An Approach for Diabetes Detection using Data Mining Classification Techniques

Sonu Bala Garg[a], Ajay Kumar Mahajan[b] and T.S.Kamal[c]
[a]PhD Scholar, IKG Punjab Technical University, Jalandhar, Punjab, India
,[b]Associate Professor, Beant College of Engineering and Technology, Gurdaspur, Punjab, India
[c]Ex-Professor,PEC University of Technology, Chandigarh.
Email: sonugarg79@yahoo.com, ajaykm_20@yahoo.co.in, tsk1005@gmail.com

ABSTRACT

Disease diagnose by expert systems, is one of the areas where tools of data mining are establishing successful results. The aim of this paper is to discover solutions for diagnosing the disease by analyzing the patterns found in the data through techniques of data mining like classification analysis. Classification is a common technique used in data mining that utilizes a set of pre-classified examples for developing a model that can help in classifying the population of records at enormous amount. There are various techniques of classification that are used for analysis of biomedical data. These include Naive Bayes, Bayes Net, J48, SMO, and Random Forest. In this paper, the comparison of different classification algorithms using Weka has been shown. Also these techniques are used to find out which algorithm is most suitable. The best algorithm based on the Cross validation is SMO classifier with an accuracy of 77.34 % and has the lowest average error at 22.65 % compared to others. The best algorithm based on the Percentage split, Decision Table classifier with accuracy of 81.99 % and has the lowest average error at 18.00 % compared to others.

Keywords: - Data mining, Bioinformatics, Data mining techniques, Weka, Diabetes

1. INTRODUCTION

In digital age a lot of data is generated daily. Managing this substantial amount of raw data and obtaining useful information from it, is one of the largest challenges of the information age. Such data is being collected gradually and stored in the form of spreadsheets and databases. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Data mining has emerged out as a very important tool for retrieving useful information from the data.

The biological data has also outgrown to a large extent due to advanced and wide biological research and increased use of computers in health care systems. It has necessitated the use of data mining tools for information retrieval and drawing meaningful conclusions from such a large amount of biological data.

1.1 Data Mining & Knowledge Discovery from Data (KDD)

Throughout the world, the way information is communicated, acquired and stored has changed a lot. A lot of data is generated every day, in almost every field. This data can be useful only if it is properly stored, organized, accessed, analysed and interpreted. Information retrieval is the task of representing, storing, organizing, and accessing information items. With the extensive use of databases and the explosive growth in their sizes, there is a requirement for effective utilization of this enormous amount of data. This is where data mining comes in handy, as it scours the databases for extraction of hidden patterns, discovering hidden information, decision making and hypothesis testing. Additionally, the enhancement of the healthcare database management systems creates a huge number of medical databases. Creating knowledge and management of large amounts of heterogeneous data has become a major field of research, namely Data mining. Data Mining is a process of identifying novel, potentially useful, valid and ultimately understandable patterns in data [1]. Many people treat the data mining as a synonym for another popularly used term, Knowledge discovery from data or KDD. Others view data mining as simply an essential step in the process of knowledge discovery [2]. The general process of KDD has been presented in Figure 1.

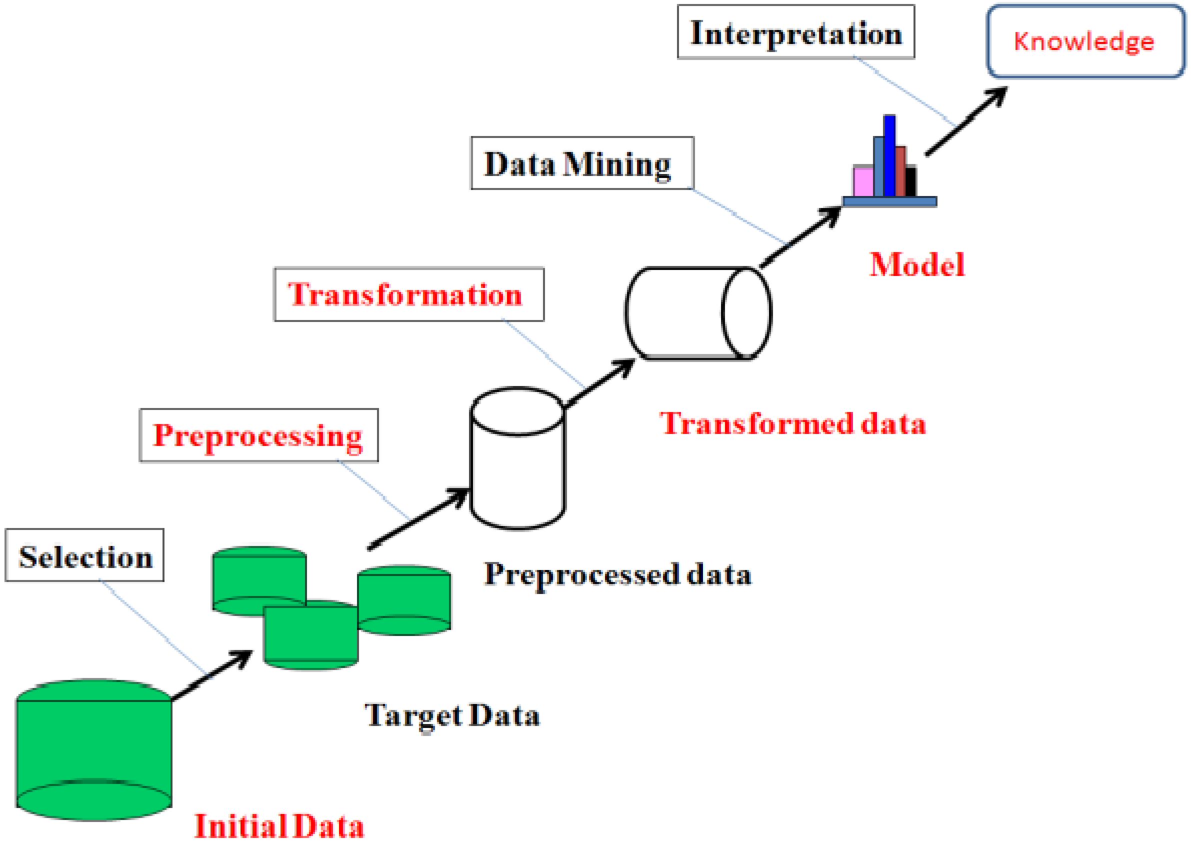

Fig.1 A general representation of KDD process [2]

Recent developments in data mining research have led to the development of various scalable and effective techniques for mining attractive patterns and knowledge in huge databases. These techniques range from effective techniques of classification to clustering; outlier analysis; frequent, sequential and structured pattern analysis method; and visualization and spatial/temporal data analysis tools.

