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][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[3] 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% [3].

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.

![Flow Chart of Problem](image)

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>
<thead>
<tr>
<th>Attribute no</th>
<th>Attribute</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>preg</td>
<td>No of times Pregnant</td>
<td>Numeric</td>
</tr>
<tr>
<td>X2</td>
<td>plas</td>
<td>Plasma glucose concentration a 2 hours in an oral glucose tolerance test</td>
<td>Numeric</td>
</tr>
<tr>
<td>X3</td>
<td>pres</td>
<td>Blood pressure</td>
<td>Numeric</td>
</tr>
<tr>
<td>X4</td>
<td>skin</td>
<td>Triceps skin thickness</td>
<td>Numeric</td>
</tr>
<tr>
<td>X5</td>
<td>insu</td>
<td>Serium insulin</td>
<td>Numeric</td>
</tr>
<tr>
<td>X6</td>
<td>mass</td>
<td>Body mass index</td>
<td>Numeric</td>
</tr>
<tr>
<td>X7</td>
<td>pedi</td>
<td>Diabetes pedigree function</td>
<td>Numeric</td>
</tr>
<tr>
<td>X8</td>
<td>Age</td>
<td>Age of person</td>
<td>Numeric</td>
</tr>
<tr>
<td>Y1</td>
<td>Class</td>
<td>Diabetes results</td>
<td>Nominal</td>
</tr>
</tbody>
</table>
Figure 2 Diabetes Dataset Used For Prediction

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

From figure 7, the confusion matrix,
\[
\begin{array}{c|c|c}
\text{a} & \text{b} & \text{classified as} \\
154 & 4 & a = 0 \\
41 & 3 & b = 1 \\
\end{array}
\]

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

Thus accuracy is calculated as
\[
\frac{(TP + TN)}{(P + N)} \text{ where } P = TP + FN \text{ and } N = FP + TN \text{ or } TP + TN / (TOTAL)
\]

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