Diabetic Prediction System Using Data Mining

L.H.S De Silva[1#], Nandana Pathirage[2] and T.M.K.K Jinasena[3]

[1,2]Department of Computer Science, General Sir John Kotelawala Defence University, Sri Lanka
[3]Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura
[#] L.H.S De Silva; <sanjayadesilva6@gmail.com>

Abstract— *Diabetes is one of deadliest diseases in the world. As per the existing system in Sri Lanka, patients have to visit a diagnostic center, consult their doctor and wait for a day or more to get their result. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. But with the rise of Machine Learning approaches, we have been able to find a solution to this problem using data mining. Data mining is one of the key areas of Machine learning. It plays a significant role in diabetes research because It has the ability to extract hidden knowledge from a huge amount of diabetes related data. The aim of this research is to develop a system which can predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treatment of the patients before it becomes critical. This research has focused on developing a system based on three classification methods namely, Decision Tree, Naïve Bayes and Support Vector Machine algorithms. Currently, the models give accuracies of 84.6667%, 76.6667%, and 77.3333% for Decision Tree, Naïve Bayes, and SMO Support Vector Machine respectively. These results have been verified using Receiver Operating Characteristic curves in a cost-sensitive manner. The developed ensemble method uses votes given by the other algorithms to produce the final result. This voting mechanism eliminates the algorithm dependent misclassifications. Results show a significant improvement of accuracy of ensemble method compares to other methods.*

Keywords— **Data Mining, Diabetes, Machine Learning**

I. INTRODUCTION

According to the World Health Organization (WHO), about 347 million people worldwide have diabetes. By the year of 2030, it has predicted to become the 7th leading reason for deaths in the world("WHO | Diabetes," n.d.). In 2012, diabetes was the direct reason for more than 1.5 million deaths in the world. The Diabetes Association of Sri Lanka (DASL) statistics reveals that by 2016, there are nearly four million diabetics in Sri Lanka(Thilakarathna, n.d.). According to DASL, almost one-fifth of the world's people with diabetes lives in the South-East Asia Region. More and more young people were being suffered by the disease and still the number of people affected by diabetes is increasing every day. Mainly there are three types of diabetes in the world(Ross, 2010).

Type 1- This results when the body fails to produce insulin. This form of diabetes was previously referred to as

"insulin-dependent diabetes mellitus" (IDDM) diabetes".

Type 2- This results because of insulin resistance, a condition in which cells fail to use insulin properly, sometimes also with an absolute insulin deficiency. This was earlier referred to as non-insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes".

Type 3- This is the 3rd main form of diabetes and it called as Gestational diabetes. It occurs when pregnant women without an earlier diagnosis of diabetes develop a high blood glucose level.

As we can realize from these facts, problems related to diabetes are many and quite costly. It is a very serious disease because, if not treated properly and on time, it could lead to very serious complications, may be the death of the patient. This makes diabetes one of the main priorities in medical science research. However, the country has not been utilized the strength of computer technology to reduce the risk of diabetes yet. With the rise of the new knowledge, scientists have discovered various kinds of new technologies that we could use to solve this problem. One of the most popular technologies today Data Mining. It is capable of predicting the risk level of a patient with significantly higher accuracy by extracting hidden patterns from historical medical records. This will help us to give timely treatment for patients by diagnosing disease early before it goes to a critical stage.

Data mining or the Knowledge Discovery in Data(KDD) is the process of exploration huge amount of data in order to discover new patterns or the trends. It is far beyond simple analytical techniques. Data mining uses many sophisticated machine-learning algorithms to discover hidden patterns in a large data set automatically. Later, such identified patterns can be used to predict future events. To do a proper diagnose on diabetic, doctors need to gather a huge amount of data about the patient. Obviously, it is harder for a human to analysis such a data volume manually. Using data-mining techniques for this is one of the best ways to enhance the accuracy and the efficiency of such process. Through both predictive (classification) and descriptive (clustering and association),data-mining techniques can be applied for this. The present study is focused on developing a diabetic

prediction system based on classification(predictive) data mining methods, namely Decision Tree algorithm, Naïve Bayes algorithm and SMO Support Vector Machine algorithms.

