Predictive Analytics in Diabetes using oneR Classification Algorithm

K. B. Priya Iyer,PhD
Associate Professor (CSC)
M.O.P. Vaishnav College for Women
Nungambakkam, ch-34

K. Pavithra
BCA Graduate
M.O.P. Vaishnav College for Women
Nungambakkam, ch-34

D. Nivetha
BCA Graduate
M.O.P. Vaishnav College for Women
Nungambakkam, ch-34

K. Kumudha Varshini
BCA Graduate
M.O.P. Vaishnav College for Women
Nungambakkam, ch-34

ABSTRACT

Data mining is becoming one of the most important tool and means for discovering useful, realistic and essential interpretations and patterns in real world scenario. The discovery of knowledge from medical datasets is crucial in order to make effective medical diagnosis. This paper helps in predicting diabetes by applying the data mining technique called oneR rule. In this paper, oneR rule in classification has been used to predict a person whether diabetic or not. The dataset which had been collected contains the information of persons with and without diabetes. Software used is Weka tool for the experiment and analysis. oneR algorithm is applied on the dataset of persons collected. Results have been obtained. The accuracy is calculated as 80.43 % which is high for predicting whether the person has diabetes or not.

General Terms

oneR classification algorithm.

Keywords

Data mining, Diabetes, Weka, oneR, Classification.

1. INTRODUCTION

Data mining is often described as the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses [1]. Data mining provides a great potential in the healthcare industry to systematically use data and analysis to identify inefficiencies and to improve care and reduce costs[2]. Diabetes is the most common endocrine disease across all population and age groups. Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high[3].

Advantages of data mining in diabetes:

- ➢ Determine the diseases and conditions are driving trends.
- ➢ Identify gaps in medical treatment

Disadvantages of data mining in diabetes:

- ➢ Misuse of information/inaccurate information
- ➢ Amount of data is overwhelming

2. LITERATURE SURVEY

The Paper[4] proposed the application Information technology of knowledge-based DSS for analysis diabetes of elder using decision tree. The NBTree model has lowest accuracy in the classification is 70.60 percent

when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454.

In another Research paper, Classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3% [5].

The paper[6] has proposed a novel learning algorithm i+Learning as well as i+LRA, which apparently achieves the highest classification accuracy over ID3 algorithm. The major limitation of their method is the adoption of binary tree rather than multi-branch tree. Such structure increases the tree size, whereas an attribute can be selected as a decision node for more than once in a tree

3. CONCEPTUAL FRAMEWORK

3.1. Data Mining

Data mining is the process of discovering, analyzing and summarising data from different perceptive into useful information.

3.2. Weka Tool

Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand and it is also free licensed under the GNU General Public License.

3.3. Classification

Classification is a data mining function that able to assign items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

3.4. OneR Rule

OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, we construct a frequency table for each predictor against the target.

3.4.1. OneR Algorithm

For each predictor,

For each value of that predictor, make a rule as follows;
Count how often each value of target (class) appears
Find the most frequent class
Make the rule assign that class to this value of the predictor
Calculate the total error of the rules of each predictor
Choose the predictor with the smallest total error.

4. PROPOSED WORK

4.1 Problem Statement

To identify whether a given person in dataset will be diabetic or non diabetic will be done on the basis of attribute values. The Dataset contains all the details of a person. Attributes like Diastolic blood pressure (mm Hg) and serum insulin values exceeding a specific value may contribute to identify whether a person is diabetic or non diabetic. A brief explanation has been given below through a flow chart.

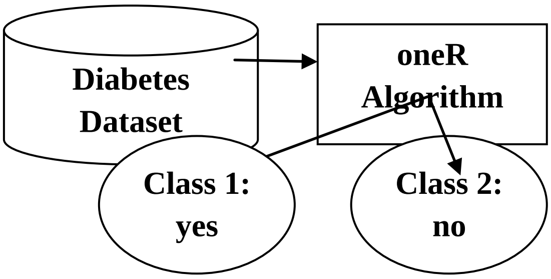


Figure 1. Flow Chart of Problem

The collected dataset is fed into the software weka which will output the total no of diabetic and non diabetic persons. The classification is based upon the primary attributes values. This technique would classify the dataset into two different classes.

4.2 Results and Discussions

The dataset that is taken for this research work contains 768 records and 9 attributes for the purpose of predicting diabetes based on the symptoms. This dataset is designed in arff format, which is considered as the most suitable input format for the weka software.