1.2 Use of data mining in bioinformatics

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures [3].The use of data mining in bioinformatics can produce very successful results. There are many patterns in biology which are not clearly understandable and data mining assists in determining novel and hopefully useful information. Both bioinformatics and data mining are rapid growing research frontiers. It is essential to scrutinize what are the significant issues of research in bioinformatics and develop new techniques of data mining for effective and scalable bio-data analysis. Not applying data mining methods in research where the model is not known might miss essential discoveries. The techniques of data mining can be used efficiently in the bioinformatics field. It is considered that data mining will provide the tools which are necessary for better understanding of gene expression, drug design, and other emerging problems in genomics and proteomics [4]. These methods can be applied to discover associations among the genes, cluster similar gene and protein sequences and draw decision trees to classify the genes. Particularly it should be analysed as to how data mining may help effective and efficient bio-medical data analysis and summarize some research issues that may encourage the future enhancement of powerful tools of data mining for bio-data analysis [5].

1.3 Diabetes

Diabetes is a disease that occurs when the insulin production in the body is inadequate or the body is unable to use the produced insulin in a proper manner. The body cells break down the food into glucose and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the glucose that is produced by breaking down the food in the body cells. Any change in the production of insulin leads to an increase in blood sugar levels and this can lead to damage of the tissues and failure of the organs. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol /L) [6].

1.3.1 Types of diabetes

There are mainly three types of diabetes:

- **Type 1 diabetes** occurs when the pancreas stops or nearly stops producing the hormone insulin. People with type 1 diabetes must take daily insulin injections. Type 1 diabetes has also been referred to as insulin-dependent diabetes and juvenile diabetes.
- **Type 2 diabetes** occurs when the body is unable to make effective use of the insulin the pancreas does make. This is often referred to as insulin resistance. Obesity is a major cause of insulin resistance in both adults and children. Type 2 diabetes has also been called non-insulin dependent diabetes and adult-onset diabetes.
- **Gestational diabetes** occurs in women who have high blood sugar during pregnancy but have not been diagnosed with diabetes previously. After delivery of the baby, many women see their blood sugar return to normal. Some women will go on to develop type 2 diabetes.

All these types of diabetes are harmful to body and are needed to be treatment. If these can be detected at an early state, the complications associated with them can be avoided.

1.3.2 Symptoms of Diabetes

Following are the most common symptoms that are usually found to be present in the persons suffering from diabetes:-

- increased thirst and urination
- increased hunger
- fatigue
- blurred vision
- numbness or tingling in the feet or hands
- sores that do not heal
- unexplained weight loss

Urine test and blood tests are conducted to detect diabetes by checking excessive body glucose. The commonly conducted tests for determining whether a person has diabetes or not are:-

- A1C Test
- Fasting Plasma Glucose (FPG) Test
- Oral Glucose Tolerance Test (OGTT).

Though both Type 1 and Type 2 diabetes cannot be cured completely, however these can be controlled and treated by special diets, regular exercise, medication and use of insulin injections. The complications of the disease include neuropathy, foot amputations, glaucoma, cataracts, increased risk of kidney diseases and heart attack and stroke and many more. By the earlier diagnosis of diabetes, risk of the complications can be reduced. Therefore a faster method of predicting the disease has been worked upon and presented in this paper.

1.4 WEKA software

Waikato Environment for Knowledge Analysis (WEKA) is a Java based machine learning software. It is as important and easily usable data mining tool developed by University of Waikato, New Zealand. WEKA is available free of cost under the GNU General Public License. It incorporates various machine learning algorithms that can be employed for data mining tasks. It is portable and can be run on almost any modern computing platform. Furthermore, this software tool is very appropriate for various types of bioinformatics analysis. It consists of general purpose environment tools for data pre-processing, regression, classification, association rules, clustering, feature selection and visualization [2]. It helps in the comparison of distinct solution strategies based on the same method of evaluation and discovering the best strategy for problem solving at hand.

2. Classification Techniques

In WEKA, there are diverse kinds of classifiers such as Tree based classifiers, Bayes, functions, Rules etc. In this paper, different classification algorithms like Decision table, Naive Bayes, Bayes Net, J48, MLP, SMO, and Random Forest have been used to study diabetes dataset. A brief description of these techniques has been presented below. Their performance, accuracy and classification are also discussed.

2.1. Decision Table

In Weka, decision tree is a good tool for performing classification. This algorithm is divided into two parts, an inducer and a visualizer. In decision tree approach the data is divided hierarchically depending upon most suitable splitter in the given dataset. The inducer selects the most important attributes for splitting the data, and the associated visualizer displays the resulting model graphically [7].

