Comparative Study of Data Mining Classifiers with Different Attributes and Different Databases Domain

P. Arumugam
Associate Professor,
Department of Statistics,
Annamalai University,
Chidambaram

Poompavai A.
Assistant Professor,
Department of Mathematics,
Apollo Arts and Science College, Chennai

Manimannan G.
Assistant Professor,
Department of Mathematics,
TMG College of Arts and Science, Chennai

R. Lakshmi Priya
Assistant Professor,
Department of Statistics,
Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai

ABSTRACT
In this paper, an attempt is made to identify and cross validate with five different classification methods in terms of precision, accuracy and kappa statistics calculated and visualized with different sets of database collected from different domain. This research paper has been implemented in R language environment and the obtained results show that which classifier is the most robust classifier method. The Accuracy based comparison of different classification for different datasets have been showed. By confusion matrix sensitivity, specificity, accuracy, true positive rate and false positive rate of different classifier for all four datasets are calculated and comparison of Kappa Statistics is also performed. The present work is about to analyze the effectiveness of the most popular classification techniques. According to the Experimental results, the Support Vector Machine model proved to have the best performance. It performed better of all datasets used. Naive Bayes Classifier, Decision Tree and Random Forest also performed well. The true positive rate and false positive rate table represent above 80% True Positive Rate and less than 20% False Positive Rate for all four datasets. Kappa Statistics basically performs the analysis between different classes. This shows the comparative analysis of different classification under the kappa statistics. Higher Value of kappa statistic is considered as good.

Keywords
Decision Tree, Random Forest, Naive Bayes Classifier, Linear Discriminant Analysis, Support Vector Machine, Confusion Matrix and Kappa Statistics.

1. INTRODUCTION
Data mining is a multi-billion dollar global market that is gaining popularity. Data mining is an inter-disciplinary field, which originated from statistics, data visualisation, data bases, and machine learning. There are many learning algorithms used in data mining – association rules, decision trees, neural networks, genetic algorithms, support vector machines etc. Anyone with a basic understanding of data visualisation techniques, statistics and computer science can easily get started with data mining. More important is an understanding of scales of measurement, data preparation and transformation techniques, data storage technologies (data bases and data warehouse), and Online Analytical Processing (OLAP).

2. DATA MINING
Data mining is the process of extracting hitherto unknown and potentially useful patterns, trends, anomalies and rules from stored historical data for business promotion, decision making or classification. Data mining is an inter-disciplinary field with roots in enterprise decision support. Exploratory Data Analysis (EDA) is a similar technique for summarising and identifying patterns in data. But EDA is often applied on small volume of data generated by sampling, direct observations or controlled measurements and analysed using purely statistical techniques.

The results obtained by a data mining process are used in marking business decisions and short-term predictions. It has diversified into many other fields that have no business context. For example, SVM is used to give a categorical label to unseen data instances using a model obtained from a set of labelled training data. It has more applications in business than in medicine, biology, genetics, etc., Similarly, genetic algorithms and neural networks are used for optimisation of empirically observed functions under constraints. Data mining is an iterative process in all fields to discover Knowledge Discovery Database (KDD).

Statisticians mostly analyzed systemically planned experiments to reply to a thoroughly formulated scientific question. These experiments lead to small amount of high quality data. Under these controlled conditions one could often derive an optimal way of collecting and analyzing the data and mathematically prove this property. The scale of data set has changed. Data are growing in two dimensions: they not only consist of more and more observations, they also contain more and more variables. Often these data are not directly sampled (for analysis)., but are merely by product of other activities. As such, they do not necessarily stem from good experimental design and some variable might contain no information. The data thus contains more and more 'noise'.

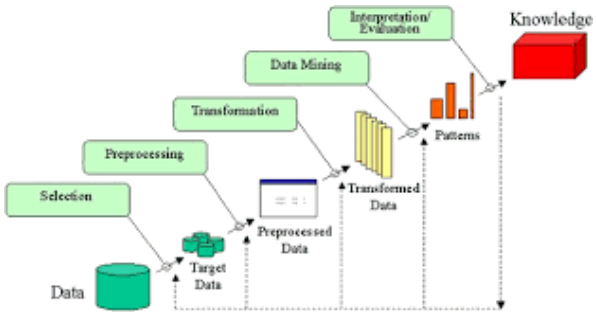


Figure 1. KDD of Data Mining Processes

Thus data mining differs from traditional statistics in several ways: formal statistical inference is assumption driven in the sense that a hypothesis is formed and validated against the data (Figure 1). Data mining in contrast is discovery driven in the sense that patterns and hypothesis are automatically extracted from data, another way, data mining is data driven,

while statistics is human driven. The branch of statistics that data mining resembles most is exploratory data analysis, although this field, like rest of statistics, has been focused on data sets far smaller than of the target of data mining researchers. Data mining also differs from traditional statistics in that sometimes the goal is to extract qualitative models which can easily be translated into logical rules or visual representations; in this sense data mining is human centered and is sometimes coupled with human-computer interfaces research (J. Han, M, Kamber and J Pei, 2012).

