Predicting Diabetes Disease using Random Forest Tree (Rft) Data Mining Technique

Gul Mohamed Rasitha Banu, N. Sasikala, Illham Bashier, Thani Babikar

Abstract: ***Diabetes is a condition that happens when the blood glucose is too high, also known as blood sugar. The primary source of energy is blood sugar, and it comes from the food you eat. Insulin, a pancreatic hormone, helps food glucose get into the cells for energy use. It also leads for an unrelated condition named, "Diabetes Insipidus", which entails complications with the processing of fluids in the kidney. Insulin is the key to the ability of the cell to use glucose. Problems with the processing of insulin or how cells perceive insulin can easily cause out of control the body's carefully balanced glucose metabolism process [1]. Diabetes emerges when either of these conditions happens, blood sugar levels rise and crash and the risk of organ damage. Earlier prediction of this diabetes condition could provide proper treatment to protect the people from un avoided illness. For this prediction we can apply data mining which is used predominantly in healthcare organizations for decision making, disease detection purpose. In this paper data have been collected from UCI repositories and the data mining tool (WEKA) is used to predict diabetes. In this database there are 768 instances in which 500 instances belongs to tested negative and 268 instances belongs to tested positive. An experimental study is carried out using data mining technique classification technique called Random Forest Tree (RFT) classifier to predict diabetes. In this research, we have used different cross fold validation to achieve better accuracy and we found that cross fold validation k= 8 gives high accuracy 76.69% while compared with other cross fold validation values.***

Keywords: Diabetes, Diabetes Insipidus, Classification, Datamining.

I. INTRODUCTION

Human body is made up of tissues and cells that set up the human body as living things that require nourishment to remain alive. Cells can only take food in a sugar type called glucose.

The blood streams are the only way to carry glucose to them, fixed in place as they are. Without access to adequate glucose, the cells of the body have nothing to fuel and die soon. People eat only food, not glucose. As part of the normal digestion process, human foods are converted into glucose. When transformed, glucose reaches the blood stream, allowing the blood to increase the level of dissolved glucose. Then the blood flow takes the absorbed glucose to the body's different tissues and cells.

Revised Manuscript Received on December 5, 2019.

Dr. Gul Mohamed Rasitha Banu , Health Informatics, FPHTM, Jazan University, KSA.
Dr. N. Sasikala, Health Informatics, FPHTM, Jazan University, KSA.
Dr.Illham Bashier, Health Education, FPHTM, Jazan University, KSA.
Dr. Thani Babikar, Health Education, FPHTM, Jazan University, KSA.

Even though glucose may be accessible in the blood, neighbouring cells cannot access the glucose without the aid of a chemical hormone called insulin. In the presence of insulin, cells consume glucose from the blood, and blood sugar levels decrease as the blood leaves sugar and reaches the cells. Insulin can be viewed as a glucose bridge between the cells and the blood stream. The body is designed to regulate and buffer the amount of dissolved glucose in the blood to maintain a steady supply to meet the needs of the cell. The pancreas, one of the many organs in your body, is creating, storing and releasing insulin into the bloodstream to bring down blood sugar levels. Insulin is the key to the ability of the cell to use glucose. Problems with the development of insulin or how cells perceive insulin can easily cause out of control, and so the body's carefully balanced glucose metabolism process. Diabetes arises when either of these complications happens, blood sugar levels increase and crash and the risk of damage to the body. RFT Algorithm is used in this work to estimate Diabetes disease. For the experimental purpose a data set of 9 features downloaded from the UCI repository site is used, all work is carried out under Windows 10 environment with WEKA open source software. Validation of K-fold is also carried out.

II. DATA SET DESCRIPTION

The data set used in this paper is downloaded from the website with the given address(https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff) for experimental purposes. The data set has 768 instances, 500 belonging to the negative category,268 belonging to the compensated diabetes category. The details of data set is given below.

