PREDICTION OF DISEASES USING VARIOUS MACHINE LEARNING MODELS

S Venkata Achuta Rao[1], G Narayana[2] & Usikela Naresh[3]

#1 Professor, CSE & Dean-Academics, SREE DATTHA Institute of Engineering and Science, Sheriguda, Ranga Reddy, Telangana-501510, India, drsvarao@gmail.com*

#2 Assoc. Professor, CSE, CVR College of Engineering, Ranga Reddy, Telangana-501510, India, narayanag.1973@gmail.com

#3 Research Scholar, CS&T, Sri Krishnadevaraya University, Anantapuramu, AP-515003, India usikelanaresh@gmail.com

Abstract:

Motivated by the rising mortality of Heart Disease, Breast Cancer and Diabetes patients every year worldwide and the handiness of enormous amount of patient's data that could be used to mine useful knowledge, researchers have been using machine learning methods to assist medical professionals in the prediction of heart disease, breast cancer and diabetes. The techniques offer a framework to study the problem of inference that is of getting the knowledge, decision making, doing predictions or developing models from a data set. To predict different diseases like diabetes, breast cancer and heart disease etc., data mining techniques were applied. In this study, classification techniques were applied to extract knowledge from the heart disease, diabetes and breast cancer data sets. The performance measures of these classification algorithms, namely, Naïve Bayes, Back Propagation, Support Vector Machines (SVM), K-Nearest Neighbor and C4.5 were examined by considering the various parameters like accuracy, specificity, sensitivity, precision, error rate, ROC area and time. For evaluating these measures Confusion matrix values were used and stratified 10-fold cross-validation method was applied.

Keywords: *Machine Learning, Diseases, Support Vector Machines, Decision Making, Datasets.*

DOI Number: 10.14704/nq.2022.20.13.NQ88222 Neuro Quantology 2022; 20(13):1786-1792

1 Introduction

Machine Learning offers different techniques, like classification, association rule mining, clustering etc. Based on the type of the problem and the size of the data set we can select appropriate data mining technique. The crucial task in data mining process is to select the right technique and right size of the data set with the attributes to optimize the process of data mining. The performance measures accuracy, sensitivity, specificity, precision, mean absolute error, area under ROC curve and efficiency of each data mining technique is based on the size of the training data set, type of task the algorithm is performing and how well the data set is represented. The current research work is to investigate the data mining algorithms capabilities applied in medical data sets. Data mining techniques were applied to analyze heart disease, diabetes and breast cancer data sets and the accuracy, efficiency and other performance measures of each data mining algorithm were examined. The performance measures of each classification algorithm were calculated and the results were displayed. And these results of the classifiers applied to the medical data sets were compared. If we consider the accuracy, sensitivity, specificity, precision and mean absolute error, then the best model is SVM. If we apply the data mining techniques which were designed for small data sets to the large data sets then there is chance of getting low quality results [31]. The researchers hope this study would enhance the prediction of diabetes, breast cancer and heart diseases by applying data mining techniques and assisting medical scientists, physicians in the treatment process of diabetes, breast cancer and heart diseases.

Data mining algorithms can be applied to many different industries for prediction and classification tasks, like in medicine, engineering and telecommunications etc. Below mentioned issues must be considered when applying the data mining techniques by using the medical data sets:

- Noise in the data
- Missing values in data
- Irrelevant data present in the data set.
- The size and the number of attributes of the data.

In the process of classification the missing values present in the data can produce low quality results in both training and testing phases. The efficiency of the data mining technique can be affected by the size of data or attributes.

Only preprocessed data of numerical input has to be used for this method. It is well suited for the considerable number of attributes. Hence the elimination of unnecessary attributes is a challenging issue in this work.

2 Related Work

Data Mining is one of the most motivating and crucial research areas with an objective of discovering significant information from huge volumes of data sets. In present period, Data mining techniques are becoming favorites in medical field because there is a requirement for an effective analytical methodology to find the unknown and useful information hidden in medical data. Data mining offers various benefits in health industry, such as detecting the unfair practices in medical insurance, various medical treatments availability for curing diseases to the patients at lower expenditure, finding the reasons for various diseases, and detection of best treatment procedures. It can also useful to the healthcare researchers for creating competent medical policies, designing automatic drug recommendation models, preparing the case studies of individual patients etc. [1].

