ISSN:0975-9646

Predictive Analysis of Diabetic Treatment Using Classification Algorithm

M.Mounika, S.D.Suganya, B.Vijayashanthi, S.KrishnaAnand

School of Computing, Sastra University
Thanjavur- 613401,India

Abstract: Data mining is a concept whose usage has grown peak over the years. Its usage could be found prevalent in a wide variety of applications. Medical science is no exception. With this view in perspective, it has been decided to explore its functioning in the specialization of diabetes. Sufficient amount of study has been made for insulin-dependent and also adult-onset diabetes. However, enough amount of work has not been expended considering both the types in perspective. This work is a step in this direction. Besides, as massive number of criterion play an essential role in its occurrence, less delicate parameters would be pruned out. As most of the criterion does not deal with definite numerical values, linear regression model is not found to be suitable. Most of the working carried out deal with classification aspects. Studies have exhibited that Naïve Bayes algorithm is found to be the most efficient while dealing with classification. Hence, this method has been embodied into the work.

Keywords: Data mining, Classification model, Weka tool, Naïve Bayes.

1. INTRODUCTION:

With rapid refinement taking place over the globe, advancement in technology is significant. The area of medicine is not far behind. However, the desired expectations have not been met as far as diagnosis of some diseases is concerned. One such disease namely diabetes has been taken into consideration. Diabetes affects millions of people and is a significant long-lasting health problem. However, safeguarding diabetes in control is a tough task as self controlling measures have to be adopted for a significantly large percentage of people across the globe.

Diabetes is a condition that causes high blood glucose. It cannot be entirely cured but can thoroughly be managed. There are primarily two types of diabetes, the first one is Type-1(insulin-dependent), and it is treated with systematic insulin injections and a nutritious diet. The second one is Type-2 (insulin resistance). The same is treated by adopting diet changes, adequate amount of exercise and prevention of smoking apart from other minor measures. Regular medication, medicines and injections play a critical role in the management of diabetics.

The classification algorithms have been used to predict blood glucose level of diabetes across different age groups. It has been observed that Naïve Bayes classification algorithm finds effective treatment for all age group patients when compared with other classification algorithms such as OneR and ZeroR. Comparative analysis has been made between classification algorithms and proven that Naïve Bayes classifies the instance with a higher degree of accuracy as compared to other classification algorithms.

2. RELATED WORK:

Data mining techniques provide a tool to generate various classifications. Danielle M. Hesseler, Lawrence Fisher, Joseph T.Mullan, Russell E. Glasgow and Umesh Masharani [5] in the year 2011 had studied the neglector factor when considering disease management in adults with Type-2 diabetes and focused on how age and life factors are associated with diabetes and its management. Fotiadis and Manis studied Automated Diagnosis of Diseases Based on Classification and presented an automated method using random forest algorithm with the help of an online fitting procedure in the year 2012.

Later in the year 2014, Mira Kania Sabariah, Aini Hanifa and Siti Sa'adah [11] proposed Early Detection of Type-2 Diabetes Mellitus with Random Forest, Classification and Regression Tree in which they studied that early detection is necessary to classify the diabetic patient. The factors include complex attributes for analyzing the characteristics of dataset. L. Xu, et al., [10] had already studied the prediction diabetes in Chinese people. Recalibration of the Framingham diabetes score on Guangzhou Biobank Cohort Study in 2014. This study is applied to readily available clinical data. Lifestyle factors have been used for diabetic prediction. Besides, Logistic regression model was used to construct the diabetic prediction.

Janice, Lopez, Robert Bailey, Marcia and Kathy Annunziata [8] dealt with the which in turn is based on patient perspective with respective to age and ethnicity and risk factors includes older age, family history, and obesity and physical inactivity. Abdul Sattar Khan, Memet Isik, Turan Set, Zekeriya Akturk and Umit Avsar [2] studied a 5-year trend of myocardial infarction, hypertension, stroke and diabetes mellitus in gender and different age groups in 2014. They used logistic regression model for the group of variables.