II. LITERATURE REVIEW

A systematic review of research findings and applications of data mining techniques in the field of diabetes has been done in order to identify the present status of the research question, the research gap, and the alternatives. Here our main objectives were to identify research goals, diabetes types, data-mining methods, data-mining software's and their technologies, data sets and outcomes. Based on that, we have developed a novel approach to predict diabetes using data mining technologies.

Huge amounts of data produced by healthcare transactions are too complex and voluminous to be processed and analysed by conventional methods (Cunningham and Holmes, 1999). However, the data mining is capable of extracting hidden knowledge from complex data repositories such as- research reports, flow charts, evidence tables and medical reports, and transform into useful information for decision making.

Breault and colleagues applied a "Classification and regression tree (CART) using the CART data-mining" software on data of 15,902 diabetes patients and detected that most important variable related to bad glycemic control (HbA1c >9.5) is age (Marinov et al., 2011). Patients below the threshold of 65.6 years old have worse glycemic control than older people, which was very surprising to clinicians. Using this knowledge, they have targeted the specific age groups that are more likely to have poor glycemic control. However, they have found age is the most valuable variable for glycemic control using the CART algorithm. There may be other important variables too. Thus, more methods have to be used to discover those.

Myiaki and colleagues (Mehrpoor et al., 2014) conducted a study to find the best predictors of diabetes vascular complications using CART on data from 165 type 2 diabetes mellitus (T2DM) patients. The authors found that age (cut-off: 65.4 years) was the best predictor, and depending on the age, the second best predictor was body weight (cut-off: 53.9kg) for the group above 65.4 or systolic blood pressure for the group below 65.4. Here they have gone more steps further.

Aiswarya and colleagues have done a research on "DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES"(Iyer et al., 2015). According them, diabetes has affected over 246 million individuals worldwide and most of them were women. According to the WHO report, by 2025, this variety is anticipated to rise to over 380 million. This paper focused on analyzing the patterns based Decision Tree and Naïve Bayes data mining algorithms for diabetes dataset. They have used 70:30 percentage split and 10-fold cross validation techniques to build their model. When using Decision Tree, they have got 76.9565% accuracy and for the Naïve Bayes, they have got 79.5652% accuracy. These results assure that classification data mining methods are better for prediction of diabetes. But there is no evidence that they have developed a system that can predict risk level of a patient in real time. They just have analysed those two algorithms using classifier models. But we have developed a system using classification data mining techniques which can diagnose the diabetic risk level of a patient.

Thirumal and Nagarajan have used Fuzzy, Neural Network, Case-Based (FNC) approach to predict rate of Diabetes. (Thirugnanam et al., 2012). They present a novel approach for the computational intelligence and knowledge engineering techniques as neural network (N), fuzzy logic (F), and case-based reasoning(C) as an individual approach (FCN). At the final prediction stage, they have applied the rule-based algorithm to the values obtained from the initial stage. They position as the benefits of applying that is the accuracy of predication rate is higher than other diabetes prediction algorithms. But using Neural network is somewhat slow because it does require more time to train the network. But this diabetic disease will be critical in some stages. So we have to find a quicker solution for this. That is one of the reasons we moved to classification based approach. Because there if we know the problem, we can get results quickly.

A Case Study on "UTILIZATION OF DATA MINING TECHNIQUES FOR DIAGNOSIS OF DIABETES MELLITUS" have done by Coimbatore Institute of Engineering and Technology (Thirumal and Nagarajan, 2006). This research has based on old diabetes patients. They have found that risk of diabetes will be low when patients are often given assessment and treatment plans that suit their wants and lifestyle. Straight forward awareness measures like low sugar diet, correct diet will avoid fatness. The Goal of this study was to urge best algorithms that describe given knowledge in multiple aspects. In this paper, several data mining algorithms have been used for test the dataset. Naïve Bayes, Decision trees, k nearest neighbor and SVM are discussed and tested with Pima Indian polygenic disease dataset. Accuracy of these models are needed to be evaluated before it is being used. If the available data are limited, it makes estimating accuracy a difficult task. Table 1 shows accuracies of the algorithms given by the confusion matrix.