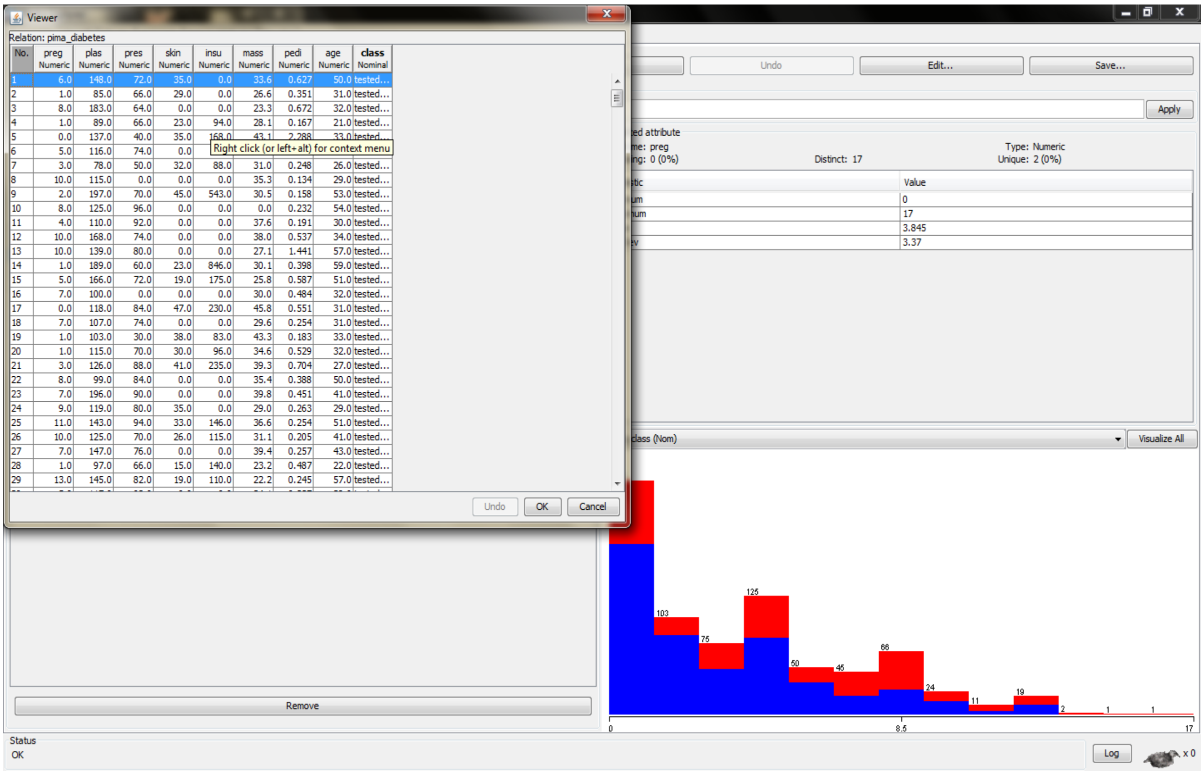
4.3 Preparing Dataset

The dataset used contains 206 instances and all instances have 9 input attributes (X1 to X8) and one output attribute (Y1).

4.4 Attributes of Dataset

Table 1 Attribute description of diabetes dataset.

Attribute no	Attribute	Description	Type
X1	preg	No of times Pregnant	Numeric
X2	plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
X3	pres	Blood pressure	Numeric
X4	skin	Triceps skin thickness	Numeric
X5	insu	Serium insulin	Numeric
X6	mass	Body mass index	Numeric
X7	pedi	Diabetes pedigree function	Numeric
X8	Age	Age of person	Numeric
Y1	Class	Diabetes results	Nominal