The data prepared by the healthcare organizations is huge and complex because of that it is hard to analyze this data in the process of taking crucial decision about patient's health condition. This data contains details about patients, hospitals, treatment expenditure, medical claims etc. So, there is a demand for generating an effective model for extracting and analyzing essential information from complex data. The analysis of medical data enhances the healthcare practices by improving the performance of management tasks of patients. The results of data mining algorithms are to cater benefits to the medical industry for grouping the patients who are having related type of health issues or diseases so that the healthcare organizations can give them effective treatment. It can also use for estimating the period of stay in the hospital, for diagnosis and preparing a medical plan for efficient information system management. Latest technologies are used in medical field to improve the medical services in reduced cost [2]. But data mining access the database in a different way from this traditional query access. Compared to traditional query the data mining produces an effective outcome and also it can be implemented in an automated manner in huge data set. It is evident from the existing data mining tools that a wide range of data analysis methods are covered by data mining [3]. The medical practitioners can predict or guess effectively various diseases of patients by applying the classification techniques of data mining on medical data sets. The literature presented here was useful in understanding the main theories which provide the basis of analysis of medical data research [4]

Classification or supervised learning algorithms became most popular and familiar because they are simple to understand and easy to implement. Various applications of classification techniques are classifying financial market trends, medical diagnosis, pattern and image recognition, etc. Prediction and estimation can be considered as kinds of classification [5]. The classification process is also called as supervised learning process because the classes were already defined before investing the data. According to Berry and Linoff [6], data mining is considered as analysis and exploration of huge volumes of data in order to detect meaningful and useful rules and patterns. The classification approach needs some knowledge about data. To build the classification model training data set can be used and the model can be validated with the help of test data set.

The data mining process has various techniques. Classification technique is one of the outstanding

techniques to predict the target attribute in data set. Classification process of a collection contains separating the items that make up the collection into different classes or categories. In the data mining context, classification is performed by utilizing a model, which is created depending on the historical data and aim of the classification process is to predict accurately the target class value for every new record in the data set [7].

A Classification task begins with training data (build data) for which the target values (class assignments) are known. Various classification methods were available that uses various kinds of techniques for discovering the relationship between the values of predictor attributes and the target attribute [8]. This literature review has helped to interpret the important theories that provide the groundwork of analysis for medical data sets research.

3 Methodology

A medical diagnosis comes under a classification process to predict the disease. Before diagnosing the diabetes disease a physician has to examine a number of features which makes physician's job difficult. In present era, data mining methods have been considered to develop automatic prediction system for diabetes disease.

The focus on the methodology used in the current research work. Various phases of the proposed work are

i. Data Collection.

ii. Data Preprocessing.

iii. Feature selection.

iv. Input the data set with reduced set of attributes to the classifiers.

v. Split the data set into training and testing data sets.

vi. Apply the selected classification method on the training data.

vii. Develop a predictive model.

viii. Evaluate the performance of developed model by using test data set.

ix. Repeat the steps vi - viii with other classification methods.

x. Compare the performances among these data mining classification methods to identify the effective method.

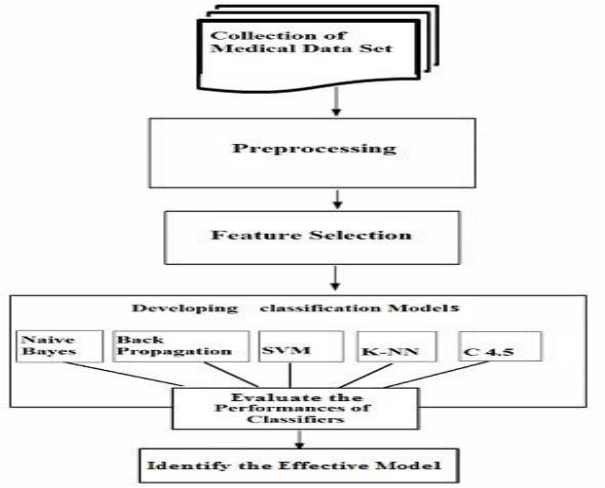

Fig.1: Steps in Identification of Effective Model for Disease Prediction

The research presents the validity and reliability of the classification methods by using confusion matrix, 10-fold cross validation technique and various performance measures like accuracy, sensitivity, specificity, precision, mean absolute error, AUC, time of computing. The main aim of this paper present a comprehensive comparison of five methods, namely, Naïve Bayes, Back Propagation, SVM, K-NN, and C4.5, and apply these classification methods to distinguish patients with either diabetes or non-diabetes. The comparative study of the results obtained from each supervised learning algorithms using Pima Indian Diabetes data set taken from the UCI machine learning repository will follow.