Table 1 accuracies of algorithms

Algori-thm	Accuracy (%)	TP	FP	Preci-sion	Recall
Naïve bayes	77.8646	0.83	0.317	0.83	0.83
C4.5	78.2552	0.864	0.369	0.814	0.864

SVM	77.474	0.775	0.309	0.77	0.775
kNN	77.7344	0.892	0.437	0.792	0.892

From the experiments, it's complete that kNN provides lower accuracy when putting next to alternative algorithms because it stores training examples and delays the processing until a new instance is classified. The speed of the algorithm is also important when we decide the efficiency of an algorithm. Which tells that classification algorithms are better that these KNN algorithms when to compare with this problem domain. Here also they only have done an analysis using weka data mining tool whether what algorithm gives better results. But this is also a case study that focused on finding best data mining algorithms for diabetic related data. So considering these facts and previous diabetic related research, we have developed a system which gives a real-time prediction about whether the patient has diabetes or not.

III. METHODOLOGY

In previous studies, they have used only single approach to identify the disease. But we have combined three classification algorithms through a voting mechanism to increase the accuracy level of the model. So if one algorithm does not predict it correctly, it doesn't affect to the final prediction because the system considers the predictions of other two algorithms too. It gives the majorities decision. Thus ensures more accuracy than a single algorithm.

A. Decision Tree J48 Algorithm

A Decision tree is basically a tree structure(Han and Kamber, 2006), which has the form of a flowchart. It can be used as a method for classification and prediction with a representation using nodes and internodes. Root and internal nodes are the test cases. Leaf nodes considered as class variables. Figure 1 shows a sample decision tree structure.

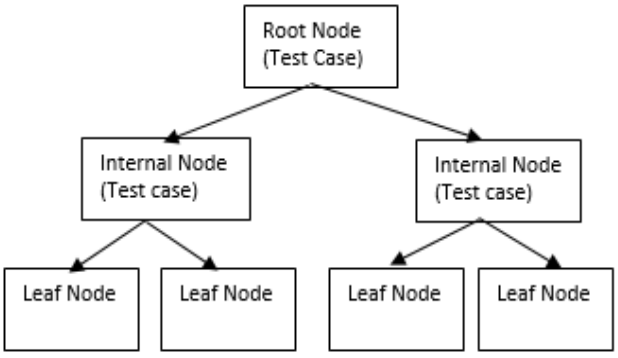

Figure 1: sample decision tree structure

Among classification data mining methods, decision tree algorithm provides powerful techniques for prediction. Among ID3, C4.5, C5, J48 and CHAIAD decision tree algorithms, we have selected J48 algorithm to develop our model. It's a java based algorithm, it works as follows. In order to classify a new item, it first creates a decision tree based on the attribute values of the available training data set. Every node of the decision tree is generated by calculating the highest information gain for all attributes. If any attribute gives an unambiguous end result (explicit classification of class attribute), the branch of that attribute will be terminated and then target value is assigned to it. We have used 12-fold cross validation technique to build the model using this algorithm. It's simply as follows.

- Break data into 12 sets of size n/12.
- Train on 11 datasets and test on 1.
- Repeat 12 times and take a mean accuracy.

In 12-fold cross-validation, the original sample is randomly partitioned into 12 equal sized subsamples. Of the 12 subsamples, a single subsample is retained as the validation data for test the model, and the remaining (12−1) subsamples are used as training data.