Viewer

Relation: pima_diabetes

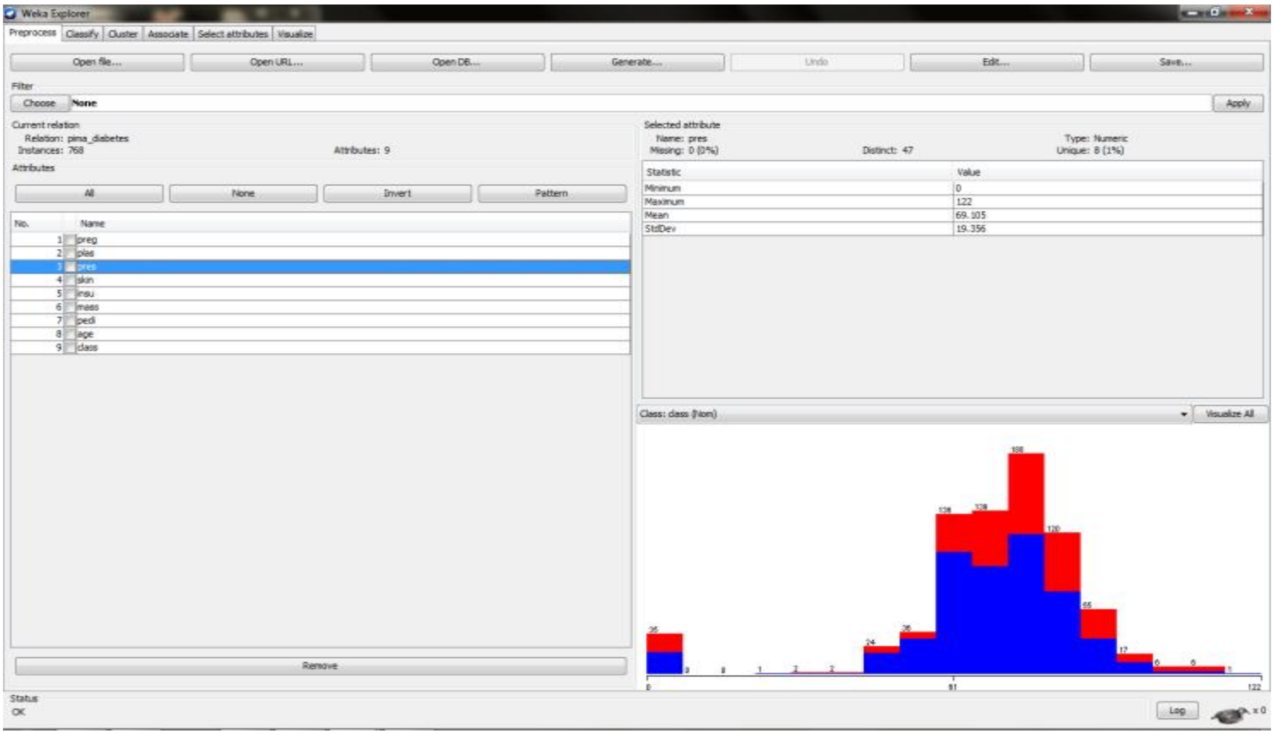
No.	preg Numeric	plas Numeric	pres Numeric	skin Numeric	insu Numeric	mass Numeric	pedi Numeric	age Numeric	class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested...
6	5.0	116.0	74.0	0.0					
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested...
25	11.0	143.0	94.0	33.0	146.0	36.6	0.254	51.0	tested...
26	10.0	125.0	70.0	26.0	115.0	31.1	0.205	41.0	tested...
27	7.0	147.0	76.0	0.0	0.0	39.4	0.257	43.0	tested...
28	1.0	97.0	66.0	15.0	140.0	23.2	0.487	22.0	tested...
29	13.0	145.0	82.0	19.0	110.0	22.2	0.245	57.0	tested...