The Pima Indian diabetes database at the UCI machine learning repository had become a standard for testing the data mining algorithms to observe their prediction accuracy in diabetes data classification [10].

The diabetes data set consists different patients data, the data set contains 768 patient records with 9 attributes including the class was used to build the predictive model. All attributes are numerical as shown in Table 1. Class variable has two values tested positive for diabetes (contains the number of instances 268) and tested_negetive (contains the number of instances 500).

Table 1: Attributes of Pima Indians Diabetes Data set

S. No	Attribute Name	Type
1	Number of times pregnant	Numeric
2	Triceps skin fold thickness (mm)	Numeric
3	Plasma glucose concentration a 2 hours in an oral glucose TT	Numeric
4	Diastolic blood pressure (mm Hg)	Numeric
5	2-Hour Serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg / (height in m)^2)	Numeric
7	Age (years)	Numeric
8	Diabetes pedigree function	Numeric

4. Results & Discussions

This data set contains no missing values. So we have considered all the instances of diabetes data set for classification. The Statistical data of each feature in the diabetes data set before normalization is shown in Table 2.

Table 2: Statistical data of each feature in PIDD

S. No	Attribute Name	Min	Max	Mean	Std. Dev
1	Number of times pregnant	0	16	4.845	4.37
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0	189	111.895	33.973
3	Diastolic blood pressure	0	112	79.105	18.356
4	Triceps skin fold thickness	0	89	30.536	16.952
5	2-Hour Serum insulin	0	946	69.799	125.244
6	Body mass index	0	67.1	32.993	8.884
7	Diabetes pedigree function Numeric	0.078	3.42	0.472	0.431
8	Age	21	91	32.241	12.76

Feature selection is a very important step in classification process. A medical data set contains several attributes because, diagnosis of most of the diseases require many tests. So, disease prediction is an expensive process as many tests are required to predict the disease. A medical specialist collects data from several sources and tests to confirm a diagnostic impression but all these tests were not necessary or useful for the prediction of the heart disease. By applying the data mining methods we can cut the cost of diagnosis, by avoiding many tests and by selection of those features which are really required for prediction of disease. In this study for selecting features information gain measure was used. Feature selection was carried out in the diabetes data set to determine the main important features. The information gain values of the features present in the diabetes data set were shown in Table 3.

Table3: Information Gain values of attributes in PIDD

S. No	Attribute Name	Info Gain
1	Number of times pregnant	0.049
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0.189
3	Diastolic blood pressure	0.015
4	Triceps skin fold thickness	0.045
5	2-Hour Serum insulin	0.070
6	Body mass index	0.085
7	Diabetes pedigree function Numeric	0.022
8	Age	0.074

In this research work we applied 10-fold cross validation technique with each of the selected classification algorithms. In the 10-fold cross validation, the original data set is partitioned into 10 mutually exclusive subsets or folds, each of approximately equal size. So, diabetes disease data set is partitioned into ten approximately equal size subsets. In these ten parts nine parts are taken as training data set and the remaining one part can be taken as test data set. In each iteration, 9 parts of diabetes disease data set are taken as training data and the left over set is taken as test data. In rotation, every subset of the diabetes disease data set treated as the test data set in exactly one iteration. So, the overall classifier accuracy is the average accuracy of ten iterations. Repeating these ten folds can make sure that every part of the diabetes disease data set was used as training data set and also as test data set thus minimizing the sample bias.

Clearly we can observe that the SVM classifier shows more Accuracy, Sensitivity, Precision and Low Specificity, Mean Absolute Error for almost all set of features. So, the SVM algorithm is identified as the best algorithm for giving better accuracy with all combinations of feature sets of Pima Indians Diabetes Data set.

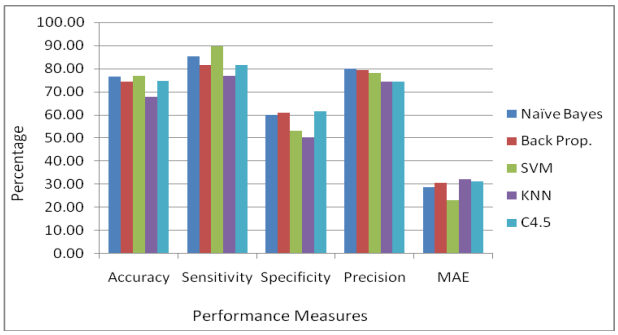

Fig.2: Comparison of performance of classification algorithms on PIDD.