B. Naïve Bayes Algorithm

Naïve Bayes classifier algorithm has been created based on the Bayes rule of conditional probability. It uses all the attributes contained in the data, and then analyses them individually as though they are equally important and independent of each other. There are various data mining existing solutions exists to find relations between the diseases and their symptoms also the medications for them. But these algorithms have their own limitations like binning of the continuous arguments, numerous iterations, high computational time, etc. But Naïve Bayes classifier affords fast, highly scalable model building and scoring. The build process for Naive Bayes is parallelized. It overcomes various limitations like the omission of complex iterative estimations of the parameter because it can be applied to a large dataset in real time. The formula used for that algorithm is simply showed here.

$$P(C_K|x) = \frac{P(C_K)P(x|C_K)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

Here we have used 70:30 percentage split technique to build the model using Naïve Bayes algorithm. This means 70 percent of the data set have been used to train the data and other 30 percent of the data set have been used to test the model.

C. SMO (Sequential Minimal Optimization)

This algorithm is commonly used for solving the quadratic programming problems that arise during the training of SVM (Support Vector Machines). SMO uses heuristics to partition the training problem into smaller problems that

can be solved analytically. SMO algorithm it replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default which helps to speed up the training process. We have used 70:30 percentage split technique to train and test the data set using this model. Here we are not only considering the accuracy but it should have the ability to handle missing values well. This algorithm does that very accurately because it uses heuristics to partition the training problem into smaller problems. That's the main reason we have selected this algorithm.

III. Experimental design

This section explains the overall design of the system and what is the process it has followed in order to get the prediction.

D. Dataset Used:

The data set we have used is a benchmarked dataset which can be used for comparing the accuracy and the efficiency of our model. Data has been obtained from Pima Indians Diabetes Database, National Institute of Diabetes and Digestive and Kidney Diseases.

Number of Instances: 600
Number of Attributes: 8 + (1 class attribute).
For Each Attribute: (all numeric-valued).

1) Inputs:

- Number of times pregnant
- Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/ (height in m) ^2)
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)

Missing Attribute Values: None

Relabelled values in attribute 'class'

- From: 0 To: tested negative
- From: 1 To: tested positive

2) Outputs:

- Predicted Results (Diagnosed State)
- Evaluation Results
- Correctly Classified Instances
- Incorrectly Classified Instances
- Kappa statistic
- Mean absolute error
- Root mean squared error
- Relative absolute error
- Root relative squared error
- Total Number of Instances

E. Procedure:

- Load previous data sets to the system (768 test cases).
- Data pre-processing has done using integrating WEKA tool(Witten et al., 2011). Following operations are performed on the dataset after that.
  a. Replace Missing Values
  b. Normalization of values.
- Then User inputs data to the system in order to diagnose whether he has the disease or not.
- Build a model using J48 Decision Tree Algorithm and train the data set.
- Build a model using Naïve Bayes Algorithm and train the data set.
- Build a model using SMO Support Vector Machine Algorithm and train the data set.
- Test the data set using these three models.
- Get the evaluation results.
- Finally, get the predicted voting from all classifiers and gives the diagnostic result.

Following diagram shows the overall procedure of this system.