2.2. Naive Bayes

The Naive Bayes Algorithm is a probabilistic algorithm and is based on Bayes theorem of posterior probability that is sequential in nature, following steps of execution, classification, estimation and prediction. Given the instance, the algorithm computes conditional probabilities of the classes and picks the class with the highest posterior. Naive Bayes classification assumes that attributes are independent. The probabilities for nominal attributes are estimated by counts, while continuous attributes are estimated by assuming all normal distribution for each attribute and class. Unknown attributes are simply skipped. Experimental studies suggest that Naive Bayes tends to learn more quickly than most induction algorithms. Therefore this algorithm was selected to compare the rate of learning. For finding relations between the diseases, symptoms and medications, there are several solutions that exist in data mining, but these algorithms hold restrictions, several iterations, binning of the constant opinion, high computational time, and can be applied on a huge dataset in valid time. The algorithm works on the simple Naive Bayes formula as shown in below [6].

$$Posterior\ Probability\ P\ (clx) = \frac{Likeli\ hood\ \ P\ (xlc\,)\ X\ Class\ \ Prior\ \ Probability\ \ \ P\ (c)}{Predictor\ \ \ Prior\ \ Probability\ \ \ P(x)}$$

2.3 Bayesian Networks

Bayesian networks are very attractive option for medical diagnostic systems as these can be useful to make assumptions, in cases the input data is incomplete. These are also called belief networks. A Bayesian network represents probabilistic association among a set of variable features and which consists of two components. The first component is directed acyclic graph (DAG). In DAG the nodes are called random variables and the edges between them represent the probabilistic dependencies between the equivalent corresponding random variables. A set of parameter is the second component which describes the conditional probability of each variable. By using the statistical and computational techniques, the conditional dependencies are estimated [8].

2.4 J48 Tree

J48 examines the normalized information gain that results from selecting an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then, the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then, a leaf node is created in the decision tree telling to choose that class. But it also happens that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs [9].

2.5 Multilayer Perceptron

The single-layer perceptron can only classify linearly separable problems. For non-separable problems it is necessary to use more layers. A Multilayer or feed forward network has one or more hidden layers whose neurons are called hidden neurons. The Fig.2 illustrates a multilayer network with one input layer, one hidden layer and one output layer [10].

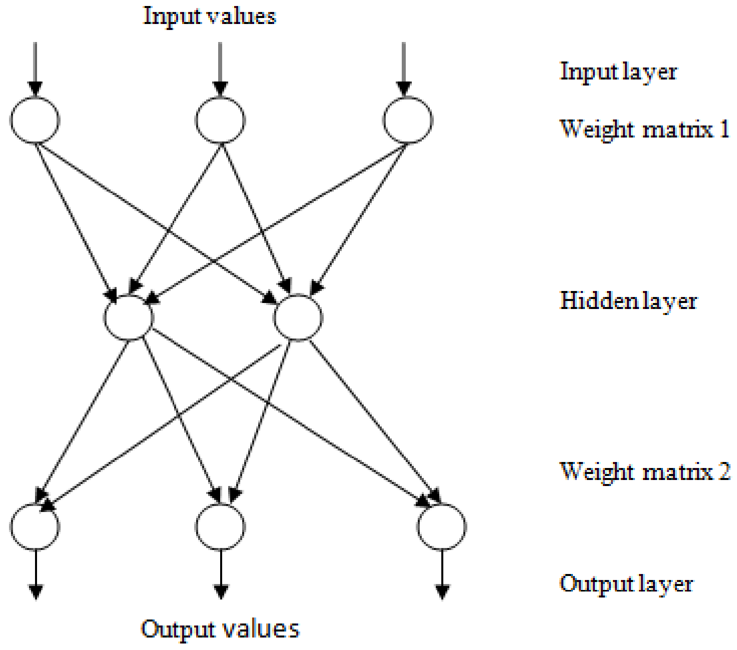

Fig.2 Multilayer Perceptron [10]

2.6 SMO

This algorithm is mainly used for solving the optimization problems. It divides the problem into a number of sub-problems, which are then solved analytically. It is used for training support vector machines. In Support Vector Machine, hyperplane is constructed that divides the data into separate classes. Then the hyperplane that is having the largest distance to the nearest training data point is selected as the best hyperplane [7].

2.7 Random Forest

It is an ensemble learning method for classification. In this, at training time a large number of decision trees are constructed and then the mode of classes of individual trees is calculated and used as the output class. It generally enhances the performance of the final model [7].

3. METHODOLOGY

3.1. Dataset Description and Pre-Processing

The data used in this study have been studied from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases having 768 instances and 9 attributes, as shown in table 1 & 2 below. Weka toolkit has been used for experimentation of different data mining algorithms. Experiments were performed using libraries from Weka machine learning environment. The Weka is an ensemble of tools for data classification, regression, clustering, association rules and visualization. As a data mining tool WEKA version 3.8 was utilized to evaluate the performance and effectiveness of the Diabetes database built from several techniques because WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models.

This paper deals with the Data Mining techniques to find out diabetes in women. The main aim is to predict if the patient has been affected by diabetes utilizing the tools of data mining by using the available medical data.

Table 1 Dataset Description

Dataset	No. of Attributes	No. of instances
Pima Indians Diabetes Database	9	768

Table 2 Attribute descriptions

S. No.	Attribute	Labelled values
1	Number of times pregnant	Preg
2	Plasma glucose concentration	Plas
3	Diastolic blood pressure in mm Hg	Pres
4	Triceps skin fold thickness in mm	Skin

5	2-Hour serum insulin	Insu
6	Body mass index in kg/m	Mass
7	Diabetes pedigree function	Pedi
8	Age in years	Age
9	Class Variable(0 or 1)	Class

4. Results and Discussion

This section summarizes the results of our experiments. In this section, the final data set was described and results of modeling from classification were provided. After that, 10-fold cross validation and percentage split for all the classifiers were performed in the following subsections. The results of experiments have been presented in Table 3 and shown in Figure 3, Figure 4, Figure 5 and Figure 6. Experiments have been carried out to evaluate the performance and usefulness of different classification algorithms for predicting Diabetes patients.

4.1 Results obtained using 10 fold cross validation method

In this sub-section, various performance measures were obtained using 10 fold cross validation method. The same have been presented in Table 3.