3. PROPOSED APPROACH:

Enough amount of research has not been conducted by clubbing both Type-1 and Type-2 diabetes. This work has provided a special emphasis towards this direction. Besides, as large numbers of parameters play a role in its occurrence less sensitive parameters would be pruned out. As most of the parameters do not deal with definite categorical values, linear regression model is not found to be suitable. Classification algorithms such as Naïve Bayes, OneR and ZeroR are found to be suitable for categorical data. The factors like diet maintenance, drug intake, extends of smoking, level of obesity, insulin deficiency are taken into consideration for predicting blood glucose level among different age groups. Various classification algorithms are implemented in weka tool for estimating the performance and the accuracy of each algorithm.

4. METHODS:

4.1 Data collection:

Dataset is collected from clinic with parameters such as 'age', 'diet' , 'smoke', 'type' , 'drug' , 'insulin', 'obesity' ,'level'. **(Table 4.2)** consists of eight columns (Age, Diet, Drug, Smoke, Insulin, Type, obesity, Level) 'age' indicates the age group of patients with three categorical values 'young', 'adult' and 'old'. 'young' indicates the patients with age range between 10-25, 'adult' indicates the patients with age range between 26-49 and 'old' indicates the patients with age range above 50.'Diet' deals with two parameters 'vegetarian' and 'non-vegetarian' based on the type of food they choose. 'Drug' includes two parameters such as 'prescribed' and 'not-prescribed' according to the oral medications prescribed by the doctors. 'Smoke' deals with two parameters such as 'yes' and 'no' based on their habits and lifestyle. 'Insulin' deals with two parameters 'needed' and 'not-needed' which is for patients who must take insulin in order to manage their blood glucose level. 'Type' indicates 'Type-1' is insulin dependent and 'Type-2' which is insulin resistant. 'obesity' deals with two parameters 'average' and 'underweight' based on body mass index value. 'level' deals with two parameters 'high' and 'normal' based on the blood sugar level of the patient.

4.2 Tools and techniques:

Various data mining tools are available each has its pros and cons. For the analysis of diabetic data classification algorithms are used to find which treatment is effective for patients with different age groups and to find the efficient classification algorithm for the analysis.

Weka version 3.6.12 is used to find the efficiency of algorithm. Weka is a graphical user interface which is used for data analysis and predictive modeling written in java. In weka, main interface called Explorer is a component based knowledge flow interface and Experimenter which provides comparison of the performance of different machine learning algorithms.**Fig.4.1** shows the architecture for data mining.

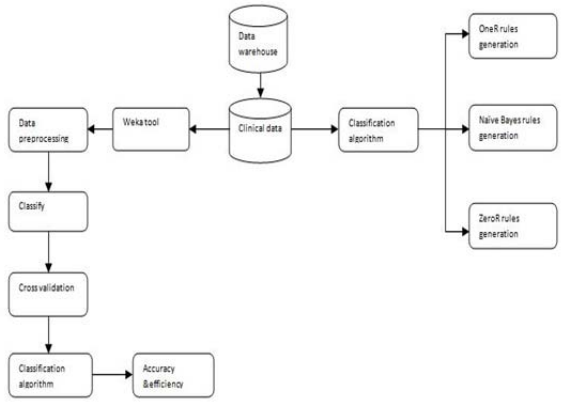

Fig.4.1: Architecture for data mining

The processing blocks includes

A. Data Warehouse:

Data warehouse stores current and historical data which is used for reporting and data analysis. It is an integration of data from various data.

B. Clinical Data:

Clinical data are the details of the diabetic patients collected from hospitals which are used for predicting the blood glucose level of patients.

C. Weka tool:

Weka is a java based machine learning tool and it is a collection of open source machine learning algorithms such as classifier, pre-processing, clustering and association rules.

D. Data preprocessing:

Data in real world are generally incomplete, inconsistent and noisy. Data preprocessing includes steps such as data cleaning, reduction, transformation and discretization helps in solving the above issues.

E. Cross validation:

Cross validation is a validation technique which is used to find the accuracy of the predictive model. It is also used to eliminate the issues like over fitting.

F. Classification algorithm:

Classification algorithms such as Naïve Bayes, OneR and ZeroR are taken into consideration. Clinical datasets are validated using weka tool for the above mentioned algorithms in order to find the accuracy and efficiency and the rules for respective algorithm are generated.

