Classification and Prediction of Diabetes Mellitus using Data Mining Techniques

M. Marimuthu, S. Deiva Rani, C. Mythili, L.Swathi
Coimbatore Institute of Technology, Coimbatore.

Abstract— *Data mining is the search of vast datasets to extract covered up and previously unknown patterns. Classification is the one of the errand in data mining. Health care data are frequently tremendous, Complex and heterogeneous because it contains distinctive variable types. These days, learning from such data is a need. Data mining can be used to separate learning by building models from Health care data such as diabetic patient datasets. Diabetes mellitus is a chronic illness and a major public health challenge around the world. Utilizing data mining strategies to help individuals to predict diabetes has gain major popularity. In this paper, we predict whether the person have diabetic or not. In this paper we use the Classification technique C tree interface to classify diabetes data.*

Keywords— **Classification, C-tree, Data Mining, Diabetes, Logistic Regression, Prediction.**

I INTRODUCTION

Data mining is the search of patterns among existing databases these patterns are hidden among these large data, for example, a connection between patient data and their medical diagnosis. Classification is a data mining technique that separate data based on classes. The objective of classification is to precisely predict the objective class for each case in the data. The issue of classification plays a vital part in analyzing any medical diagnosis. Medical diagnosis is a problem complicated by many factors and concerning all of human abilities including intuition and the subconscious. Diabetes mellitus which is often simply referred to as diabetes is a disorder that is caused by decreased production of insulin or by decreased ability to use insulin, for this reason glucose levels in the blood increases. Diabetes increases the risks of developing heart disease, kidney disease, blindness, nerve damage and blood vessel damage.

According to a survey, Indonesia became the fourth country in the world that has the highest diabetes rate and increased up to 14 million people. It is based on the report of World Health Organization (WHO), in which the number of diabetics in Indonesia in 2000 was 8.4 million people followed by India (31.7 million), China (20.8 million) and the United States (17.7 million). WHO reported that there are more than 143 million people who suffered diabetes. This number projected the prevalence that will double in the 2030 and as much as 77% of which occur in developing countries. Techniques such as Binary logistic regression, support vector machine algorithm and C tree interface are used.

II LIETURTURE SURVEY

This section showed that there have been several studies on the prediction problem using statistical approaches and many data mining techniques. However, a few studies related to medical diagnosis using data mining approaches have been reported.

In [1] authors have constructed an artificial neural network model for diagnosis of diabetes, they used certain combination of preprocessing techniques to handle the missing values and compared the results of accuracy of the model for each technique, however the method of handling missing values presented in this paper wasn't employed in that study.

In [6] the SVM implementation gives the prediction accuracy of 94%. Another implementation of the SVM in detecting the diabetes is given in [7]. Here, the SVM classifier, however, performs only 78 % of accuracy. A method for prediction of diabetes by using Bayesian network is given in [8] while the authors in [9] separately use Naïve Bayes and k-nearest neighbor algorithm.

Hai Wang et. al. [10] performed a study in medical knowledge acquisition using Data mining. It has been widely considered as an effective tool for knowledge discovery. This paper discuss about the critical part of medical specialists for medical data mining, and shows a model of medicinal learning procurement through data mining. According to American Diabetes Association (ADA) in 2010, diabetes mellitus is a group of metabolic diseases with hyperglycemia characteristic that occurred because of abnormalities of insulin secretion, insulin action, or both [11]. The term data mining refers therefore to the overall process consisting of data gathering and analysis, development of inductive learning models and adoption of practical decisions and consequent actions based on the knowledge [12].

III PROPOSED METHOD

A. Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. Logistic regression as a special form of linear regression when the result variable is categorical, where we are utilizing log of odds as dependent variable. In simple words, it predicts the

probability of event of an occasion by fitting data to a logistic function.

Binary Output Variable: This may be clear as we have just specified it, yet logistic regression is planned for binary (two-class) classification issues. It will predict the probability of an instance having a place with the default class, which can be snapped into a 0 or 1 classification.

Remove Noise: Logistic regression expect no error in the output variable, consider evacuating outliers and possibly misclassified examples from your training data.

Gaussian Distribution: Logistic regression is a linear algorithm. It assumes a linear relationship between the input variable with the yield. Data changes of your input factors that better uncover this linear relationship can bring about a more exact model. For instance, you can utilize log, root, Box-Cox and other univariate changes to better uncover this relationship.

Remove Correlated Inputs: Like linear regression, the model can over fit if you have multiple highly-correlated inputs. Consider calculating the pair-wise correlations between all inputs and removing highly correlated inputs.