3. REVIEW OF LITERATURE

In recent days the amount of data stored in educational database is increasing fast. Many research scholar and scientist dealt with the classification of certain diseases using artificial neural network (ANN) and fuzzy equivalence relations. The heart rate variability is used as the base signal from which certain parameters are extracted and presented to the ANN for classification. The same data is also used for fuzzy equivalence classifier. These study of ANN and fuzzy classifier accuracy is nearing 85 percent to 90 percent. (U. Rajendra Acharya, P. Subbanna Bhat, S.S. Iyengar , Ashok Rao, Sumeet Dua (2003). The another scholar used Bayes classification for prediction model to identify the difference between high learners and slow learners student (Brijesh Kumar Bhardwaj, Saurabh Pal ,2011)

The application of data mining is highly noticeable in fields like e-business, marketing, text mining, linguistic studies, etc. and retail has led to its application in other industrials sectors. Among these subdivisions just discovering is healthcare. The healthcare surroundings is still „information wealthy, but knowledge very meagre. Healthcare data is available within the healthcare environment. This research paper propose to provide a survey of current techniques of knowledge discovery in databases using data mining techniques in Heart Disease Prediction. The researcher applied many data mining techniques like, KNN, Classification, Clustering methods, Deacons Trees and Bayesian Classification. All these techniques not perform well except Decision Tree classification and some time Bayesian classification is having similar accuracy as that of decision tree. (Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, 2011)

In other study the Text mining classification is the process of classifying documents into predefined categories based on their input content. The document split into two categories Training and testing documents. Text classification is primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data. Active supervised learning algorithms to automatically classify the text need sufficient documents to learn accurately Using Naïve Baye[s Classifier and Genetic algorithm. These two experimental results show that projected system works as a successful in the text document classifier. (S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan, 2010)

Three data mining classifiers like, Logistic Regression, SVM and Neural Network classifiers are considered for classification of performance analysis of different data mining classifiers before and after feature selection on binomial database. The Congressional Voting Records data set is a binomial data set investigated in this study is taken from UCI machine learning repository. The classification performance of all classifiers is presented by using statistical performance measures like accuracy, specificity and sensitivity. Gain chart and R.O.C (Receiver Operating Characteristics) chart are also used to measure the performances of classifiers. A comparative study is carried out among the data mining classifiers. Experimental result showed that without feature selection Logistic Regression and SVM classifiers provides 100 percent accuracy and neural network provides 98.13 percent accuracy on test data set. With feature selection SVM classifier provides 100% accuracy. The performance of SVM classifier is found to be the best among all classifiers with reduced number of features. (Pushpalata Pujari, 2013)

4. DATABASES

4.1 Dataset 1

The secondary database was collected from UCI website. The number of instances in this study is 650 and number of attributes are 32. The attributes used in this study are school, internet, romantic, address, gender, age, parent status, mother education, father education, travel time to school from home, study time, activities, health and absences. These data mining classification model were developed using R language. Initially dataset had 32 attributes. After attribute selection (internet, romantic, address, sex, age, Status, Medu, Fedu, traveltime, studytime, activities, health and absences) missing values records were identified and deleted from dataset. After deleting records with missing values, 649 were left out. On these 649 records data mining classification techniques such as Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis were applied.

4.2 Dataset 2

The data is a secondary data taken from DATA.GOV website. The number of instances in this study is 1565 and number of attributes are 13. The attributes used in this study are state, record test iodine, age, bmi, hb, fasting sugar. The data mining classification model were developed using R language. Initially dataset had 13 attributes. After attribute selection (state, area, age, record test iodine, bmi, hb, fasting sugar) missing values records were identified and deleted from dataset. After deleting records with missing values we were left with 1565 records. On these 1565 records data mining classification techniques such as Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis were applied.

4.3 Dataset 3

The data is a secondary data and taken from UCI website. The number of instances in this study is 4521 and number of attributes are 17. The attributes used in this study are age, job, marital status, education, default, housing, loan, contact, day, month, duration, campaign, poutcome and dependent variable. The data mining classification model were developed using R language. Initially dataset had 17 attributes. After attribute selection (age, job, marital status, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome and dependent variable) missing values records were identified and deleted from dataset. After deleting records with missing values we were left with 4522 records. On these 4522 records data mining classification techniques such as Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis were applied.