G. OneR rules generation:

OneR is one of the classification algorithms that generate a specific set of rules that test only one attribute for entire dataset.

H. Naïve Bayes rules generation:

Naive Bayes is one of the most efficient classification algorithms that generate a set of rules for each attribute in the dataset.

I. ZeroR rules generation:

ZeroR is one of the classification algorithms which depends on the target and ignores rest of the predictors. It only predicts the majority of the category

4.2.1 Naïve Bayes:

The Bayesian classification represents a supervised learning method as well as statistical classification. It can solve diagnostic and predictive problems. Naïve Bayes classifier is very adaptable, obliging various parameters straight in the quantity of variables in the learning issue. Naïve Bayes takes linear time, as opposed to iterative time as used by many of the other types of classifiers. Naïve Bayes generates rules for each attribute. Using the training datasets collected from the clinic, model is generated.

The conditional probabilities for each feature value in the test data are calculated by getting the count of instances with that feature value in a particular class and dividing it by the count of instances with the same class in the training set. This is done for each class in the data set.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P (A) is the prior probability of A

P (B) is the prior probability of B

P (A|B) is the posterior probability of A given B

4.2.1.1 Experimental analysis:

R programming language is used to perform the analysis of

datasets. Datasets are classified into young, adult and old according to their ages of the diabetic patients.The sample clinical dataset for young age patients as shown in **Table 4.2**

AGE	DIET	DRUG	SMOKE	TYPE	LEVEL	INSULIN	OBESE
Young	Non vegetarian	prescribed	yes	two	High	Not needed	Underweight
Young	Vegetarian	Not prescribed	no	one	normal	Not needed	Average
Young	Non vegetarian	Not prescribed	yes	one	normal	Needed	Underweight
Young	Non vegetarian	Not prescribed	no	one	normal	Needed	Average
Young	Vegetarian	Not prescribed	no	one	High	Needed	Average
Young	Vegetarian	Not prescribed	no	one	normal	Needed	Average
Young	Vegetarian	prescribed	no	two	normal	Not needed	Average
Young	Vegetarian	prescribed	no	two	High	Not needed	Average

Table 4.2: Sample dataset

Naïve Bayes classification is one of the best algorithms that suits both nominal and numerical data for evaluation. It generates rules for all the parameters in the dataset. The rules generated for the dataset mentioned in **Table 4.2** as shown in **Fig 4.2.**

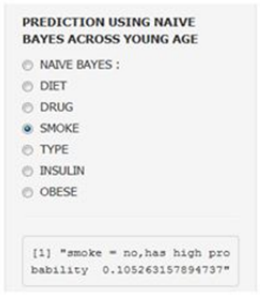

Fig 4.2: Probability value for smoke

For example, the factors like diet, drug, smoke, obese, type, insulin are taken into consideration for young age patients. As smoke is concerned, the people belonging to young age group are advised not to smoke which has the highest probability value. Similarly, the highest probability for all the factors can be analyzed by applying Naïve Bayes classification algorithm. This in turn will help in prediction of the changes in lifestyle needed for maintaining the blood sugar level. The probability value of diet for young age patients is predicted in RStudio as shown in **Fig 4.2.**

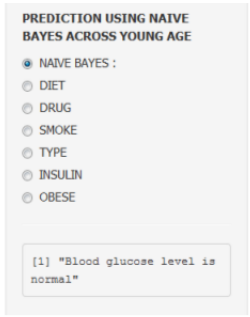

Fig 4.3: Blood glucose level for young age patients

By applying Naïve Bayes algorithm on collected clinical datasets, the highest probability value for all the factors can be identified. Depending upon the Probability, Naïve Bayes rule has been generated. From **Fig 4.3**, it has been found that the blood glucose level of young age patients is ***normal,*** if they are intimated by the doctor to follow the above treatments.

4.2.2 OneR:

OneR is a one rule algorithm which is utilized to foresee one rule for every attribute and picks the characteristic with most minimal error rate. OneR constructs frequency table for counting the most occurrences of each attribute. It then finds the net class weight for each attribute, by comparing new values of every attribute.