Table 3 Ten Fold Cross validation

Technique	**Accuracy**	**Kappa statistics**	**MAE**	**RMSE**	**RAE**	**RRSE**	**Error**
Decision table	71.22%	0.3492	0.3448	0.4277	75.85	89.72	28.77
Naïve Bayes	76.30%	0.466	0.284	0.416	65.50	87.43	23.6979
Bayes Net	74.34%	0.429	0.2987	0.4208	65.711	88.28	25.651
J48	73.82 %	0.4164	0.3158	0.4463	69.4841	93.6293	26.1719
MLP	75.39%	0.4484	0.2955	0.4215	65.013	88.427	24.6094
SMO	77.34%	0.4734	0.3094	0.3954	68.081	82.9651	22.7865
Random Forest	75.78%	0.4682	0.2266	0.476	49.848	99.862	22.6563

The results obtained have been analyzed below:

4.1.1 Classification Accuracy

Fig 3 shows the graphical representation of accuracy obtained by various classifiers on different datasets.

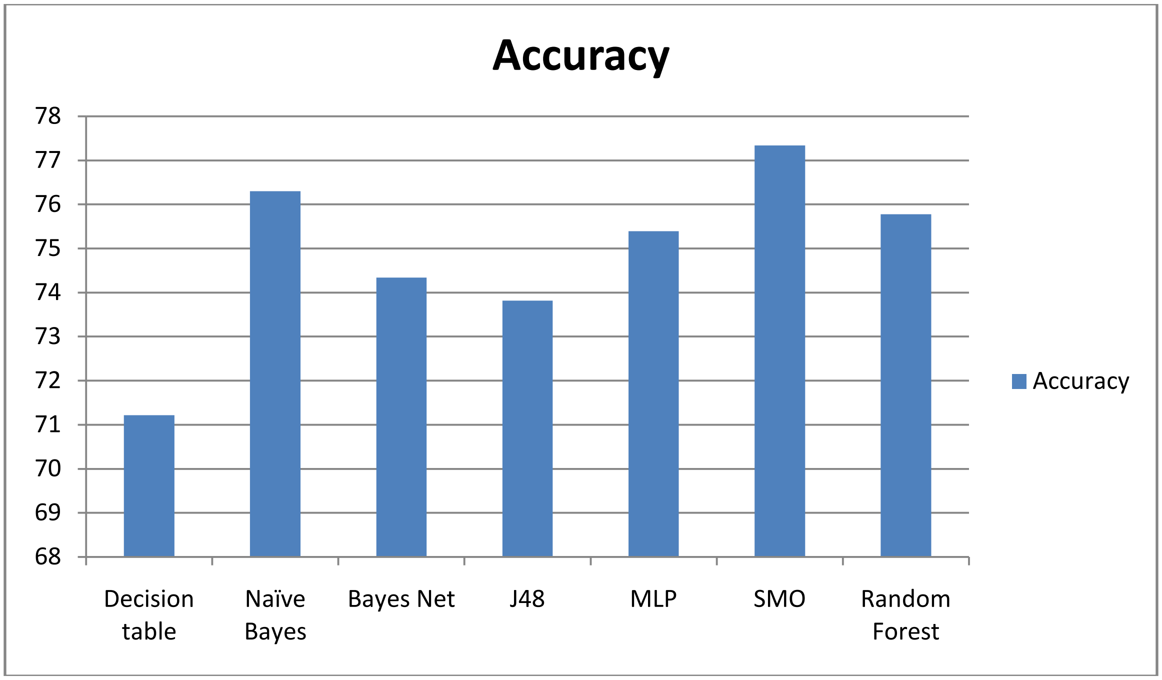

Fig. 3 Graphical representation of Accuracy obtained by various classifiers on different datasets

From Fig. 3 it can be seen that SMO algorithm gives highest classification accuracy i.e. (77.34%) and Decision Table algorithm gives least classification accuracy i.e.(71.22%). From the above facts it can be concluded that SMO performs best in case of diabetes dataset.

4.1.2 Kappa Statistics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

Fig. 4 shows the Graphical representation of Kappa statistics, Mean Absolute Error and Root Mean squared Error of various classifiers on different datasets.

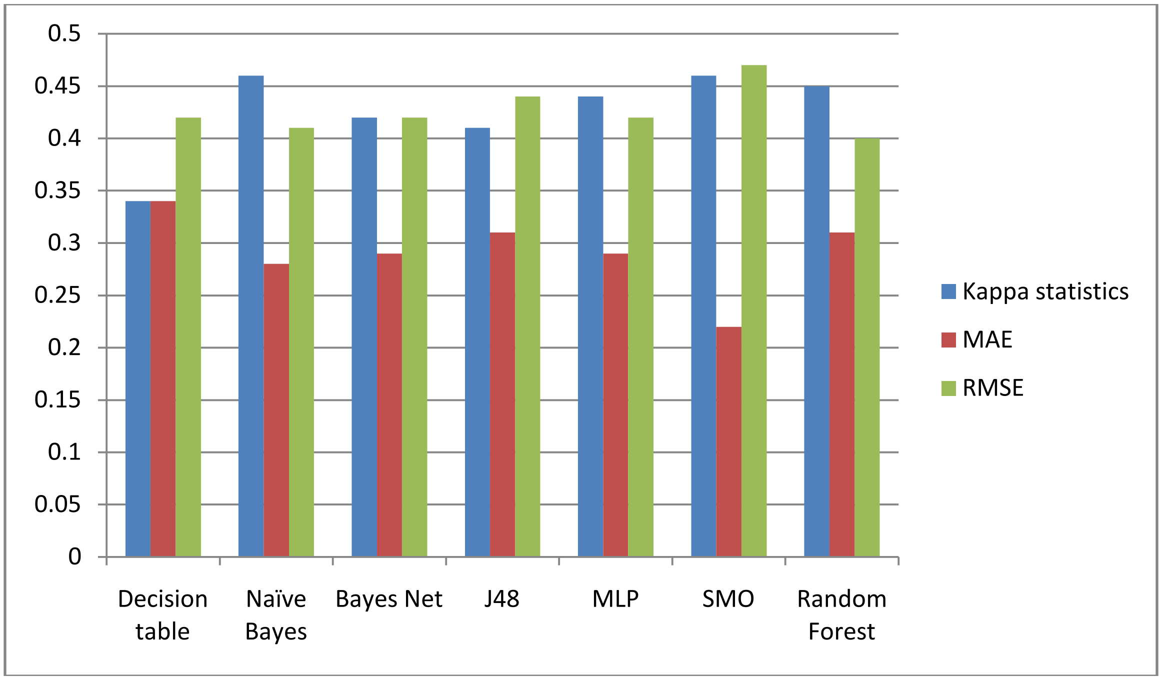

Fig.4 Graphical representation of Kappa statistics and error rates of various classifiers on different datasets