Fail to Converge: It is feasible for the normal probability estimation process that takes in the coefficients to fail to converge. This can happen if there are various highly related inputs to your data.

B. Support Vector Machine Algorithm

This is a machine learning algorithm used to analyze data for classification and regression analysis. SVM comes under supervised learning in machine learning that takes a gander at data and sorts it into one of two sections. A SVM yields a guide of the arranged information with the edges between the two as far separated as possible. SVMs are utilized as a part of content categorization, picture classification, handwriting recognition and in the sciences. A support vector machine is otherwise called a support vector organize (SVN).

Maximal-Margin Classifier: The Maximal-Margin Classifier is a classifier that verifies the work of SVM function for the numeric input variables in your data. A hyper plane is a line that parts the input variable space. In SVM, a hyper plane is chosen to best separate the focuses in the input variable space by their class, either class 0 or class 1. In two-measurements you can visualize this as a line and how about we accept that the greater part of our input points can be totally isolated by this line. For instance:

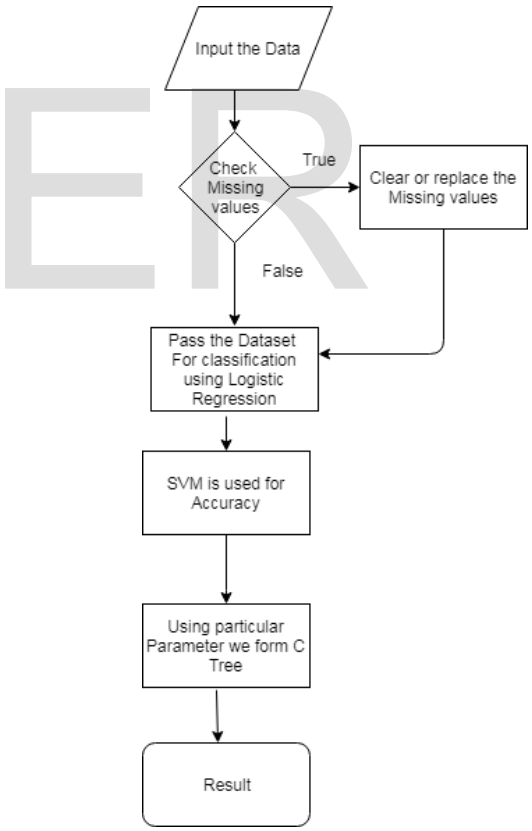
$$B0 + (B1 * X1) + (B2 * X2) = 0 \qquad (1)$$

In equation (1) Where the coefficients (B1 and B2) that determine the slope of the line and the intercept (B0) are found by the learning algorithm, and X1 and X2 are the two input variables.

C. C-tree

C-tree is one of the classification algorithm. For this the needed package is Party, arules, aruleviz we need to bind this three packages and then start to classify the dataset. In this we split the dataset into two parts of data that is train and test data. Give separate index id for the both train and test data. Based on one major variable we can add or subtract (i.e., any mathematical action) can be done for the other variable. Using C-tree function with the details of train data and test data print the graphical representation of the C-tree. By using the function predict print the tabular view of the predicted data. Using plot function we can plot the graphical representation of the C-tree. Before all this process we have to set the node size of the tree by using the function seed.

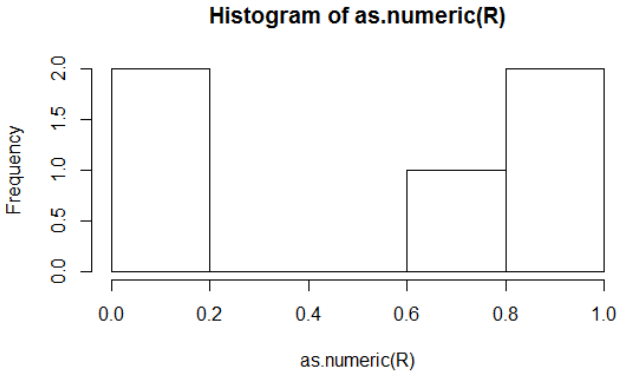
D. Flow Chart



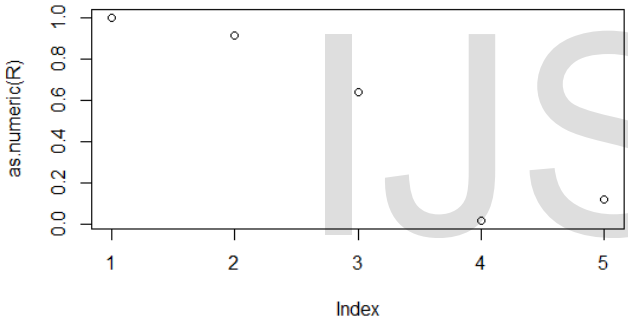
IV RESULTS AND DISCUSSIONS

A. Logistic Regression

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	*0.9172*	*0.6398*	*0.0203*	*0.1223*



