ISSN 2395-1621

Comparison of Classifiers for the Risk of Diabetes Prediction

[#1]Varsha Satpute, [#2]Dr. Swati Bhavsar

[1]varshasatpute22@gmail.com
[2]swati_bhavsar4@rediffmail.com

[#1]PG Student, Dept. of Computer Engineering
[#2]Professor, Dept. of Computer Engineering

Matoshri College of Engineering, Nashik.

ABSTRACT

The system allows the user to make use of algorithms to predict the risk of diabetes mellitus in the human body. The various classification models such as Decision Tree, Artificial Neural Networks, Logistic Regression, Association rules, and Naive Bayes are used in this system. Then the Random Forest technique is used to find the accuracy of each model in the project. The dataset used is the Pima Indians Diabetes Data Set, which has the information of patients, some of them have developing diabetes, therefore, this project is aimed to create a mobile application for predicting a person's class whether present in of the diabetes risk or not.

Index Terms—decision tree, artificial neural network, logistic regression, naive bayes, random forest.

ARTICLE INFO

Article History

Received: 1st May 2019

Received in revised form :

1st May 2019

Accepted: 2nd May 2019

Published online :

05th May 2019

I. INTRODUCTION

Diabetes Mellitus is one of the most common and growing diseases which occurs due to the high glucose level in blood of human being. It is growing in many countries all over the world and it is necessary to prevent this disease at an early stage by identifying the symptoms of diabetes using several methods and by taking precautions to not to happen. This paper focused on various classification algorithms and the purpose of this is to compare the performance of every algorithm using the data mining techniques and to predict the accurate result. The main focus of this study is to give the prediction of the risk of diabetes for everyone without going to a hospital and without testing the blood samples. This project has the aim to encourage and promote good health of human being in society. This diabetes prediction application will be created as a simple application for diagnosis and it will give results that whether a human tends to have diabetes in the future. However, this application is used only for initial diagnosis. This can be used by people who find themselves in the diabetes risk group, these people should go to see a doctor for a formal diagnosis and should take proper treatment on the disease by an expert doctor.

II. REVIEW OF LITERATURE

Abundant literature has been dedicated to the classifiers of data mining and tremendous progress has been made ranging from efficient and scalable algorithm for different datasets. This section provides a brief overview of the current status of diabetes prediction and discusses a few promising research directions. We believe that mining research has substantially Broadened the scope of data analysis and will have deep impact on mining methodologies and applications in the future. However, there are still some challenging research issues that need to be solved in searching using concept of various classifiers of data mining in the research of diabetes prediction.

A. K-Means Algorithm And Logistic Regression

They improved the accuracy of the each prediction model, and made the model more acceptable and adaptive to more than one available dataset. This method involves a series of pre-processing procedural tasks, this model has two main parts, first is the improved K-means algorithm and second is the logistic regression algorithm. It uses the Pima Indians Diabetes Dataset. The Waikato Environment is used as Knowledge Analysis toolkit to utilize and to compare the results.

B. Electromagnetism-Like Mechanism

The use of data mining techniques related to artificial intelligence for various medical database classifications, prediction and diagnosis has increasing popularity. They

presented a novel method for feature selection by making use of op-posite sign test (OST) which works as a local search for the electromagnetism-like mechanism algorithm, which can work as improved electromagnetism- like mechanism (IEM) algorithm. The wrapper method is used in this algorithm for nearest neighbor algorithm classifier. The proposed IEM algorithm is compared with algorithms of feature selection and classification. In this study, 54 UCI datasets are used to evaluate the performance of various classification algorithms. These datasets are characterized according to data sizes, features, and classes.

C. Temporal Abstraction with Data Mining

The Hemodialysis patients may affected by unhealthy care behaviors and they are needed to be hospitalized for many days6. If the rate of hospitalization of a hemodialysis center is very high, and it has low service quality. Therefore, decreasing of rate of hospitalization is a important problem for health care centers6. This study comprises temporal abstraction along with data mining techniques for analyzing patients having dialysis the related biochemical data needed to develop a decision support system 11. The mined temporal patterns are helpful for clinicians to predict hospitalization of hemodialysis patients and to suggest immediate treatments to avoid hospitalization.

D. Pre-Diabetes Detection By Risk Factors

The characteristics of each participant and Pearson Chi-square test is performed between two groups. The sensitivity analysis is performed, along with logistic regression model, ANN model, C5.0 and decision tree model. This study is aimed to determine the order detailed predictions produced from the training dataset and testing datasets which are presented in the form of confusion matrices.