Right click (or left+alt) for context menu

ted attribute
me: preg
ing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

tic	Value
um	0
num	17
	3.845
ev	3.37

Figure 2 Diabetes Dataset Used For Prediction



Weka Explorer

Current relation
Relation: pima_diabetes
Instances: 768
Attributes: 9

Filter: None

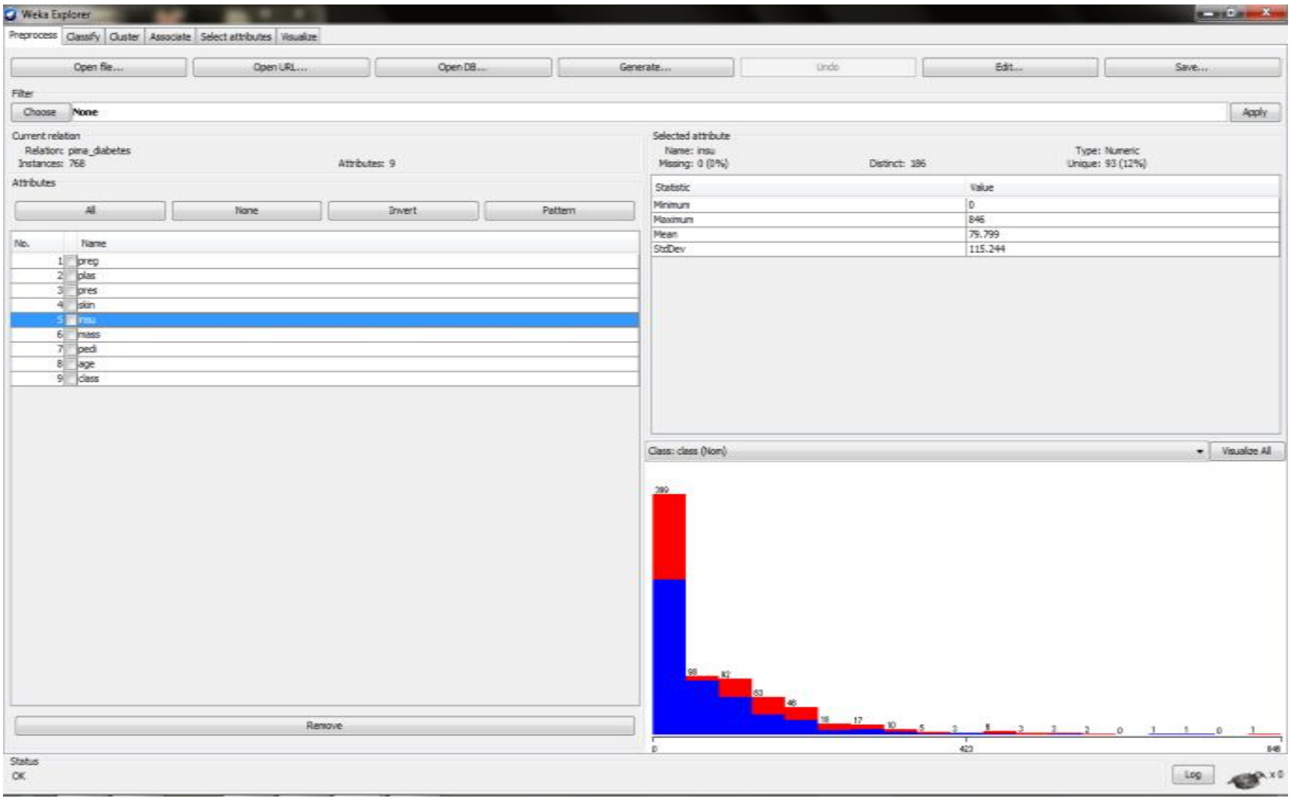
No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

Selected attribute
Name: pres
Missing: 0 (0%)
Distinct: 47
Type: Numeric
Unique: 8 (1%)

Statistic	Value
Minimum	0
Maximum	122
Mean	69.105
StdDev	19.356

Class: class (Nom)

Figure 3. Attribute Histogram with Proper Values (Diastolic Blood Pressure)