Histogram format for the Logistic Regression



Graphical representation of Logistic regression

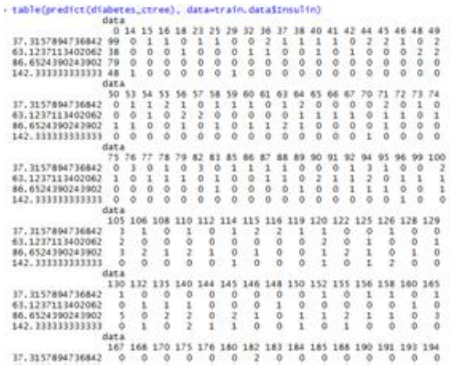
B. Support Vector Machine

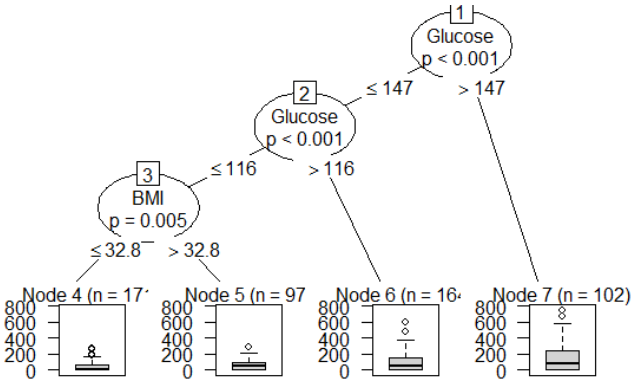
sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
1	0.04333523		0.25	0.8790943	0.3299309
0.01182008	0.06978217				
2	0.04333523		0.50	0.9056552	0.5146615
0.03064493	0.19972432				
3	0.04333523		1.00	0.9143509	0.5970924
0.03212887	0.18607146				

C. C Tree

In this Method we have drawn C Tree for the formula

c-tree <- insulin ~ Bloodpressure + BMI+ Glucose





Diagramatic Representation of C Tree

V CONCLUSION

This work focused the implementation of Logistic Regression, Support Vector Machine Algorithm and C Tree for the diabetes data. From the analysis, it is examined that the formation of classifications will be different for classification methods. From the histogram, it is seen that the Logistic Regression accuracy is 0.91, Support Vector Machine is 0.92.

REFERENCES

[1] Al Jarullah, A.A, "Decision Tree Discovery for the Diagnosis of Diabetes", Innovations in Information Technology (IIT), International Conference.

[2] Aini Hanifa, Mira Kania Sabariah and Siti Sa'adah, "Early Detection of Diabetes Mellitus Random Forest, Classification and Regression Tree", IEEE Transaction.

[3] Swasti Singhal, Monika Jena, "A Study on Weka Tool for Data Preprocessing, Classification and Clustering", IJITEE.

[4] Mrs.S.S.Sherekar, Tina R.Patil, "Performance Analysis Of Naïve Bayes and J Classification Algorithm for Data Classification", Journal of Computer Science and Application.

[5] Vijayarani.S, Muthulakshmi.M., evaluating the efficiency of rule techniques for file classification, International Journal of Research in Engineering and Technology.

[6] R. Aishwarya, P. Gayathri, and N. Jasinkar, "A Method for Classification Using Macnine Learning Technique for Diabetes", International Journal of Engineering and Technology.

[7] V. A. Kumari and R. Chitra, "Classification of Diabetes Using Support Vector Machine," International Journal of Engineering Research and Applications.

[8] A. Arora, M. Kumari, R. Vohra, "Prediction of Diabetes Using Bayesian Network", International Journal of Computer Science and Information Technologies.

[9] A. N. R. Nurhayati, "Implementation of Naïve Bayes and K-Nearest Neighbor Algorithm for Diagnosis of Diabetes Mellitus", Applied Computational Science.

[10] Hai Wang, "Medical Knowledge Acquisition through Data Mining", IEEE International Symposium.

[11] PERKENI, "Konsensus Pengendalian dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia", Jakarta.

[12] C. Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making", Chennai.