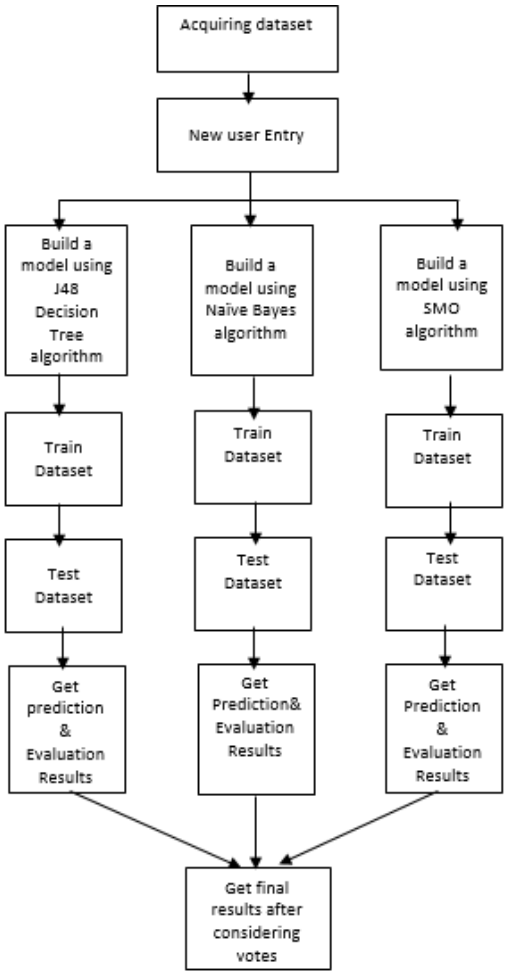

Figure 2: Overview of procedure

Number of times pregnant	8
Plasma glucose concentration 2 hours in an oral glucose tolerance test	183
Diastolic blood pressure (mm Hg)	64
Triceps skin fold thickness (mm)	0
2-Hour serum insulin (mu U/ml)	0
Body mass index (weight in kg/ (height in m) ^2)	23.3
Diabetes pedigree function	0.672
Age (years)	32
Class variable (0 or 1)	?

Followings are some of the user interfaces of developed system

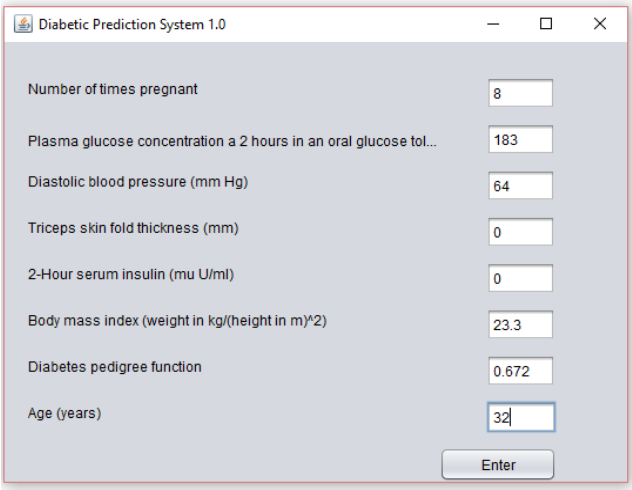

Figure 3 : User interface for the input attributes

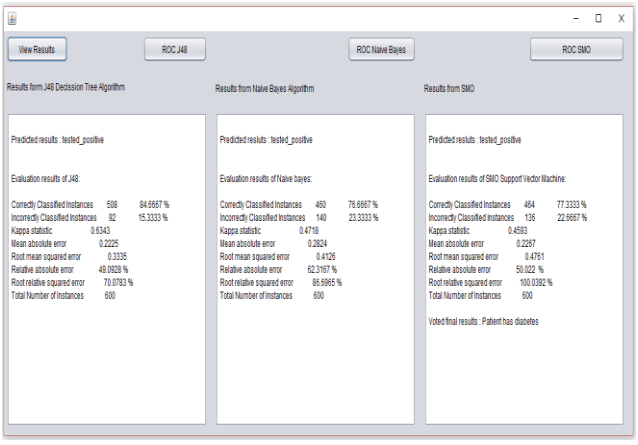

Figure 4: prediction and evaluation result interface

V. RESULTS AND DISCUSSION

Although it could produce different results for local data set due to varies differences such as DNA, foods they are eating or may be because of the life style of local people. We have chosen a benchmark diabetic dataset in order to compare our results with the other results.

Following are input readings from diabetic affected person.

You can see that this automated system also has diagnosed it correctly. You can see the diagnosed results from Figure 4 user interface also. It gives final result as this:

Voted final results: Patient has diabetes.

Voted final result is the final results generated after voting of all algorithms. If two or more models gives same diagnostic results it will identify that as the final result. That way the system ensures it always produces a correct iagnose. Because even though one algorithm blindly ives a wrong prediction, other two algorithms also check their results and prevents providing wrong diagnostic results.