III. PROPOSED SYSTEM

The proposed system is comprised of four stages in the overall framework in the study. The very first step of the framework is data manipulation. Next, there are four models will be examined for determining a prediction model. Then, the accuracy of each and every model will be calculated and compared with each other for getting the best model from them so that an accurate result will be predicted to the user. Lastly, it provides the result to the user whether he will be prone to have diabetes in the future. The study ends up creating a mobile application.

Healthcare information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer-based analysis are very needed. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytical tools. It has been proven that the benefits of introducing data mining into the medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources.

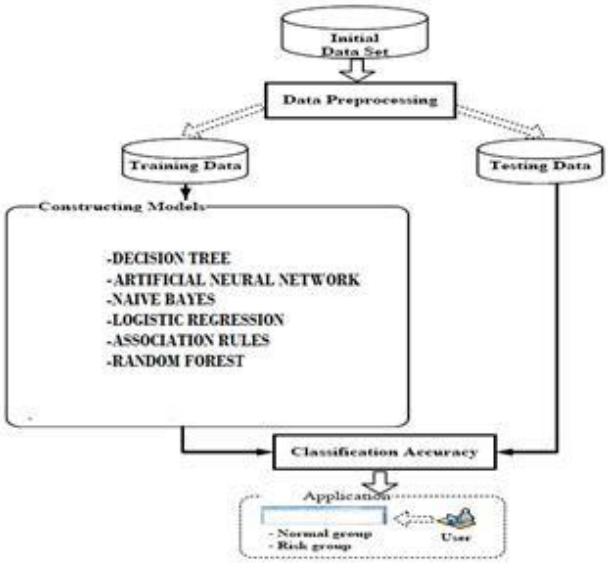

Fig.1. System Architecture

B. Data Pre-processing

An initial data set was collected from PIMA INDIAN Dataset. Each person has to fill a form that will be used to get input data from user. In this framework, the data set is used as follows; Dataset is collected from each PCU and merged together before giving to application, then some variables such as BMI and age needs a transformation, are converted using other design parameters . for example if These variables are correlated and depending on each other by giving factors. Hence, the number of input variables are eleven and number of output variable is one i.e. class output variable as shown in Table 1. The table has all variables used in the prediction model, their description and values of each variable which are allowed. Variables from one to eleven are input variables where from one to six variables are numeric values and the remaining are categorical variable and the last variable is the dichotomous, called as a class variable which defines output of application. The name of the output variable is the fasting blood sugar (FBS) is divided into two main groups, one is normal group and another is risk group. The normal group is a people who are having variable FBS which is less than 100 mg/dl whereas a people who are having FBS between 100-125 mg/dl will be categorized into the risk group. In this project, people who are having FBS more than 125 mg/dl are not included because they are categorized into a diabetes group so these people will be treated as patients so that they should visit to doctor for further treatment.

IV. ALGORITHM

1. Decision Tree

Decision tree induction is the simple learning method of decision trees from class training tuples. A decision tree has a flowchart-like tree-like structure, where it has internal nodes also called as non-leaf nodes which denotes a given

test on attribute, decision tree has branch form root to leaf node which represents an output of the test, and it also has leaf node also called as terminal node which holds a class label. The top node in a tree is called as the root node. While constructing decision tree classifiers it does not require any technical knowledge or setting of parameters, and therefore it is appropriate for discovery of knowledge. Definition of the algorithm starts with specifying a root node from the given most relationship between each input and output variables. Generate decision tree

1. Check if algorithm satisfies termination criteria.

2. Computer information-theoretic criteria for all attributes.

3. Choose best attribute according to the information-theoretic criteria for construction of tree.

4. Create a first node i.e decision node based on the best attribute in step 3.

5. Split the dataset based on newly created decision node in step 4 as per decision node.

6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call).

7. Attach the tree obtained in step 6 to the decision node in step 4 for tree construction.

8. Return tree.

2. Artificial Neural Network

ANN is computational system which is inspired by the method of processing of structure and ability of learning of a biological human brain. The Characteristics of Artificial Neural Network includes a large number of very simple processing neuron-like processing elements in computational system. There is a large number of weighted connections between the elements. Distributed representations of knowledge over the connections are given.

$$f(x) = \frac{1}{(1+e^{x})}$$

$$Where\ x=(input\_val1\ X\ wt1)+(input\_val2\ X\ wt2) + \ldots\ldots\ldots+(input\_val3\ X\ wt\ n)$$

This algorithm is used to evaluate a function which has of large number of set of input variables. It can be used not only for numeric output variables but also for categorical variables in the system. First of all a calculation of number of hidden layers is done by adding number of input and number of output variables which is then divided by two. For example, if there are twenty-seven input variables and only one dichotomous class output variable. Then there will be the number of hidden layers is equal to fourteen. Then, values of these hidden layers are given by sigmoid function such as.