Filter: None

Current relation
Relation: pima_diabetes
Instances: 768
Attributes: 9

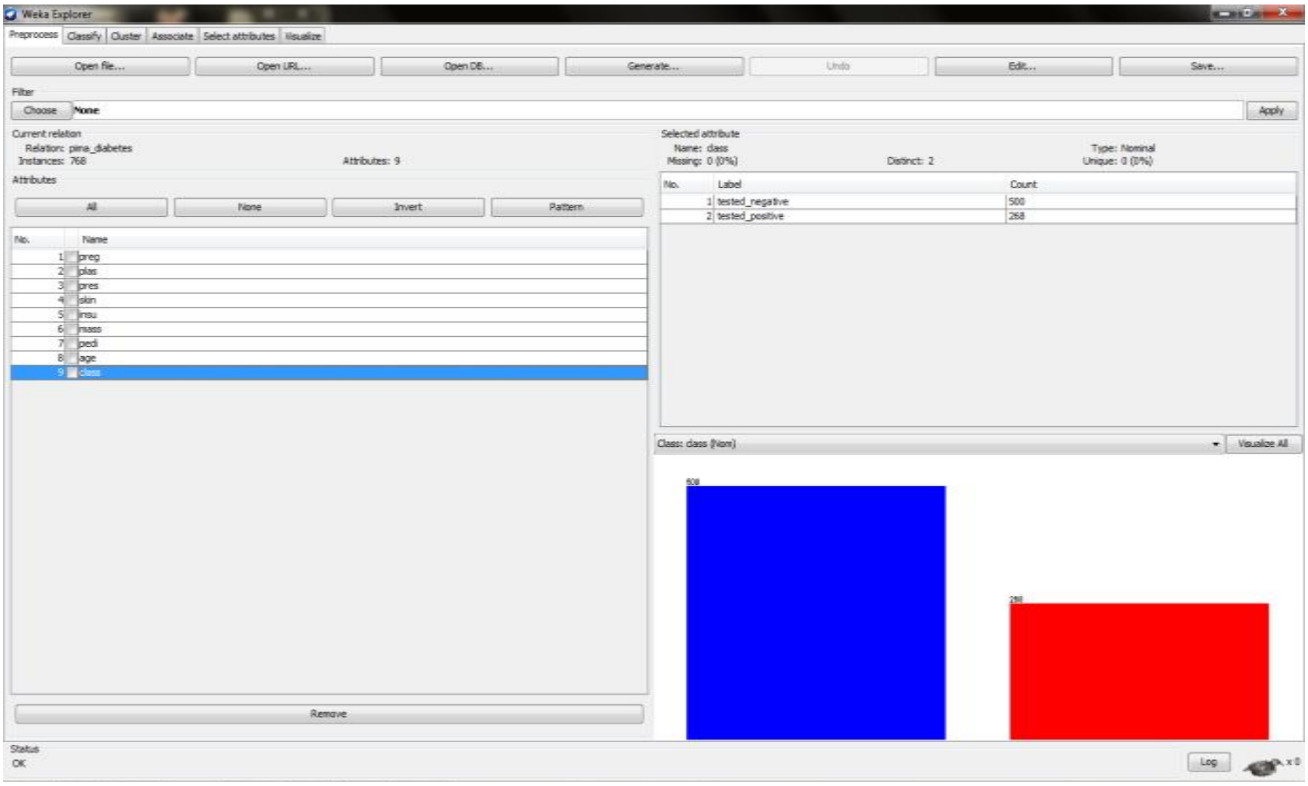
No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

Selected attribute
Name: insu
Type: Numeric
Missing: 0 (0%)
Distinct: 186
Unique: 93 (12%)

Statistic	Value
Minimum	0
Maximum	846
Mean	79.799
StdDev	115.244

Status
OK

Figure 4 Attribute Histogram with Proper Values (Serum Insulin)



Filter: None

Current relation
Relation: pima_diabetes
Instances: 768
Attributes: 9

No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

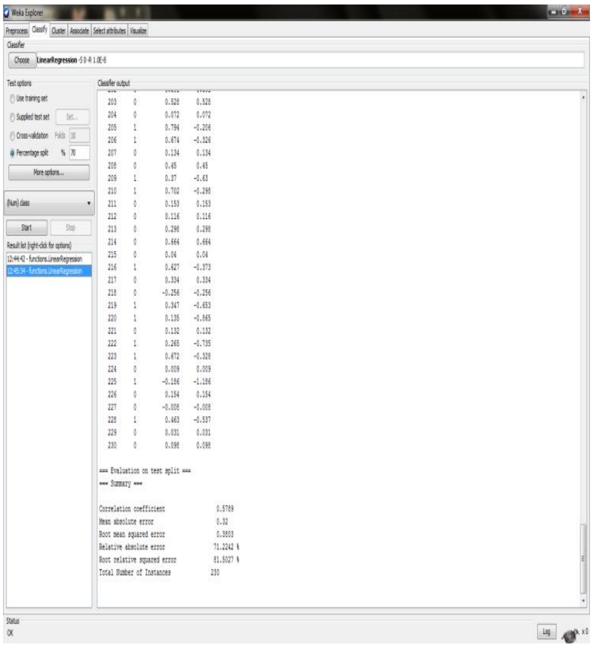
No.	Label	Count
1	tested_negative	500
2	tested_positive	268

Status
OK

Figure 5 Class Histogram with Proper Values

4.5. Problem

Linear regression is used for the prediction of diabetes from the collected dataset. This is considered as the best suitable technique for prediction process of any dataset. Linear regression uses binary data for its processing. The data collected was nominal data. In order to convert nominal data into a binary data, filters had been used.The linear regression had been applied onto the binary diabetes dataset. The result obtained is given in figure 6.