It can be observed from fig.4 that SMO classifier has largest value of K (0.4734) as compared to other classifiers of diabetes dataset.

Mean Absolute Error (MAE)

According to Table 3, Decision Table classifier achieves maximum MAE values i.e. (0.3448) and Random Forest classifier has minimum MAE values (0.2266).

Root Mean Squared Error (RMSE)

The result shows that SMO has the lowest error rate i.e. (0.3954) as compared to other classifiers of diabetes dataset.

4.1.3 RAE (Random Absolute Error) and RRSE (Root Relative Squared Error)

Fig.5 shows the Graphical representation of RAE and RRSE of various Classifiers on different datasets

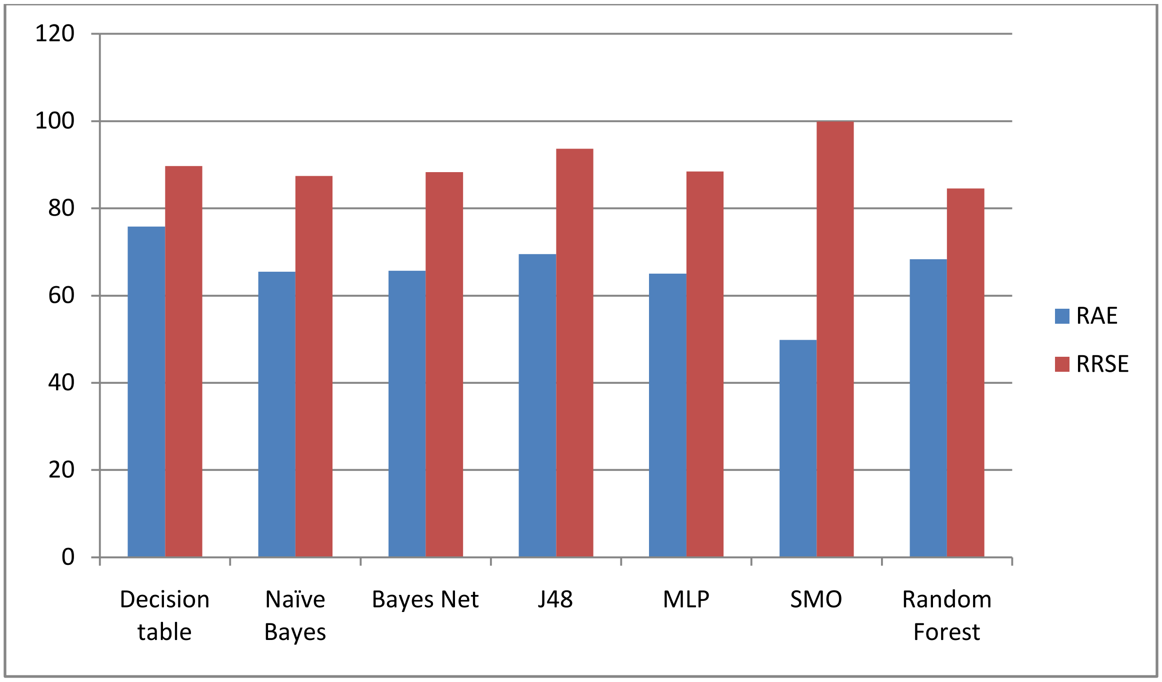

Fig. 5 Graphical representation of RAE and RRSE of various classifiers on different datasets

RAE (Random Absolute Error)

The result shows that Random Forest has the lowest error rate i.e. (49.848)

RRSE (Root Relative Squared Error)

The result shows that SMO has the lowest error rate i.e. (82.9651)

4.1.4 Error

Fig.6 shows the Graphical representation of Error of various Classifiers on different datasets

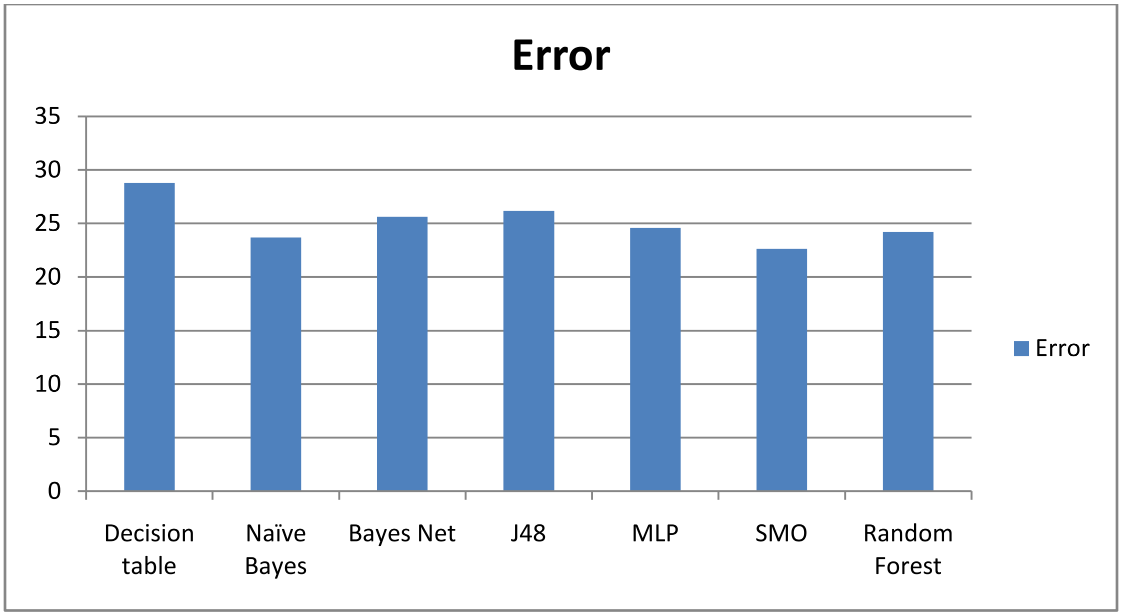

Fig.6 Graphical representation of Error of various classifiers on different datasets