We have produced evaluation results of this system after testing it using 600 diabetic records. Obtained results are as follows

Table 2: Evaluation results of three classifier models

Evaluation Results	J48 Decision Tree	Naïve Bayes	SMO Support Vector Machine
Predicted results	tested_positive	tested_positive	tested_positive
Correctly Classified Instances	508(84.6667 %)	460(76.6667%)	464(77.3333 %)
Incorrectly Classified Instances	92(15.3333 %)	140(23.3333%)	136(22.6667 %)
Kappa statistic	0.6343	0.4718	0.4593
Mean absolute error	0.2225	0.2824	0.2267
Root mean squared error	0.3335	0.4156	0.4761
Relative absolute error	49.0928%	62.3167%	50.022%
Root relative squared error	70.0783%	86.6965%	100.0392
Total Number of Instances	600	600	600

These results show that most higher accurate results are given by the J48 Decision Tree and SVM Support vector

machine algorithms. J48 has more than 84% accuracy and other two also have more than 76% accuracy. So it has more accuracy when comparing with most of other systems that have developed. Furthermore, because the voting process that we have used in this system, it ensures that it gives higher accurate results than when considering accuracies of the classifiers separately. Because it first considers all the diagnosed results of three classifiers and gives the final prediction results after that.

F. Confusion Matrix

Confusion Matrix has the information about predicted and actual classification results of the classifiers. The Performance of the classifiers has been evaluated using this matrix.

	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN

Table 3: Confusion Matrix

TP- Positive tuples that were correctly labelled by the classifier
TN-True Negative tuples that were correctly labelled by the classifier.
FP- False Positive tuples that were incorrectly labelled as positive.
FN- False Negative tuples that were mislabelled as negative.

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

But if your data set is heavily skewed, there is a problem in the confusion matrix. E.g. 90% of its instances are positive instances and 10% of instances are negative instances, in that case, any classifier you are producing that will be very bad. Because it doesn't know the difference between positive and negative instances. It just blindly tells everything as positive. So even though it has 90% of accuracy, it's just wrong because that is not reflecting the appropriate evaluation of your classifier. Because of that, we have used ROC (Receiver Operating Characteristic) curves in order to justify it has the correct accuracy. ROC curves are visualization tools which you can tell easily that in a cost sensitive manner whether your classifier is really appropriate or not. Because of that, we have generated ROC curves for each of our classifier models to test the accuracies in cost sensitive manner.

Here in Figure 6, 7 and 8 you can see the ROC curves generated for our J48 Decision Tree algorithm, Naïve Bayes and SVM Support Vector algorithm. Here X and Y axis are representing followings.

X axis: True positive rate
Y axis: False positive rate

You can see in all the ROC curves, that they are skewed to the True positive side. Which proves that our accuracies of the all three classifier models are high. Which conclude that our all three classifiers are appropriate ones and also have good accuracies.