LinearRegression -S 0 -R 1.0E-8

○ Use training set ○ Supplied test set ○ Cross-validation Folds 10 ● Percentage split % 70

(Num) class

12:44:42 - functions.LinearRegression
12:45:34 - functions.LinearRegression

```
    203      0         0.528     0.528
    204      0         0.072     0.072
    205      1         0.794    -0.206
    206      1         0.674    -0.326
    207      0         0.134     0.134
    208      0         0.45      0.45
    209      1         0.37     -0.63
    210      1         0.702    -0.298
    211      0         0.153     0.153
    212      0         0.116     0.116
    213      0         0.298     0.298
    214      0         0.664     0.664
    215      0         0.04      0.04
    216      1         0.627    -0.373
    217      0         0.334     0.334
    218      0        -0.256    -0.256
    219      1         0.347    -0.653
    220      1         0.135    -0.865
    221      0         0.132     0.132
    222      1         0.265    -0.735
    223      1         0.672    -0.328
    224      0         0.009     0.009
    225      1        -0.186    -1.186
    226      0         0.154     0.154
    227      0        -0.008    -0.008
    228      1         0.463    -0.537
    229      0         0.031     0.031
    230      0         0.098     0.098

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.5709
Mean absolute error                      0.32
Root mean squared error                  0.3803
Relative absolute error                 71.2242 %
Root relative squared error             81.5027 %
Total Number of Instances              230
```

Figure 6 Result of Linear Regression

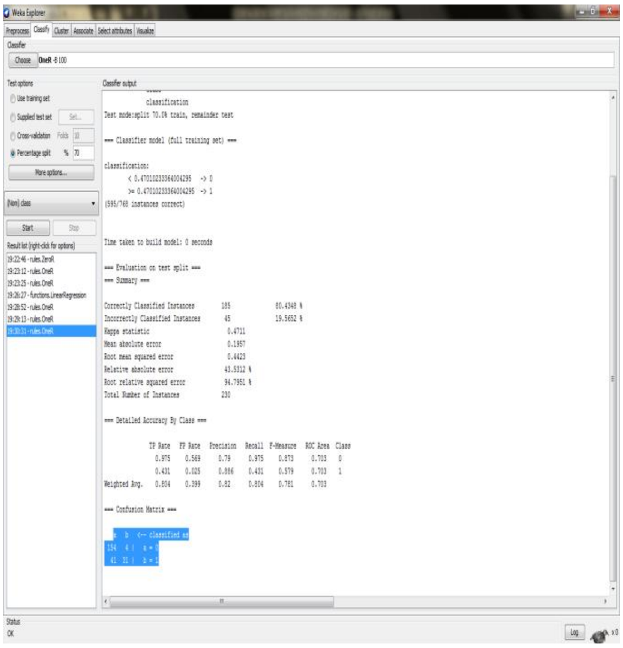
The result has been summarized as follows,

Correlation coefficient is 0.5709, Mean absolute error is 0.32, Root mean squared error is 0.3803, Relative absolute error is 71.2242 %, Root relative squared error is 81.5027 % and the total number of instances is 230. This is not considered as a better result for diabetes dataset. Hence alternative method has been used.

The applied technique for obtaining better prediction for diabetes dataset is classification by regression method. Since, classification by regression is applied to the class attribute that are nominal values, the remaining other attributes are removed. To apply this method, a new attribute called classification has been added. Thus, the regression is applied for class and classification.

4.6 OneR Method For Prediction

oneR method is applied to get the prediction value for the diabetes dataset. By increasing bucket size we have a better prediction.



OneR -B 100

○ Use training set ○ Supplied test set ○ Cross-validation Folds 10 ● Percentage split % 70

(Nom) class

19:22:46 - rules.ZeroR
19:23:12 - rules.OneR
19:23:25 - rules.OneR
19:26:27 - functions.LinearRegression
19:28:52 - rules.OneR
19:29:13 - rules.OneR
19:30:31 - rules.OneR

```
              classification
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===

classification:
        < 0.47010233364004295   -> 0
        >= 0.47010233364004295  -> 1
(595/768 instances correct)


Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         185               80.4348 %
Incorrectly Classified Instances        45               19.5652 %
Kappa statistic                          0.4711
Mean absolute error                      0.1957
Root mean squared error                  0.4423
Relative absolute error                 43.5312 %
Root relative squared error             94.7951 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.975     0.569      0.79      0.975     0.873      0.703    0
                 0.431     0.025      0.886     0.431     0.579      0.703    1
Weighted Avg.    0.804     0.399      0.82      0.804     0.781      0.703

=== Confusion Matrix ===

   a   b   <-- classified as
 154   4 |   a = 0
  41  31 |   b = 1
```

Figure 7 Result of OneR Rule

From figure 7, the confusion matrix,

$$\begin{array}{ll} a \quad b & <-- \ classified\ as \\ 154 \quad 4 \,| & a = 0 \\ 41 \quad 3 \,| & b = 1 \end{array}$$

4.7 Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system.