The result shows that SMO has the lowest error rate i.e. (22.6563)

Depending on above parameters performance of the algorithms are compared and shown in Fig.3, Fig.4, and Fig.5 & Fig.6. It is observed from the figure that SMO gives near accurate results.

4.2 Results obtained based on Percentage split method

The various performance measures obtained based on Percentage split are presented in table 4.

Table 4 Based on Percentage split (66:34)

Technique	Accuracy	Kappa statistics	MAE	RMSE	RAE	RRSE	Error
Decision table	81.99%	0.5608	0.3064	0.3776	67.94	80.66	18.00
Naïve Bayes	77.01%	0.463	0.266	0.382	58.974	81.643	22.988
Bayes Net	78.16%	0.5220	0.280	0.378	62.15	80.85	22.839
J48	76.24 %	0.0.434	0.312	0.405	69.29	86.71	23.75
MLP	74.32%	0.4319	0.3186	0.4445	70.648	94.948	25.670
SMO	79.31%	0519	0.298	0.374	66.062	79.888	19.923
Random Forest	78.54%	0.490	0.206	0.454	45.873	97.169	20.68

The results obtained have been analyzed below:

4.2.1 Classification Accuracy

Fig.7 shows the Graphical representation of accuracy of various classifiers on different datasets

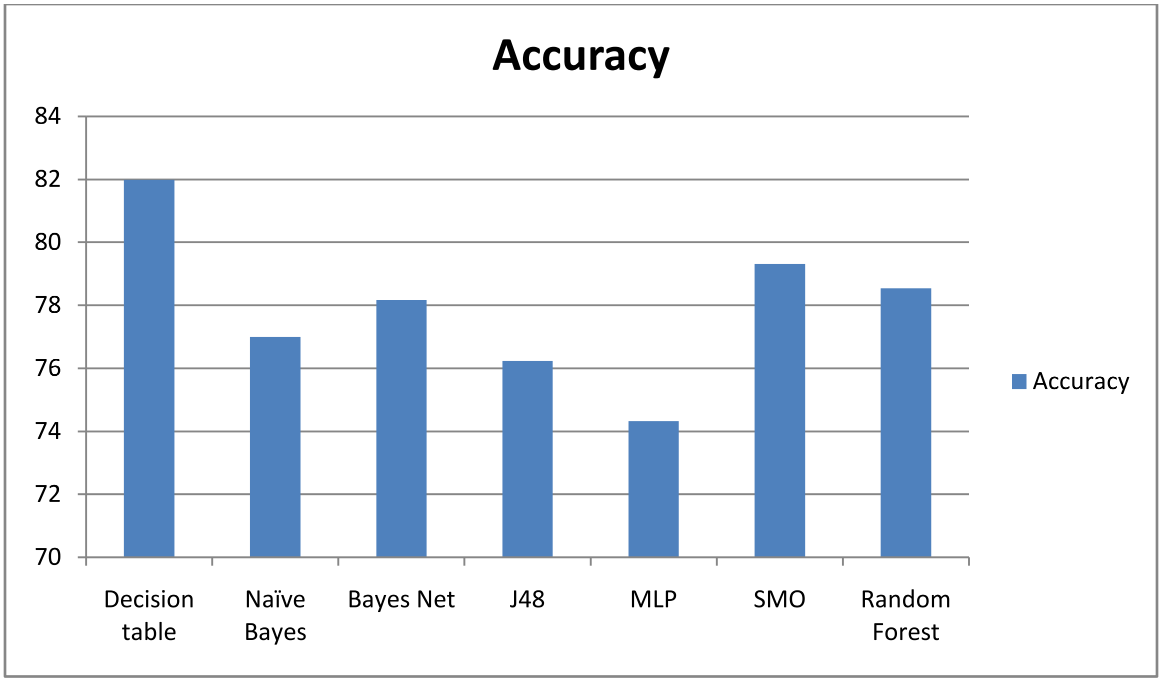

Fig.7 Graphical representation of accuracy of various classifiers on different datasets

Fig.7 shows that Decision Table gives highest classification accuracy i.e. (81.99 %) and MLP gives least classification accuracy i.e. (74.32 %). From the above facts it can be concluded that Decision Table performs best in case of diabetes dataset.