Table - I: Data set Description

Data Description SN	Attribute Name	Value Type
1	Number of times pregnant	Continuous
2	Plasma glucose concentration	Nominal
3	Diastolic blood pressure	Nominal
4	Triceps skin fold thickness	Nominal
5	serum insulin	Nominal
6	Body mass index	Numeric
7	pedigree function	Nominal
8	Age	Numeirc
9	Class variable	0 1

III. REVIEW OF STUDY

[Habtamu Wondifraw Baynest] The worldwide prevalence of diabetes has continued to increase dramatically. Globally, as of 2011, an estimated 366 million people had DM, with type 2 making up about 90% of the cases [9,10]. The number of people with type 2 DM is increasing in every country with 80% of people with DM living in low- and middle-income countries. [5]

[Quan Zou,1,2,] * Kaiyang Qu,]According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. [6]

[Tejas N. Joshi*, Prof. Pramila M. Chawan**] Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.[7]

[Md. Aminul Islam, Nusrat Jahan] With the default configuration, logistic regression had the highest accuracy (78.01%) and AUC (0.833) whereas IBK obtained lowest accuracy (70.99%) and simple classification algorithm OneR obtained the lowest AUC (0.642).[8]

IV. METHODOLOGY

A. Random Forest Tree Algorithm:

This Algorithm Description of the following evolved version can assist doctors to take right choice to present better treatment to the patients. The RFT Algorithm is defined beneath.3.1.2 Random Forest Tree Algorithm:

Data mining technology helps several category algorithms along with sensible regression, helping vector machine, naive classifier Bayes, and choice timber. But the random forest classifier is near the top of the classifier hierarchy. Random forests or random choice forests are an ensemble gaining knowledge of technique for classification, regression and different tasks that works through constructing a large number of selection trees on the time of training and generating the magnificence that is the magnificence mode (classification) or suggest prediction (regression) of the individual tree.[1]

B. K – Fold cross -Validation:

In K fold pass-validation, the original pattern is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained because the validation statistics for checking out the version, and the last k − 1 subsamples are used as training records [2]. The pass-validation method is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation facts. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that everyone observations are used for both training and validation, and each observation is used for validation exactly the records is loaded into WEKA software.[3] After pre-processing, various facts mining category techniques are applied at the statistics set to expand the predictive models. And the system is educated the usage of the training set. After the gadget is educated it's far tested the use of up to 10 fold pass validation method. Evaluation is carried out using certain overall performance measures.

V. EXPERIMENTS WITH WEKA

Weka is a collection of algorithms for data mining tasks in machine learning. Weka comprises tools for pre-processing, classification, regression, clustering, rules of association, selection and visualization of features. Weka can download from the website. [4] The pre-processing stage of Weka is shown in below fig.1

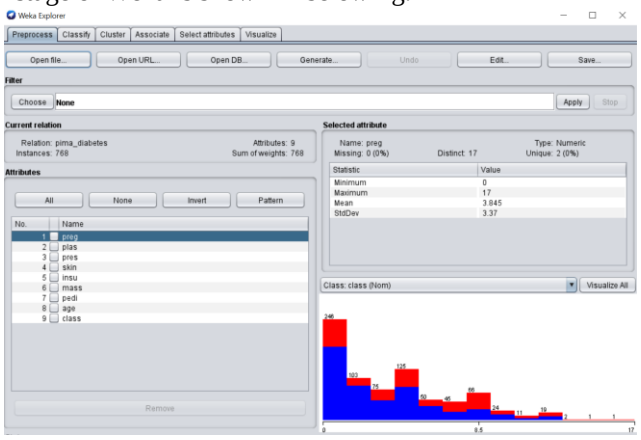

Fig:1 Pre-processing stage of Diabetes data set

VI. RESULT AND ANALYSIS

There are in total 768 records in the diabetic data set.All the records are classified as tested negative, and tested positive. In our experiment data is supplied to Random Forest Tree classifier algorithm. The following fig.2 shows the output of Random Forest classifier with cross fold k=10.

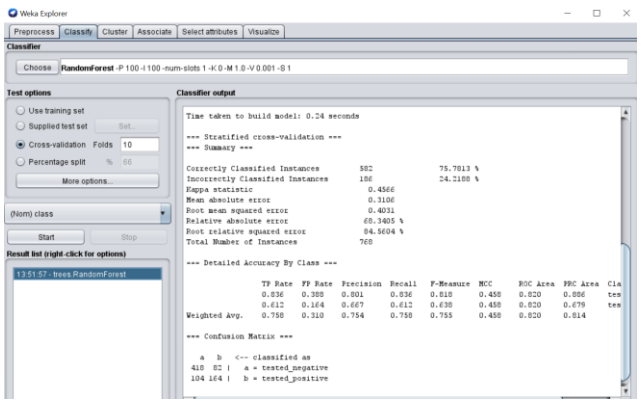

Fig:2 Output of RFT classifier with Cross Fold Validation k=10

The following Table 2 represents confusion matrix for Random Forest Tree Algorithm with k=10 fold.