4.2.2.1 Experimental analysis:

OneR algorithm creates one rule for every attributes. In order to create a rule the frequent class for every attribute should be determined. By using the sample dataset in **Table 4.2** the rules for OneR is generated as shown in **Fig 4.4.**

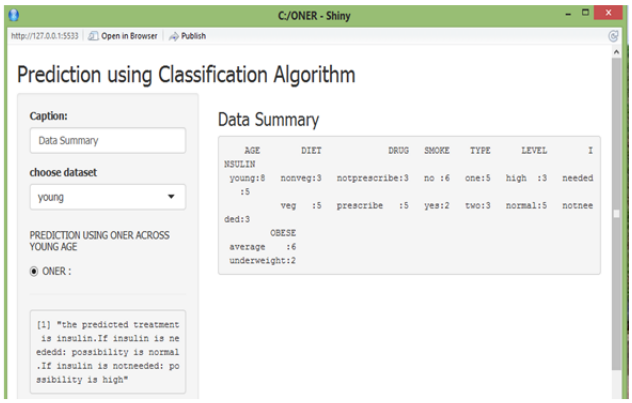

Fig 4.4: Rule generated for young age patients

For example, the factors like diet, drug, smoke, obese, type, insulin has been taken into consideration for young age patients. But OneR algorithm can generate only one rule for all the attributes depending on the smallest error rate. It has been observed from **Fig 4.4,** that the people having high glucose level are advised to undergo insulin treatment whereas the people with normal blood glucose level are not advised to undergo insulin treatment.

4.2.3 ZeroR:

ZeroR is one of the classification algorithms which depends on the target and ignores rest of the predictors. It only predicts the majority of the category. It is the baseline performance for all other classification algorithm.

4.2.3.1 Experimental analysis:

ZeroR algorithm predicts only the majority class. ZeroR does not make use of predictor classes rather it considers only the target class. From **Table 4.2**, we have target class as **Level.** The rule generated as shown in **Table 4.3.**

Level	
High	**Normal**
3	4

Table 4.3: Rule generated for ZeroR

4.2.4 Performance analysis:

It has been observed that Naïve Bayes algorithm classifies the instances more efficiently than OneR and ZeroR. This has been clearly highlighted in **Table 4.4**. Error rate of classification is also less for Naïve Bayes algorithm when compared with other two algorithms is highlighted in **Table 4.5**

Algorithm	Naive Bayes	OneR	ZeroR
Correctly classified Instances	97.43 %	88.20 %	67.17 %
Incorrectly classified Instances	2.56 %	11.79 %	32.82 %
TP Rate	0.974	0.882	0.672
Precision	0.975	0.895	0.451
F Measure	0.974	0.884	0.54
ROC Area	0.975	0.892	0.5
Kappa Statistics	0.941	0.745	0

Table 4.4 Performance evaluation of different algorithm using weka tool

Algorithm	MAE	RMSE	RAE	RRAE
Naive Bayes	0.10	0.20	24.44	43.24
OneR	0.11	0.34	26.70	73.12
ZeroR	0.44	0.46	100	100

Table 4.5 Error rate of different algorithm using weka tool

5. RESULT AND DISCUSSION:

The data mining tool and classification algorithm is applied to clinical datasets for the prediction of blood glucose level. In this work, the successful diabetic treatment and effective classification calculation for predictions are discussed. Utilizing Naïve Bayes calculation the effective treatment for different age group such as young, adult and old are recorded and prediction is done accordingly. It is cleared that drug prescription is effective for old age groups with type-2 diabetes, drug prescription and dietary controls is effective for adult age groups where as young age groups need to concentrate on other factors such as dietary controls, physical exercise, smoke cessation and insulin. Therefore this prediction gives a positive mode of treatment for different age groups

Drug prescription is strongly recommended for old and adult age groups but for young age group drug prescription is not required, insulin is strongly recommended for young age groups with type-1 diabetes. There is a risk of side effects due to intake of drugs so young age groups are advised to follow other modes of treatments like dietary controls, exercising, insulin and smoke cessation but for adult and old age groups despite of side effects they need to take drugs to control their blood glucose level.