4.2.2 Kappa Statistics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

Fig. 8 shows the Graphical representation of Kappa statistics,Mean Absolute Error and root Mean Squared Error of various classifiers on different datasets

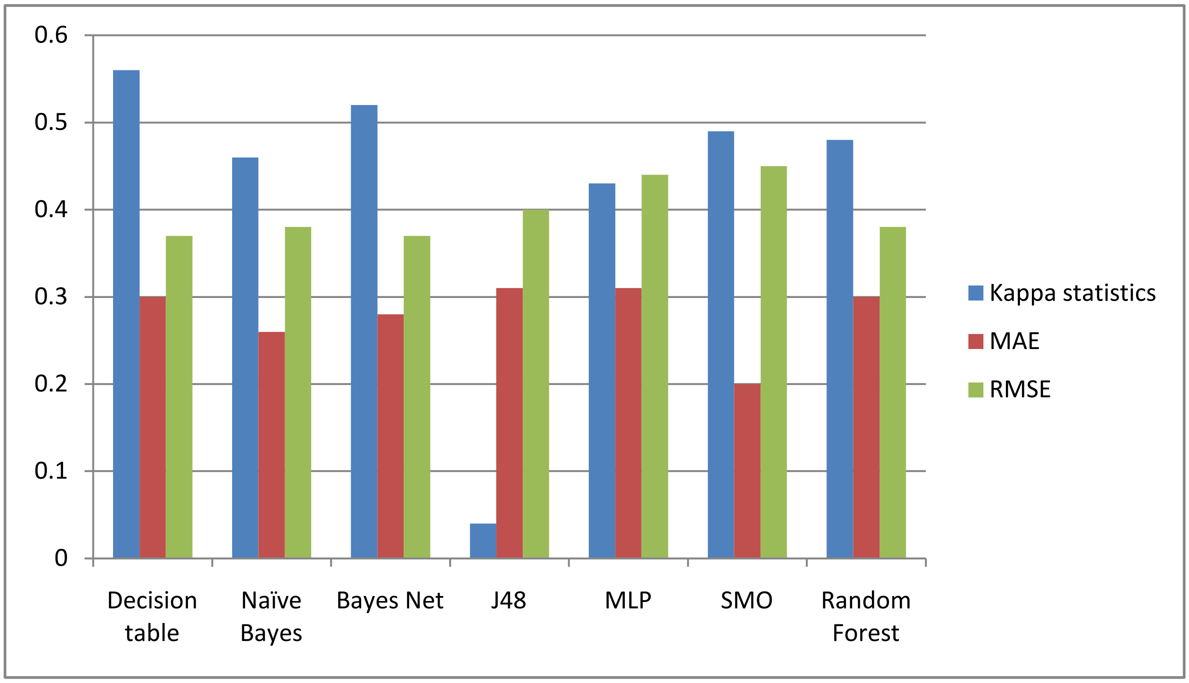

Fig. 8 Graphical representation of Kappa statistics and Error rate of various classifiers on different datasets

It can be observed from Fig.8 that Decision table classifier has largest value of K (0.5608) as compared to other classifiers of diabetes dataset.

Mean Absolute Error (MAE)

According to Fig.8, Random forest classifier has minimum MAE values (0.206).

Root Mean Squared Error (RMSE)

According to Fig.8, SMO has the lowest RMSE value i.e. (0.374).

4.2.3 RAE (Random Absolute Error) and RRSE (Root Relative Squared Error)

Fig.9 shows the Graphical representation of RAE and RRSE of various classifiers on different datasets.

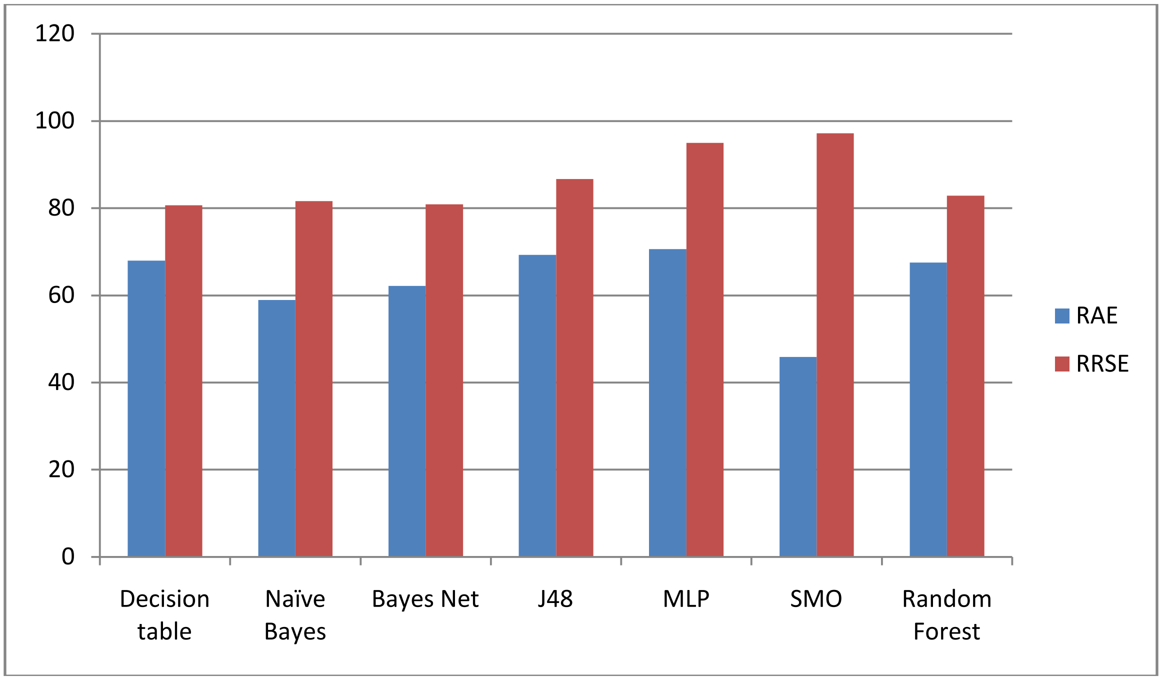

Fig.9 Graphical representation of Error rates of various Classifiers on different datasets

According to Fig.9, the result shows that Random forest has the lowest error rate i.e. (45.873).

RRSE (Root Relative Squared Error)

According to Fig. 9, SMO has the lowest error rate i.e. (79.888).