3. Naive Bayes Theorem

Naive Bayes algorithm represents a various methods for classification such as supervised learning method and statistical method for classification. It considers probabilistic model in classification and it allows user to calculate the uncertainty of the model which can be done by calculating probabilities of the outputs. It can solve various problems of diagnostic predictive and predictive sytems. Let class i is the group of people lying in diabetes risk group i and V is the set of input variables that are used in a system where is it the assumed that each variable is independent. To predict a class of a person whether lies diabetes risk, an approach of Naive Bayes can be defined by

Where P($class_i$ V) is defined as a probability of a

$$P(Class_i \mid V) = \frac{P(V \mid Class_i) * P(Class_i)}{P(V)}$$

training data set with variable V that will be the output class $class_i$. P(V $class_i$) is a probability of a training data set of $class_i$ and variable V where V = V1∩ V2

$$P(\widetilde{Class_i} \mid \hat{V}) = \frac{P(V_1 \mid Class_i) * P(V_2 \mid Class_i) * \ldots\ldots * P(V_M \mid Class_i) * P(Class_i)}{P(V)}$$

VM. P($class_i$) is a probability of person who is the part of diabetes risk group i. Thus, the prediction of a person will be $class_i$ when it predicts the highest value of P($class_i$ |V) among all classes for the given prediction model.

4. Logistic Regression

Logistic regression analysis studies the association between a set of categorical dependent variable and a set of independent variables. Let z is the variable which defines a probability of occurrence of an event and 1-z is a probability of nonoccurrence of an event. Hence, the logistic regression model can be given by the formula where 0 is the intercept and β0, β1, ..., βp are the regression coefficients.

$$Z = \frac{e(\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta pxp)}{1+e(\beta 0+\beta 1x1+\beta 2x2+\ldots+\beta pxp}$$

5. Random Forest Algorithm

Random forests is an ensemble learning algorithm. The basic concept of the algorithm is that building a number of small decision-tree having less features and which is a computationally very cheap process. After building many small and weak decision trees parallelly, then these trees can be combined to form a single and strong tree by averaging or taking the majority vote. The algorithm is started by merging a combination of trees which each tree will vote and emerged for a class1. Suppose that there are N number of data and M number of input variables in a data set where the real data used in this system which comprised of data and input variables. Let k is the number of sampling groups in the system, ai and bi be number of data and number of variables in group i where i is given

as 1, 2, 3, 4 ... and k. Each sampling group is defined as follows

1. ai is the data where ai is not greater than N variables which are selected randomly from N input variable.
2. bi are the variables where bi is not greater M data which are selected randomly from M input variable.
3. A tree increases to leaf nodes and predicts class.

After repetition of steps from 1 to 3 for the given k times, this tree will take a structure like a forest. Then the classification of many models will be selected by considering a majority vote of all trees in the forest.

E.g

$$Age>55 \wedge BPH= yes \rightarrow class= diabetes\_risk\_group \quad [support=20\%, confidence=93\%]$$

6. Association Rules

Association rules gives a strong relationship between attribute- value pairs (or items) that occur frequently in a given data set. Association rules are commonly used to analyse the patterns of two or more frequent things that will happen together and two things which are frequently used together.

E.g. purchase pattern of customer in store.

Such analysis is useful in many decision-making processes, such as product placement, catalogue design, and cross- marketing. Discovery of association rules are based on associative classification where association rules are generated and analyzed for classification and prediction of diabetes.

Association rule mining is two-step process. First

step is that it searches for attribute-value pair that occurs repeatedly in data-set. Each pair in dataset is called as item. Group of these items is called as frequent item-set. Next step is to analyse frequent item sets to generate association rules.

V. SYSTEM REQUIREMENTS

A. Software Requirement

1) Operating System: Windows7 AND ABOVE
2) Application Server: Tomcat5.0/6.X, Glassfish
3) Front End: HTML, Java, Jsp
4) Scripts: JavaScript
5) Server side Script : Java Server Pages.
6) Database: My sql 5.5
7) Database Connectivity: JDBC.

