An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network

S.Priya
PG Scholar, Department of CSE
Kongu Engineering College
Perundurai, Erode-638 052

R.R.Rajalaxmi
Professor, Department of CSE
Kongu Engineering College
Perundurai, Erode-638 052

ABSTRACT
People in today's world get affected by many diseases that do not have a complete cure. The development of one disease may lead to various other complications. One such disease is Type-2 Diabetes. It is a global health problem. This is the most common type of diabetes usually developed at the age of 40 and older. This increases the risk factors like kidney failure, heart disease, blindness, nerve damage and blood vessel damage. It is predicted from the characteristics of the patients. A Hybrid Prediction Model (HPM) has been developed using k-means clustering and C4.5 classifier. In that model, the dataset is initially cleaned and then Z-score normalization is applied on this dataset. A pattern is extracted from this using clustering. A model was built on this extracted pattern using the c4.5 classifier. This produced an accuracy of 92.38%. This classification accuracy can be improved by using Neural network. This improved model separates the dataset into either one of the two groups. This model has revealed an accuracy of 97.93%. This earlier detection will help the physicians to reduce the probability of getting that disease.

Keyword
classification, clustering, Z-score normalization.

1. INTRODUCTION
Human body needs energy for activation. The carbohydrates are broken down to glucose. Glucose is the important energy source for the human body cells. Insulin is needed to transport the glucose into body cells. The blood glucose is supplied with insulin and glucagon hormones produced by pancreas. Insulin hormones produced by the beta cells of the islets of langerhans and glucagon hormones are produced by the alpha cells of the islets of langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated and insulin given to the blood. Insulin enables blood glucose to get in to the cells and this glucose is used for energy. So blood glucose is kept in a narrow range [4].

Diabetes is a serious global health problem. It is a disease in which body does not produce insulin or use it properly. This increases the risks of developing kidney disease, heart disease, blindness, nerve damage and blood vessel damage [5]. There are two types of diabetes namely: Type-1 and Type-2.

Type 1 diabetes can occur at any age. However, it is most often diagnosed in children, adolescents, or young adults.Type-1 diabetes occurs when the body's immune system is attacked and the beta cells of pancreas are destroyed. This results in insulin deficiency. The only treatment for this is insulin.

Type-2 diabetes occurs when the pancreas does not produce enough insulin to meet the body's needs. This is the most common type of diabetes developed at the age of 40. Recent studies have shown that 80% of type-2 diabetes complications can be prevented by earlier identification.

2. RELATED WORK
Data mining is the extraction of useful information from the large volume of data. Data mining has been applied in various fields like medicine, marketing, banking, etc. In medicine, predictive data mining is used to diagnose the disease at the earlier stages itself and helps the physicians in treatment planning procedure.

Majority of the work has already been carried out in the area of predictive classification for the Pima Indian diabetes datasets. Decision tree using the Rapid miner tool [1] has been used to build the prediction model on this datasets. This model considered that the Plasma Glucose attribute as the main attribute in predicting the disease. It has produced an accuracy of 72%. Another decision tree model has been built using the Weka's J48 decision tree classifier [2] on this dataset with an accuracy of 78.18%. Association rule mining has also been implemented on the same datasets. This produced large number of rules from the combinations of attributes.

Other classification methods like neural network approach obtained an accuracy of 75.4%, Bayesian approach achieved an accuracy of 79.5%. A number of classification algorithms on the same dataset have performed classification with accuracies in the range of 59.5-77.7% [3]. The improvement in the prediction accuracy can also be obtained by using improved Weighted Least Squares Support Vector Machine (WLS-SVM) based on Quantum Particle swarm Optimization (QPSO) algorithm as stated in [7].

The predictive data mining has also been applied in dosage planning for the Type-2 Diabetes [8]. Two models namely Adaptive Neuro Fuzzy Inference System (ANFIS) and Rough Set Theory based methods have been developed. ANFIS method was found to be reliable as it produced better sensitivity and lesser Root Mean Square Error values than Rough Set Theory model for the dataset. These models have been built from the real dataset with some more additional features like Creatinine, Cholesterol, urine level, etc. These features were also helpful in assessing the risks related to diabetes.

Earlier researches have also been carried out in predicting the risks related to the diabetes with some additional features and has revealed different accuracies. Association rule mining and classification techniques were useful in identifying the relationships edema and diabetes, and wheezes and diabetes as stated in [9].