4.2.4 Error

Fig. 10 shows the Graphical representation of Error Rates of various classifiers on different datasets

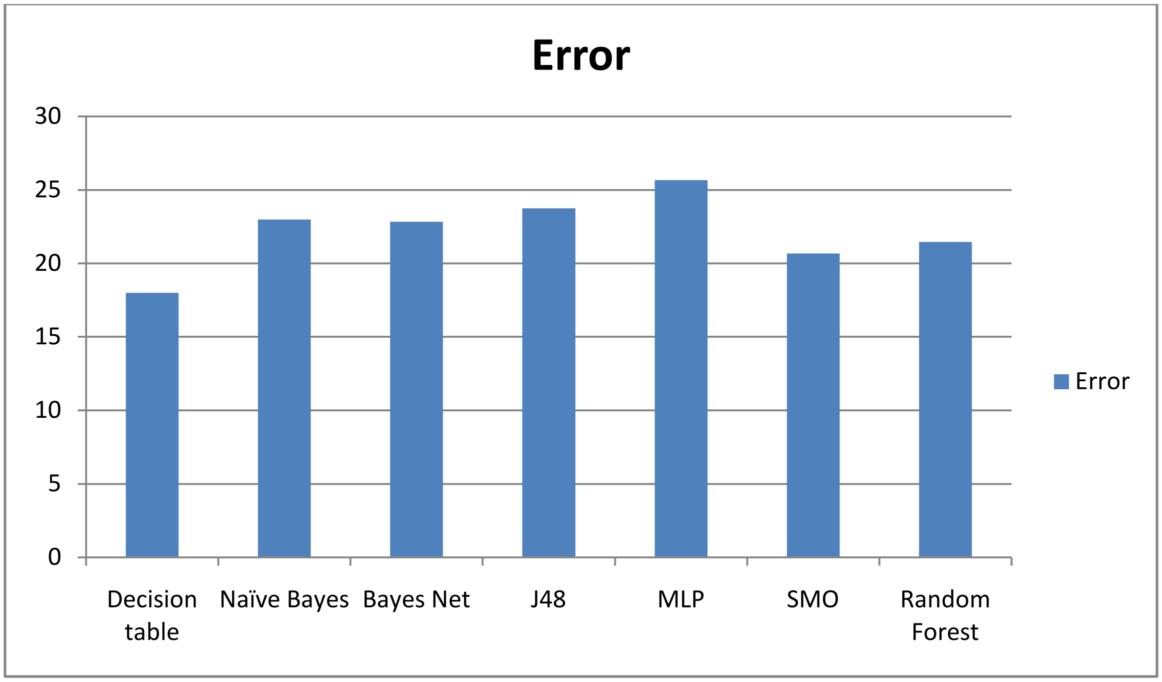

Fig. 10 Graphical representation of Error Rates of various classifiers on different datasets

According to Fig. 10, Decision Table has the lowest error rate i.e. (18.00).
It is observed from the figure that Decision Table gives near accurate results.
Depending on above parameters performance of the algorithms are compared and shown in Fig. 7, Fig.8, Fig.9 and Fig.10. It is observed from the figure that SMO gives near accurate results by using 10 fold cross-validation method and Decision Table gives accurate results by using Percentage split method for the same parameters.

5. Conclusion & Future work

In this paper, performance of various classifiers on the basis of accuracy, execution time, type of dataset and domain has been compared. The best algorithm based on the Cross validation is:

- SMO classifier with an accuracy of 77.34%.
- SMO classifier with the lowest average error at 22.65%.

These results proposed that among the tested machine learning algorithms, SMO classifier has the capability to considerably enhance the standard techniques of classification for use in the field of bioinformatics. The best algorithm based on the Percentage split is

- Decision Table classifier with an accuracy of 81.99%.
- Decision Table has the lowest average error at 18.00%.

  These results proposed that among the tested machine learning algorithm, Decision Table classifier has the capability to considerably enhance the standard techniques of classification for use in the field of bioinformatics. Therefore no particular algorithm is best suited for specific situation, the performance of the classification algorithms depends on the type and size of data sets, one algorithm is more appropriate for one data set while other algorithm is not appropriate for the same data set.

REFERENCES

[1] David Satish Kumar , Amr T.M Saeb, Khalid AI Rubeaan , “Comparative Analysis of Data Mining Tools and Classification Techniques using Weka in Medical Bioinformatics”, Computer Engineering and Intelligent Systems (2013) Vol. 4, No.3, pp. 28-38.

[2] Aher Sunita B, L.M.R.J LOBO, “Data Mining in Educational System using Weka,” International Conference on Emerging Technology Trends (2011) pp.20-25.

[3] Gangwar Vivek, Singh Yogendra, Ghose Udayan, “Data mining of Biological Data in Bioinformatics using Transcription ,translation Algorithm and Pattern Matching of Protein Sequences”, International Journal of Advanced Research In Computer Science (2012) Vol.3, No.3, pp.479-482.

[4] Mohammed J. Zaki, George Karypis, Jiong Yang, “Data Mining in Bioinformatics”, Algorithms for molecular biology (2007), Vol.2, No.4.

[5] Han J , “How can data mining help bio-data analysis”, 2nd International Conference on Data Mining in Bioinformatics (2002) Springer-Verlag, pp. 1-2.

[6] Lyer Aiswarya , Jeyalatha S, Sumbaly Ronak, “Diagnosis of diabetes using classification mining techniques”, International Journal of Data Mining & Knowledge Management Process (2015) Vol.5, No.1, pp.1-14.

[7] Bedi Rajni, Sharma Ajay Shiv, “Classification Algorithms for Prediction of Lumbar Spine Pathologies”, Springer, ICAICR (2017), pp. 42-50.

[8] Saini Nisha, Monica, Kumar Vijay, Kumbhar S, “Churn Prediction in Telecommunication Using Classification Techniques Based on Data Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, (2015) Vol. 5, No. 3.

[9] Salama I Gouda, Abdelhalim M. B, Zeid Magdy Abd-elghany “ Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers”, International journal of Computer and Information Technology (2012), Vol.1, No.1, pp.36-43.

[10] Amin Md. Nurul, Habib Md. Ahsan, “Comparison of Different Classification Techniques Using WEKA for Hematological Data”, American Journal of Engineering Research (2015) Vol. 4, No. 3, pp. 55-61.