4.4 Dataset 4

The data is a secondary data and taken from UCI website. The number of instances in this study is 9910 and number of attributes are 15. The attributes used in this study are listing of

attributes, age, work class, Final weight, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week and native-country. The data mining classification model were developed using R language. Initially dataset had 15 attributes. After attribute selection (listing of attributes, age, work class, Final weight, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week and native-country), missing values records were identified and deleted from dataset. After deleting records with missing values we were left with full records. On these records data mining classification techniques like Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis were applied.

In this study, four datasets were considered which are from the UCI Repository and DATA.GOV. These datasets are effective enough to show classification process. These datasets are analysed under different classification parameters. The detailed descriptions of these datasets in terms of features and data points are given below.

Table 1 Description of the Four Databases

Sl. No	Dataset	Instances	Attributes
1	Dataset 1	650	32
2	Dataset 2	1565	13
3	Dataset 3	4521	17
4	Dataset 4	9910	15

Every dataset has different types of data, including numbers, text and other domain data points. Each of the dataset is explored explicitly due to their uniqueness in terms of their varying attributes, discrete or continuous nature of data etc. These datasets are analyzed for classification task by using R tool under different classification approaches (Table1).

R contains number of built-in data mining classification so that different mining operations can be performed directly. R is used by researches to analyze effectiveness of different algorithms. In this study, R tool is used to perform analytical study of classification on datasets.

5. METHODOLOGY

5.1 Decision Tree

A decision tree (DT) is a nonparametric classification and prediction model organised in the form of a rooted tree with two types of nodes called decision nodes and class nodes.

DT is a supervised data mining model, which originated in managerial decision theory, gambling, and theory of games. The input to a DT algorithm is the labelled training data and output is the hierarchical structure hidden in training data. An advantage of DT is that it decomposes a complex decision making problem into smaller manageable sub-problems (corresponding to each of the subtrees). Complex decision is based upon a large number of factors.

These factors are represented by simple binary digits (0 and 1), categorical variables, integers, reals, complex numbers or structured data types. Variables are categorical or quantitative in most data mining applications. In web and text mining, we also come across structured data. We will assume that the data are first captured into a flat file (without any hierarchical structure on it), with each row representing data about one subject. Attribute (columns) value can be comma separated or tab separated. The chosen format depends upon the software to be used for processing. The class labels (categories) into which samples get assigned should be known for training data. The classes must be mutually exclusive and collectively exhaustive. In other words, each item should belong unambiguously to a single class. The number of cases should be more than total classes. Data must be sufficient for a reasonable number of splits (Figure 2.).

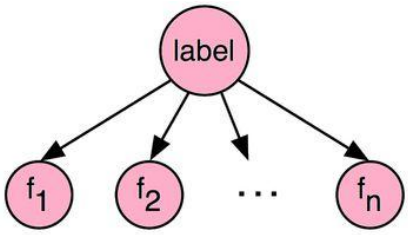


Figure 2. Decision Tree

If the node are numbered from top to bottom sequentially, we will denote the size of node i by s_i.

5.2 Random Forest

The random forests algorithm is a machine learning technique that is increasingly being used for image classification and creation of continuous variables such as percent tree cover International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences 2010 and forest biomass. Random forests are an ensemble model which means that it uses the results from many different models to calculate a response. In most cases the result from an ensemble model will be better than the result from any one of the individual models (Dahinden 2009). In case of random forests, several decision trees are created (grown) and response is calculated based on the outcome of all the decision trees.

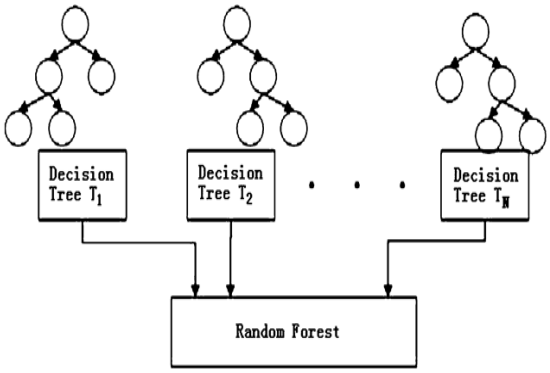


Figure 3. Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is mode of the classes (classification) or mean prediction (regression) of individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest comes at an expense of some loss of interpretability, but generally greatly boosts the performance of final model(Figure 3.).