B. Hardware Requirement

1) Processor: Intel
2) CPU Speed: 1.1 GHz or Higher
3) RAM: 2 GB or Higher
4) Hard Disk: 100 GB or Higher

VI. EXPERIMENTAL ANALYSIS

In this section, the classification accuracy of thirteen models in the previous section is presented to assess the performance of each model as shown in Table 1.

Models	**Accuracy(%)**
Decision Tree(DT)	85.09
Artificial Neural Network(ANN)	84.53
Assosition Rule(AR)	82.30
Naïve Bayes(NB)	81.01
Random Forest(RF)	**85.558**

Table 1.:Comparision Result

Table 1 displays the results of comparison of the classification accuracy of thirteen models. The top five accuracy are Random Forest, Decision Tree, Artificial Neural Network, Decision Tree and Boosting with Decision Tree models which are 85.09%, 84.53%, 82.30%, 81.01% and 85.55% respectively. It can be seen that most of them are based on Decision Tree algorithms. Hence, Decision Tree model works well with this data set. The least accuracy is from the model of Bagging with Naïve Bayes, 80.960%. The results also suggest that Bagging and Boosting techniques improve the accuracy of Decision Tree, Artificial Neural Network. The accuracy of Naïve Bayes model is only improved by Boosting technique. On the other hand, the accuracy of Bagging with Naïve Bayes model, 80.960%, is less than the accuracy of Naïve Bayes only, 81.010%, but not by much. However, these accuracies are not greater than the Random Forest accuracy. Note that Random Forest was developed from the combination of Trees and Bagging algorithms therefore the accuracy of Bagging with Decision Tree model, 85.333%, is closely to the accuracy of Random Forest model, 85.558%. In order to be confirmed the accuracy of prediction, the comparision result as shown in below Fig. 2.

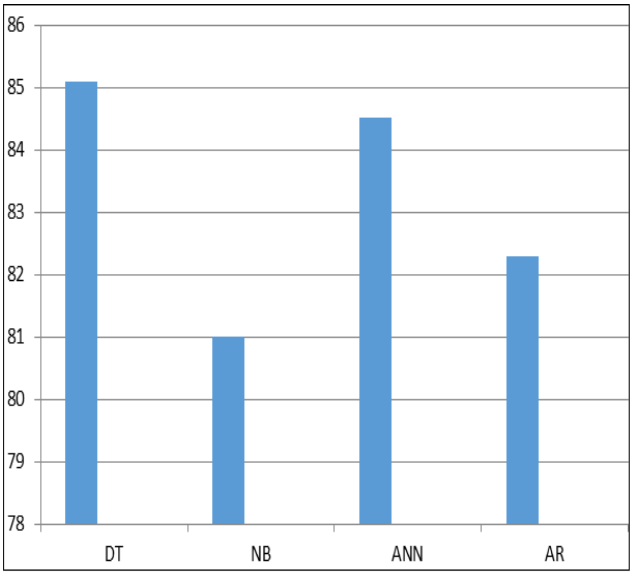

Fig. 2: Comparison Of 5 models.

