-6938.

[6] Panigrahi Srikanth., Dharmaiah Deverapalli.,2016. A critical study of classification algorithms using diabetes diagnosis. 2016 6th International Advanced Computing Conference IEEE DOI 10.1109

[7] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

[8] S. S. Amiripalli and V. Bobba, “An Optimal TGO Topology Method for a Scalable and Survivable Network in IOT Communication Technology,” Wireless Pers Commun, vol. 107, no. 2, pp. 1019–1040, Jul. 2019.

[9] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[10] Esposito, F., Malerba, D., Semeraro, G., Kay, J., 1997. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[11] Amiripalli, S. S., Kumar, A. K., & Tulasi, B. (2016, February). Introduction to TRIMET along with its properties and scope. In AIP Conference Proceedings (Vol. 1705, No. 1, p. 020032). AIP Publishing LLC.

[12] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[13] Garner, S.R., 1995. Weka: The Waikato Environment for Knowledge Analysis, in: Proceedings of the New Zealand computer science research students conference, Citeseer. pp. 57–64.

[14] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.

[15] Amiripalli, S. S., & Bobba, V. (2019). Impact of trimet graph optimization topology on scalable networks. Journal of Intelligent & Fuzzy Systems, 36(3), 2431-2442