Confusion Matrix:

Table-II: Confusion matrix for Random Forest Tree Algorithm for k=10 fold

Target Class	Tested Negative	Tested Positive
Tested Negative	418	82
Tested Positive	104	164

The following fig.3 shows the output of Random Forest classifier with cross fold k=8.

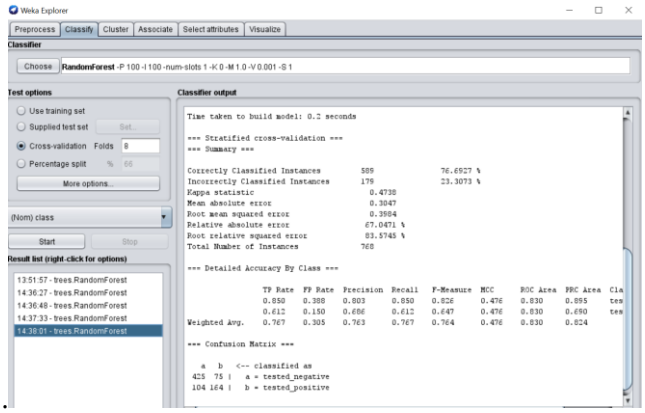

Fig:3 Output of RFT classifier with Cross Fold Validation k=8

The following Table 2 represents confusion matrix for Random Forest Tree Algorithm with k=8 fold.
Confusion Matrix:

Table -III: Confusion matrix for Random Forest Tree Algorithm for k=8 fold

Target Class	Tested Negative	Tested Positive
Tested Negative	425	75
Tested Positive	104	164

The following Table 4 depicts detailed accuracy for different k-folds for Random Tree algorithm.

Table-IV : Accuracy for different k-folds for Random Tree algorithm

K=n	TP Rate	FP Rate	Precision	Recall	F-Measure	Time taken to build Modeel	Error rate	Accuracy
10	0.836	0.388	0.8	0.83	0.81	0.24Secs	24.2	75.78
9	0.836	0.403	0.66	0.83	0.62	0.22 Secs	24.7	75.26
8	**0.85**	**0.388**	**0.8**	**0.85**	**0.82**	**0.02Secs**	**23.3**	**76.69**
7	0.84	0.422	0.78	0.84	0.81	0.21 Secs	25.1	74.86
6	0.852	0.403	0.79	0.85	0.82	0.22 Secs	23.6	76.3

VII. CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is a very challenging task in the field of health care. There are different techniques in data mining that has proven to be very helpful in decision making. In this paper we have applied Random Forest Tree data mining classification techniques is used to classify the diabetic disease. K-fold cross validation is also performed. The Random Forest Tree Algorithm gives 76.69% accuracy,23.30%error rate and less time 0.02 Seconds with k=8 folds cross validation while compared with K=6,7,9,10. As a future work the same technique is used to apply for other disease datasets such as heart disease, Breast Cancer, hepatitis and so on.

REFERENCES

1. www.wekipedia.com
2. https://www.cs.waikato.ac.nz/ml/weka/
3. https://stats.stackexchange.com/q/90288
4. https://stackoverflow.com/questions/10437677/cross-validation-in-weka/10539247
5. Habtamu Wondifraw Baynest, Classification, Pathophysiology, Diagnosis and Management of Diabetes Mellitus,ISSN: 2155-6156 JDM, an open access journal Volume 6 • Issue 5 • 1000541 J Diabetes Metab.
6. [Quan Zou,1,2,* Kaiyang Qu,1 Yamei Luo,3 Dehui Yin,3 Ying Ju,4 and Hua Tang5,*],Predicting Diabetes Mellitus with Machine Learning Techniques, Published online 2018 Nov 6. doi: 10.3389/fgene.2018.00515.
7. [Tejas N. Joshi*, Prof. Pramila M. Chawan**] Diabetes Prediction Using Machine Learning Techniques, ISSN: 2248-9622, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13
8. [Md. Aminul Islam, Nusrat Jahan] Prediction of Onset Diabetes using Machine Learning Techniques, International Journal of Computer Applications (0975 – 8887), Volume 180 – No.5, December 2017