6. CONCLUSIONS AND FUTURE IMPROVEMENTS:

There are three different classification algorithms used in this experiment namely, ZeroR, OneR and Naïve Bayes. This experiment shows that Naïve Bayes is the fastest and ZeroR is the slowest. The performance comparison is found using weka data mining tool. Each model is converted into rules and those rules are incorporated into this application. It is found Naïve Bayes outperforms OneR and ZeroR. Classification algorithm such as Naïve Bayes, OneR and ZeroR is applied to diabetes datasets collected from the clinic and blood sugar level for young, old and adult patients is predicted using the rule generated by each models. This framework could be stretched out further to discover conceivable outcomes of different diseases. In spite of the fact that the planned framework is exceedingly productive and matches the doctor's finding, different strategies like data mining algorithm could likewise be gone for. These aspects could be left for further examinations.

ACKNOWLEDGMENTS:

We would like to thank almighty god and our parents. We extend thanks to Our Guide Prof. Dr.S.Krishna Anand was the driving force behind this whole idea from the start. We would also like to thank Dr.Sankara narayanan, Dean, Thanjavur Medical College for helping us out in providing patient records. We would like to extend our gratitude to all the staff of our university who has either directly or indirectly helped us in the completion of the project. We would like to acknowledge each and every one who had a role to play in ensuring our success.

REFERENCES:

[1] Abdullah A.Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, Application of data mining: Diabetes health care in Young and old patients, *Journal of King Saud University*, 25(2013), 127 – 136.

[2] Abdul Sattar Khan, Memet Isik, Turan Set, Zekeriya Akturk and Umit Avsar, 5-year trend of myocardial infarction, hypertension, stroke and diabetes mellitus in gender and different age groups, *Journal of Taibah University Medical Sciences*, 3(2014), 198-205.

[3] American Diabetes Association., *Living with Diabetes: Insulin Basics*, June7, 2013:http://www.diabetes.org/living-with-diabetes/treatment-andcare/medication/insulin/insulin-basics.html

[4] Bandana Garg, Design and Development of Naïve Bayes Classifier, *North Dakota State University*, 2013.

[5] Danielle M.Hessler, lawrence Fisher, Joseph T.Mullan, Russel E.Glasgow, Umesh Masharani., Patient age: A neglected factor when considering disease management in adults with type 2 diabetes, *Elsevier Journal*,85 (2011), 154–159.

[6] Dario Antonelli, Elena Baralis, Giulia Bruno,Tania Cerquitelli, Silvia Chiusano, Naeem Mahoto, Analysis of diabetic patients through their examination history, *Elsevier Journal*,40 (2013), 4672-4678.

[7] Gaya Buddhinath, Damien Derry., A Simple Enhancement to One Rule Classification, *Department of Computer Science & Software Engineering University* of Melbourne.

[8] Janice.M.S.Lopez, Robert.A.Bailey, Marcia.F.T.Rupnow, KathyAnnunziata., 2014, Characteri-zation of type2 diabetes Mellitus Burden by age and ethnic Groups Based on a Nationwide Survey, *Elsevier Journal*,36 (2014).

[9] Joseph L.Breault, Colin R.Goodall, Peter J.Fos, Data mining a diabetic data warehouse, *Elsevier Journal*, 26 (2002), 37-54.

[10] L.Xu,C.Q.Jiang,C.M.Schooling,W.S.Zhang,K.K.Cheng,T.H.Lam., Prediction of 4-year incident diabetes in older chinese: Recalibration of Framingham diabetes score on Guangzhou Biobank Cohort Study, *Elsevier Journal*, 69 (2014), 63–68.

[11] Mira Kania Sabariah, Aini Hanifa and Siti Sa'adah, Early Detection of Type-2 Diabetes Mellitus with Random Forest, Classification and Regression Tree, *IEEE Transaction*, 2014, 238-242,.

[12] Swasti Singhal, Monika Jena, A Study on Weka Tool for Data Preprocessing, Classification and Clustering, *IJITEE*, 2 (2013), Issue-6.

[13] Tina R.Patil, Mrs.S.S.Sherekar , Performance Analysis Of Naïve Bayes and J Classification Algorithm for Data Classification, *Journal of Computer Science and Application*, 6 (2013), No.2.

[14] Vijayarani.S, Muthulakshmi.M., evaluating the efficiency of rule techniques for file classification, *International Journal of Research in Engineering and Technology*, 02(2013).