3. PROPOSED MODEL
3.1. Data Preprocessing
The Pima Indian Diabetes dataset is collected from the UCI Machine Learning repository [6]. The data set contains 768 samples with eight attributes namely Pregnant (number of times of pregnant), Plasma-Glucose (Plasma glucose concentration), DiastolicBP (Diastolic blood pressure), TricepsSFT (Triceps

skin fold thickness), Serum-Insulin, BMI (Body Mass Index), DPF (Diabetic Pedigree Function) and Age. This dataset is found to have some missing values. So, some of the preprocessing techniques like data cleaning, feature selection and normalization are performed in order to improve the quality of the mining result.

Out of these eight attributes, TricepsSFT and Serum-Insulin attributes are found to contain large number of missing values. So these two features are removed. After removing these two attributes and the records with missing values only 625 instances remain from 768 samples. Z-score normalization as given in Eq.(1) is done on this data set using the Rapid Miner tool in order to bring the data set in a suitable form for mining. This is shown in Fig.1.

$$v' = (v - A')/\sigma \qquad (1)$$

Where v' is the mean for the variable
σ is the standard deviation for the variable and
v is the new normalized value

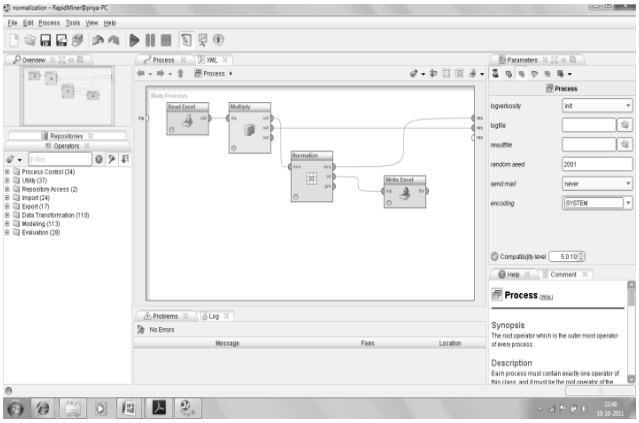

Figure. 1. Z-Score Normalization

3.2. Data Clustering

A simple K-Means clustering is performed on this data set using the Rapid miner tool. K-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. It uses Euclidean distance as the dissimilarity measure. The k-means algorithm can be divided into two phases: initialization phase and iteration phase. The k-means clustering algorithm is as follows:

Let D be a data set with n instances and C_1, C_2... C_k be the k disjoint clusters of D.

Initialization Phase:

1. (C_1, C_2…C_k)=initial partition of D.

Iteration Phase:

2. Repeat
3. D_{ij}=distance between case i and cluster j;
4. N_i=arg $\min_{1<=j<=k}$ d_{ij};
5. Assign case i to cluster n_i;
6. Recomputed the cluster means of any changed clusters above;
7. Until no further changes of cluster membership occur in a complete iteration.

The k-means clustering is done to validate the class label. This data set is grouped into two clusters as shown in Fig.2. Then a validation is performed against the class label. Out of 625 samples, 192 samples are found to be incorrectly classified as shown in Table 1. After removing these samples only 433 instances remain from the original data set.

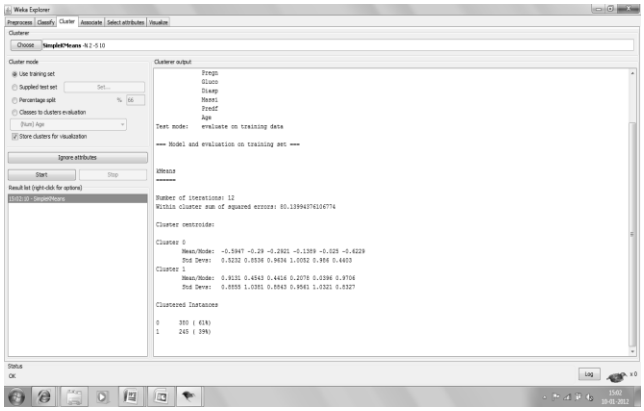

Figure 2. K-means clustering with k=2

Table 1. Clustering results

Samples	clusters	Clustered instances	Incorrectly classified instances
625	Cluster 1(yes)	245	192
	Cluster 0 (no)	380	

3.3. Data Classification

Classification is carried out using Neural network. This network learns a model by means of a feed-forward neural network trained by a back propagation algorithm. General structure of this network refers to the large number o highly interconnected processing elements (neurons) working together. The operations performed in the learning process are:

- Fixing the weights
- Simulating the network which equals the input signal flow through the network to the output
- Output is manipulated which is often the binary representation
- Based upon single operation and the threshold the network formulates fast decision and real time response.