The SVM classifier performances for different feature set combination of PIDD were shown in the Table 4

Table.4. Comparison performance of various algorithms with quantifiable measures

No. of Attributes	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
8-attr.	77.47	90.00	54.10	77.10	22.66	0.72	0.41
7-attr.	76.95	89.80	53.00	76.50	23.05	0.71	0.11

5 Conclusion & Future Scope

There are various data mining classification methods that can be used to predict the diabetes disease in the patients. In this thesis the five classification techniques were used to predict the diabetes disease. They are Naive Bayes, Back Propagation, SVM, K-NN and C4.5. These techniques were compared by using the measures accuracy, sensitivity, specificity, precision, mean absolute error, SVM outperformed Naïve Bayes, Back Propagation, KNN and C4.5 algorithms based on accuracy and other performance measures. Results were compared to find the most appropriate model for predicting the disease, and it is also identified that the efficiency of each classification method based on the size and quality of the data set. Future studies could explore whether accuracy could be improved by increasing or reducing the training data set, or by adjusting the parameters of the algorithms or by implementing a new predictive model with reduced historical data as it saves the effort required for the overall classification task.

References

1. Ali Adeli, & Mehdi Neshat. (2010), A Fuzzy Expert System for Heart Disease Diagnosis, Proceedings of International Multi Conference of Engineers and Computer Scientists, Hong Kong, Vol. 1, pp.134 -139.
2. B.M. Patil, R.C.Joshi, & Durga Toshniwal (2012), Hybrid prediction model for type II diabetic patients, Expert Systems with applications, science direct, 37(12), pp. 8102-8108.
3. Batista, G.E.A.P.A., & M.C. Monard (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence: An International Journal, 17(5), pp. 519-533.
4. Biswadip Ghosh. (2012), Using Fuzzy Classification for Chronic disease, Indian Journal of Economics & Business, 11(1), pp. 231-240.
5. Cortes, C., & Vapnik, V. (1995), Support-vector networks, Machine Learning, 20 (3), pp. 273-297.
6. Hand, D., H. Mannila & P. Smyth, (2001), Principles of Data Mining, MIT.
7. Fayyad Usama, Gregory Piatetsky Shapiro, & Padhraic Smyth (1996), From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence Press, pp.37-54.
8. Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufman.
9. Harsh Vazirani, Rahul Kala, Anupam Shukla, & Ritu Tiwari. (2010), Use of Modular Neural Network for Heart Disease, International Journal of Computer and Communication Technology, 1(2), pp. 88-93.
10. Koh, H.C., & G. Tan. (2005), Data Mining Application in Healthcare, Journal of Healthcare Information Management, 19(2), pp.64-72.
11. S.V.Achuta Rao etc.al.,A Predictive Model for Classification of Brest Cancer Data Sets", Springer Nature Publications, Advances in Computing Data Sciences, 5th International Conference, ICACDS 2021, Nashik, India, April 23–24, 2021
12. Mitchell, T.M., (1999). Machine learning and data mining. Communications of the ACM, 42(11), pp.31-36.
13. Parthiban, G., A. Rajesh, & S.K.Srivatsa (2011), Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method, International Journal of Computer Applications, 24(3), pp.7-11.
14. Sattler, K., & Dunemann, O. (2001). SQL database primitives for decision tree classifiers. Proceedings of the tenth international Conference on information and knowledge management.
15. Atlanta, Georgia. pp. 379-386. Yildirim, E.G., A.Karachoca, & T. Ucar. (2011), Dosage Planning for diabetes patients using data mining methods, Procedia computer science, science direct, Vol.3, pp. 1374-1380.
16. Zhou, Q.Q., M. Purvis, & N. Kasabov, (2007), Membership Function Selection Method for Fuzzy Neural Networks, University of Otago, Dunedin.

17. S. Nimmala, B. Vikranth, S. Muqthadar Ali, R. Usha Rani, and B. Rambabu, "An intelli MVC: An intelligent majority voting and heuristic-based hybrid classifier to predict high blood pressure," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 3, pp. 3003–3020, 2022.

18. N. Satyanarayana, Y. Ramadevi, K. Manasa, and B. Rambabu, "An intelli AFM: An Intelligent Association based fuzzy rule miner to predict high blood pressure using bio-psychological factors," *Intelligent Decision Technologies*, vol. 14, no. 2, pp. 227–237, 2020.