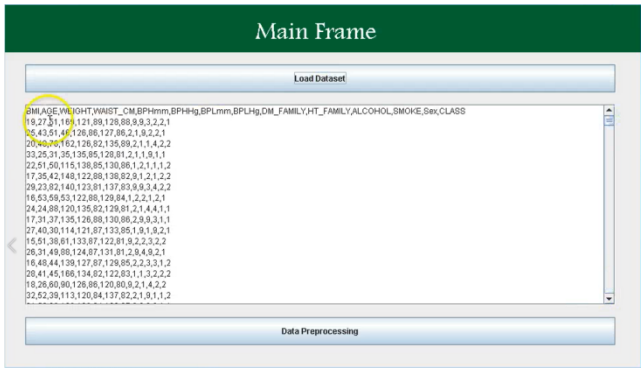

Fig. 3: Load Dataset

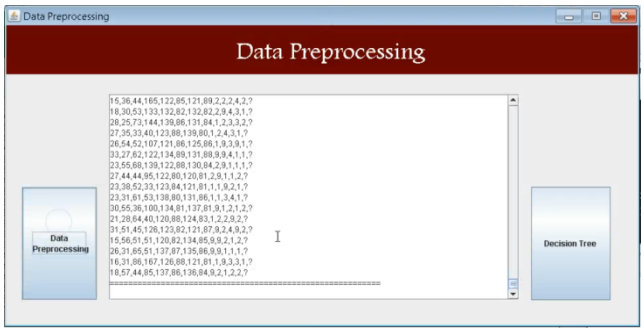

Fig. 4: Data Pre-processing

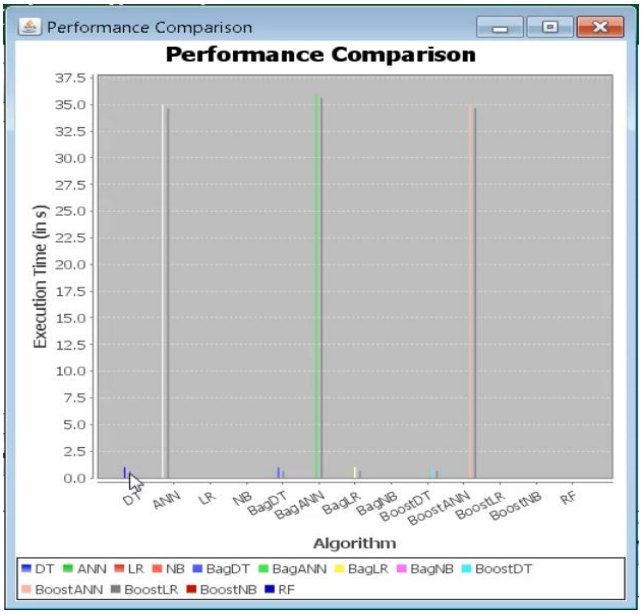

Fig. 5: Performance Comparison

VII. CONCLUSION

A versatile application is proposed by utilizing an utilization of sickness classifiers and a genuine informational collection. The information utilized in this creation are general data of individuals who were gathered from PIMA Indian Dataset. Before making the application, grouping models are assessed for developing expectations demonstrate. This model comprises of Decision Tree, Neural Network, Logistic Regression, Naive Bayes, Association Rules and Random Forest calculations. So as to give precise outcomes to client, exactness of every single model is determined. The ROC Curve is attracted to affirm to precision of each model. The model which is having most elevated exactness esteem will be exhibited to client thus. In procedure of Random Forest comprises of choosing the information in arbitrary path as well as choosing input factors in irregular way. That is the reason the application expands the precision of forecast. Hence this task has plan to create and looks at different order models for diabetes hazard forecast.

REFERENCES

1. Stefano Bromuri , Serban Puricel, Rene Schumann , Johannes rampf,Juan Ruiz and Michael Schumacher, -An expert Personal Health System to monitor patients affected by Gestational Diabetes Mellitus: A feasibility study,— Journal of Ambient Intelligence and Smart Environ-ments 8 (2016) 219237G [1].

2. Gyorgy J. Simon,Member, IEEE,Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha,M. Regina Castro and Peter W. Li ,-Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus,—IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEER-ING [2].

3. Han Wu, Shengqi Yang , Zhangqin Huang, Jian He, Xiaoyi Wang,- Type 2 diabetes mellitus prediction model based on data mining,— Informatics in Medicine Unlocked 10 (2018) 100107 [3].

4. Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, Qing Liu b, - Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,— Kaohsiung Journal of Medical Sciences (2013)29, 93e99 [4].

5. Kung-Jeng Wang, Angelia Melani Adrian a, Kun-Huang Chen a, Kung-Min Wang b,- An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus,— Journal of Biomedical Informatics 54 (2015) 220229 [5].

6. Jinn-Yi Yeh, Tai-Hsi Wu b, Chuan-Wei Tsao, -Using data mining tech-niques to predict hospitalization of hemodialysis patients,— Decision Support Systems 50 (2011) 439448 [6].

7. Nongyao Nai-arun, Rungruttikarn Moungmai , - Comparison of Classifiers for the Risk of Diabetes Prediction,— 7th International Conference on Advances in Information Technology [7].

8. Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath, - Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients,— IJEAT Volume-1, Issue-3, February 2012 [8].

9. Darshan K R, Anandakumar K R, A Comprehensive Review on Us-age of Internet of Things (IoT) in Healthcare System — 2015 IEEE ICERECS&T [9].

10. Chih-Hua Tai, Daw-Tung Lin, A Framework for Healthcare Every-where: BMI Prediction using Kinect and

Data Mining Techniques on Mobiles, 2015 16th IEEE International Conference on Mobile Data Management[10].

11. Ritika Chadha , Shubhankar Mayank, Prediction of heart disease using data mining techniques, 2016 OF CSIT Springer [11].

12. Eleni I. Georga, Vasilios C. Protopappas, Stavroula G. Mougiakakou,, Short-term vs. Long-term Analysis of Diabetes Data: Application of Machine Learning and Data Mining Techniques, IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 2003 [12].