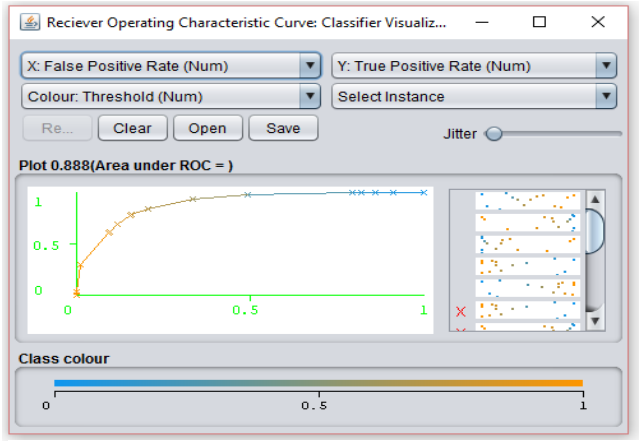

Figure 5: ROC Curve of J48 Decision Tree model

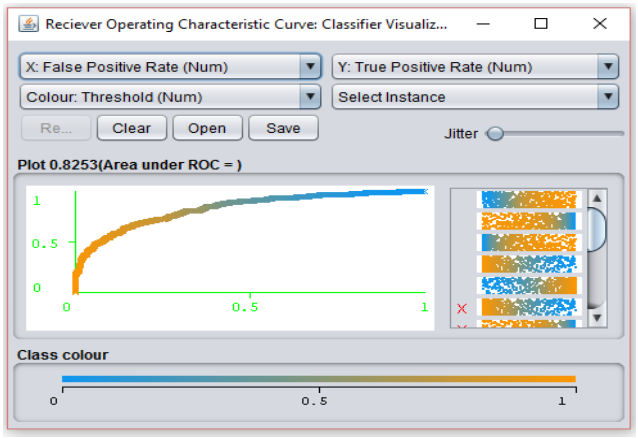

Figure 6: ROC Curve of Naïve Bayes mode

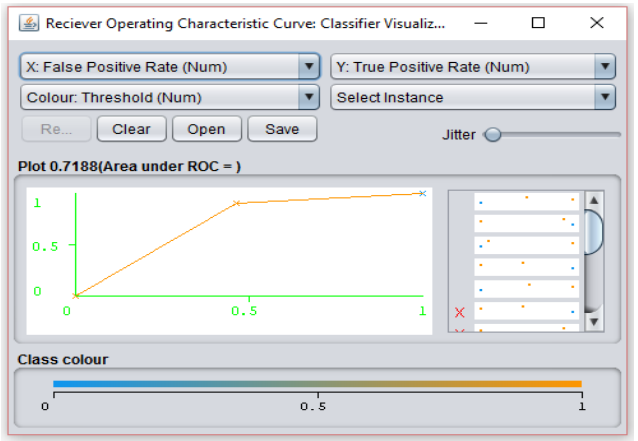

Figure 8: ROC Curve of SMO model

Furthermore, we have planned to gather more data from different locales over the country and develop more precise and general prescient model because increasing the data set also cause to increase the accuracy of the results. This data set that we have used is a benchmark dataset, which is gathered from other countries. It will be very helpful to build a precise model when we can use a data set from our country because the DNA patterns are different for every region.

V. CONCLUSION

Although all the methods have given more than 75% accuracy, the Decision Tree and the SMO Support Vector Machine give more accurate results than the Naïve Bayes algorithm. However, the ensemble method gives the highest accuracy from all due to the voting process of all the algorithms.

References

Cunningham, S.J., Holmes, G., 1999. Developing innovative applications in agriculture using data mining, in: The Proceedings of the Southeast Asia Regional Computer Confederation Conference.

Han, J., Kamber, M., 2006. Data mining: concepts and techniques, 2nd ed. ed, The Morgan Kaufmann series in data management systems. Elsevier ; Morgan Kaufmann, Amsterdam ; Boston : San Francisco, CA.

Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. Int. J. Data Min. Knowl. Manag. Process 5, 01–14. doi:10.5121/ijdkp.2015.5101

Marinov, M., Mosa, A.S.M., Yoo, I., Boren, S.A., 2011. Data-mining technologies for diabetes: a systematic review. J. Diabetes Sci. Technol. 5, 1549–1556.

Mehrpoor, G., Azimzadeh, M.M., Monfared, A., 2014. Data Mining: A Novel Outlook to Explore Knowledge in Health and Medical Sciences. Int. J. Travel Med. Glob. Health 2, 87–90.

Ross, T.J., 2010. Fuzzy logic with engineering applications, 3. ed. ed. Wiley, Chichester.

SureshikaThilakarathna, n.d. Nearly four Million Diabetics in Sri Lanka [WWW Document]. URL http://www.news.lk/news/business/item/5701-nearly-four-million-diabetics-in-sri-lanka (accessed 1.26.16).

Thirugnanam, M., Kumar, P., Srivatsan, S.V., Nerlesh, C.R., 2012. Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach. Procedia Eng. 38, 1709–1718. doi:10.1016/j.proeng.2012.06.208

Thirumal, P.C., Nagarajan, N., 2006. UTILIZATION OF DATA MINING TECHNIQUES FOR DIAGNOSIS OF DIABETES MELLITUS-A CASE STUDY.

WHO | Diabetes [WWW Document], n.d. URL http://www.who.int/mediacentre/factsheets/fs312/en/ (accessed 5.22.16).

Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: practical machine learning tools and techniques, 3rd ed. ed, Morgan Kaufmann series in data management systems. Morgan Kaufmann, Burlington, MA.