The Neural network consists of three layers:

- **Input layer:** Input neurons define all the input attribute values for the data mining model, and their probabilities.
- **Hidden layer:** Hidden neurons receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. The greater the weight that is assigned to an input, the more important the value of that input is. Weights can be negative, which means that the input can inhibit, rather than favour, a specific result.
- **Output layer:** Output neurons represent predictable attribute values for the data mining model.

Classification is performed on these 433 instances by using the Neural network. This is performed in the Rapid miner tool as shown in Figure. 3. The output of this network is shown in Figure. 4.

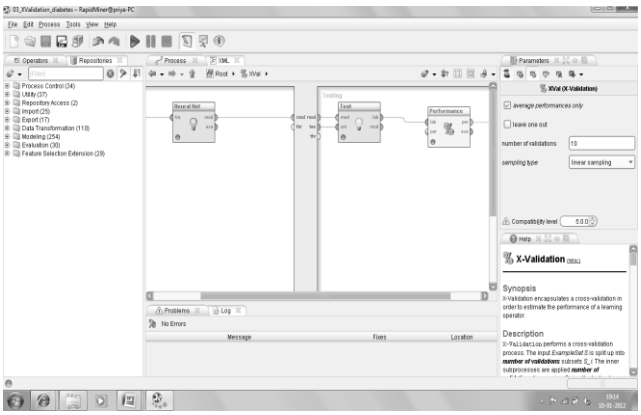

Figure 3. Classification using Neural Network

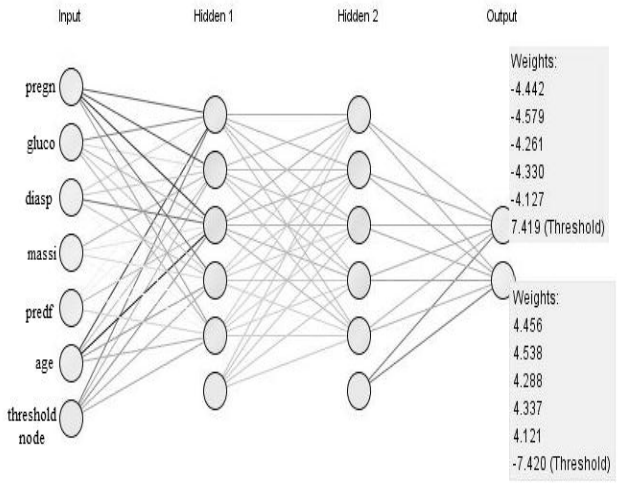

Figure 4. Output of the Neural network

The parameters considered in building this neural network are shown in Table 2. Since it can represent an arbitrary decision boundary the number of hidden layer is set to 2.

Table 2 Parameters used

No. of hidden layers	2
Training cycles	500
Learning rate	0.3
Momentum	0.2

The dataset collected from the hospitals are preprocessed and clustered into two groups. A pattern is extracted and a classifier model is constructed using this extracted data. Finally the performance is measured using the accuracy, sensitivity and specificity. The framework is shown in Figure. 5.

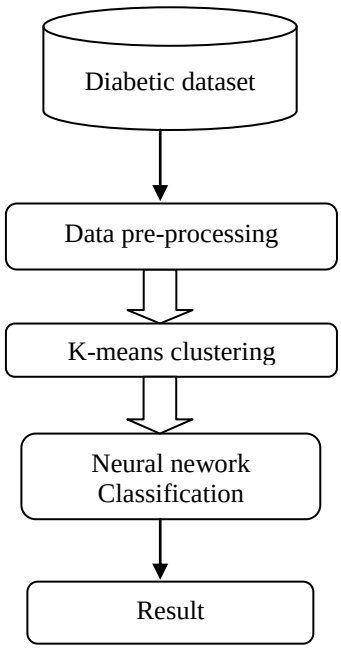

Figure 5. The frame work of the improved data mining model

4. PERFORMANCE EVALUATION

The performance of this model is measured using the following metrics:

4.1. Accuracy, Sensitivity, Specificity

For a single prediction there are possibilities of four outcomes namely: true positive, true negative, false positive and false negative. With these values the accuracy, sensitivity and specificity can be calculated. The accuracy as given in Eq. (2) determines how effectively the model predicts the development of Type-2 Diabetes for a newly diagnosed patient. Sensitivity as given in Eq. (3) identifies the actual positives i.e. correctly identifying the diabetic patients. Specificity as given in Eq. (4) identifies the actual negatives i.e. correctly identifying the non-diabetic patients. The equations are given by:

$$\text{Accuracy} = TP+TN/\ (TP+TN+FP+FN) \quad (2)$$
$$\text{Sensitivity} = TP/\ (TP+FN) \quad (3)$$
$$\text{Specificity} = TN/\ (TN+FP) \quad (4)$$

where

- TP is True Positive: Diabetic patients correctly diagnosed as Diabetic
- FP is False Positive: Healthy people incorrectly identified as Diabetic
- TN is True Negative: Healthy people correctly identified as healthy
- FN is False Negative: Diabetic patients incorrectly identified as healthy.