	Predicted	
	Positive	Negative
Actual True	**TP**	**FN**
Actual False	**FP**	**TN**

Figure 8: Confusion Matrix

- ➢ True positive (TP) - These are the positive tuples that were correctly labeled by the classifier [7].
- ➢ True Negative (TN)-These are the negative tuples that were correctly labeled by the classifier [7].
- ➢ False Positive (FP)-These are the negative tuples that were incorrectly labeled as positive [8].
- ➢ False Negative (FN)-These are the positive tuples that were mislabeled as negative [7].

Thus accuracy is calculated as

$$\frac{(TP + TN)}{(P + N)} \quad where$$

$$P = TP + FN\ and$$

$$N = FP + TN\ Or\ TP + TN\ /\ (TOTAL)$$

According to experimental results, correctly classified instances for oneR method is 595 and the accuracy is 80.43 % which is high. oneR method is a promising technique for this type of dataset

From the figure of detailed accuracy by class has the calculated TP rate and FP rate which gives weighted average accuracy. This gives the correct split point with correct instances for the supplies diabetes data set.

5. CONCLUSION

The knowledge discovery from medical dataset is important to make crucial medical diagnosis. Diabetes is a disease that has many different complications which has an impact on society. This work aims at the discovery of new set of values to predict the presence or non presence of diabetes in a person. Diabetes dataset is collected and Pre-processing is used to improve the quality of data. The dataset is processed using oneR method to have accuracy 80.43 % which is higher than linear regression in weka. Correctly classified instances for oneR method is 595 and the accuracy is 80.43 %. The Mean absolute error is 0.1957, Root mean squared error is 0.4423, Relative absolute error is 43.5312 %, and Root relative squared error is 94.7951 %.

6. FUTURE SCOPE

There are some limitations of this study.

Firstly, as far as the diabetes dataset is concerned, there might be other risk factors that the data collections did not consider which include family history, metabolic syndrome, smoking, inactive lifestyles, certain dietary patterns etc. The proper prediction model would need more data gathering to make it more accurate. This can be achieved by collecting diabetes datasets from multiple sources, generating a model from each dataset.

Secondly, in this study linear regression has been used for prediction, but it did not supplied expected result. So, oneR method is used, which in turn offered high accuracy than former to predict diabetes. Therefore, in the near future there might be another method which can offer better result (accuracy) than oneR.

7. REFERENCES

[1] Mukeshkumari, Dr.Rajanvohra And Anshularora, “Prediction Of Diabetes Using Bayesian Network”, Mukeshkumari Et Al, / (Ijcsit) International Journal Of Computer Science And Information Technologies, Vol. 5 (4) , 2014, 5174-5178.

[2] Https://Www.Healthcatalyst.Com/Data-Mining-In-Healthcare

[3] Https://Www.Niddk.Nih.Gov/Health-Information/Diabetes/Overview/What-Is-Diabetes

[4] Sudajailowanichchai, Saisuneejabjone, Tidanutputhasimma, ”Knowledge-Based Dss For An Analysis Diabetes Of Elder Using Decision Tree”.

[5] Yang Guo , Guohuabai , Yan Hu School Of Computing Blekinge Institute Of Technology Karlskrona, Sweden, “Using Bayes Network For Prediction Of Type-2 Diabetes”.

[6] Beckles Gla, Thompson-Reid Pe, Editors. Diabetes And Women’s Health Across The Life Stages: A Public Health Perspective. Atlanta: U.S. Department Of Health And Human Services, Centers For Disease Control And Prevention, National Center For Chronic Disease Prevention And Health Promotion.

[7] Jiawei Han, Michelinekamber, Jian Pei, “Data Mining Concepts And Techniques” Third Edition.

[8] Sapna Jain 2.M Afshar Aalam3. M. N Doja,”K-Means Clustering Using Weka Interface”, Proceedings Of The 4th National Conference; Indiacom-2010 Computing For Nation Development, February 25 – 26, 2010.