4.2. k- Fold Cross-Validation

K-fold cross-validation divides the data set into k subgroups. Each subgroup is tested with the classification rule generated from the remaining (k-1) subgroups. K different test results are obtained for each train-test configuration. The average result

gives the accuracy of the algorithm. 10-fold cross-validation is used in this model.

4.3. Kappa statistics

Kappa statistics gives an agreement between two different observers on the same data for qualitative items. If the value is 1, then there exists a complete agreement between the observers. The kappa value Eq. (5) can be calculated from the following equation

$$K=[P(A)-P(E)]/[1-P(E)] \quad (5)$$

$$P(A)=(TP+TN)/N \quad (6)$$

$$P(E)=[(TP+FN)*(TP+FP)*(TN+FN)]/N^2 \quad (7)$$

where N is the total number of instances used. P(A) is the percentage of agreement between the classifier and underlying truth calculated by Eq. (6). P(E) is the chance of agreement calculated by Eq. (7).

4.4. Confusion Matrix

Confusion matrix is the visualization tool used in the supervised learning. It is a matrix with each column representing the predicted class and each row representing the actual class.

5. EXPERIMENTAL RESULTS

An improved data mining model is built for the prediction of Type-2 Diabetes using the Weka and Rapid miner. Weka (Waikato Environment for Knowledge Analysis) is an open source machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

RapidMiner, formerly YALE (Yet Another Learning Environment), is an open source environment for machine learning**,** data mining**,** text mining and predictive analytics**.**
It is used for research, education, training, rapid prototyping, application development, and industrial applications. It supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

The result produced by this model is higher than the other models since it performs classification using Neural networks in the Rapid miner tool. This has produced an improvement in the accuracy when compared to the existing work as shown in Table 2 and the confusion matrix is shown in Table 3.

Table 3. Result of data classification

	C4.5 Classifier	**Neural network**
Accuracy	92.38%	97.93%
Sensitivity	90.38%	96.32%
Specificity	93.29%	98.65%
Kappa statistics	0.8249	0.953

Table 4. Confusion matrix

a	b	Classified as
124	11	a = one
10	288	b = zero

6. CONCLUSION AND FUTURE WORK

Predictive data mining plays a major role in extracting hidden information in medicine. This model is built for the Pima Indian Diabetes dataset and has produced an improved accuracy of 97.93%. The prediction method can be improved by other such classification techniques. This can also be implemented in the real dataset with some additional features. With this the risks related to diabetes can also be predicted.

7. REFERENCES

[1] Jianchao Han, Juan C.Rodriguze, Mohsen Beheshti, "Diabetes Data Analysis and Prediction Model Discovery Using Rapid miner", IEEE, Second International Conference on Future Generation Communication and Networking,pp.96-99,2008.

[2] Asma A. AlJarullah, "Decision tree discovery for the diagnosis of type-2 Diabetes", IEEE, 2011 Internaional Conference on innovations in information technology, pp. 303-307, 2011.

[3] B.M. Patil, R.C. Joshi, Durga Toshniwal, "Hybrid Prediction Model for Type-2 Diabetic patients", Expert Systems with Applications, Science direct, pp. 8102-8108,2010.

[4] "Diabetes Mellitus: a guide to patient care", Williams L and Wilkins, Philadelphia, 2007.

[5] effectsofdiabetes **:**http://www.effectsofdiabetes.org/

[6]http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data

[7] C.Yue, L. Xin, X.Kewen and S Chang, "An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO algorithm and WLS-SVM", IEEE, International Symposium on Intelligent Information Technology Application Workshops,2008.

[8] E.G. Yildirim, A. Karahoca and T. Ucar, 'Dosage planning for diabetes patients using data mining methods', Science Direct , Procedia Computer science, pp.1374-1380,2010.

[9] S.M Nuwangi, C.R. Oruthotaarachchi, J.M.P.P Tilakaratna and H.A. Caldera, 'Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes' IEEE, pp.372-377,2010.