5.3 Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods (Figure 4.).

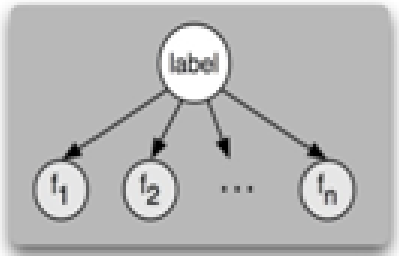
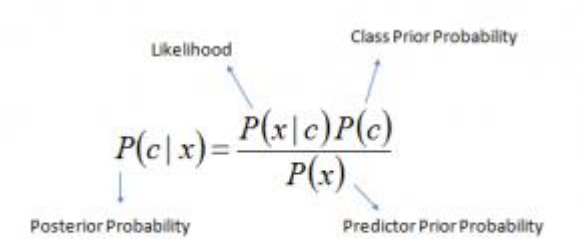


Figure 4. Random Forest Classifier

5.3.1 Naive Bayes Algorithm

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). The equation is:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where, P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

5.4 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (*supervised* learning), the algorithm outputs an optimal hyper plane which categorizes new examples.

There are many classifiers that originated in statistics. Examples, naive Bayes classifier, maximum entropy classifier, Fisher's discriminant classifier, partial least squares classifier, and Mahalanobis distance based classifier. In addition, multiple (linear and nonlinear) regression and logistic regression models can be used as classifier. Some of these classical models for pattern classification and prediction have assumptions on the data distributions. For instances, multiple regression models assume that error terms are normally distributed, and that independent variables are correlated. Similarly, normality is assumed in discriminant analysis, canonical correlation, etc. The Support Vector Machine (SVM) is a supervised classification model without any assumptions on the data distribution. Another name for SVM is kernel machines (as nonlinear SVM uses a kernel mapping). A machine learning algorithm tries to learn the relationship (X→y) from the training data X to the classes or categories y, so that it can be used to classify new data instances. It is used for pattern recognition (eg: face, retina, fingerprint and other images, handwritings and speech recognition), classification (eg: medical classification), clustering (web page and image clustering) and regression (SVR). There could exist multiple separating hyper plane when the number of data points is larger than the dimensionality (Figure 5.)y.

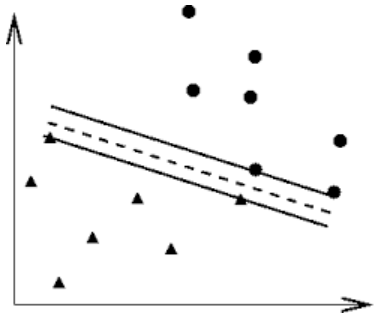


Figure 5. Support Vector Machine

5.5 Linear Discrminant Analysis

Originally developed in 1936 by R.A. Fisher, Discriminant Analysis is a classic method of classification that stood as the test of time. Discriminant analysis often produces models whose accuracy approache more complex modern methods. Discriminant analysis can be used only for classification (i.e., with a categorical target variable) and not for regression. The target variable may have two or more categories.

Discriminant analysis is a classification involving two target categories and two predictor variables. The following figure shows a plot of the two categories with the two predictors on orthogonal axes (Figure 6.):

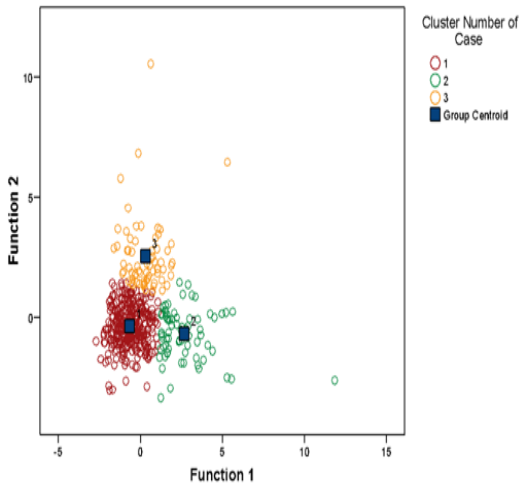


Figure 6. Discriminant Classification

Linear discriminant analysis finds a linear transformation of the two predictors, X and Y which yields a new set of transformed values that provides a more accurate discrimination than either predictor alone:

$$\text{Transformed Target} = C_1 * X + C_2 * Y$$

6. RESULT AND DISCUSSION

6.1 Dataset 1

6.1.1 Classification Tree:

In the dataset result established that the root node error: $\frac{226}{649} = 0.34823, and\ sample\ \ Size\ \ n = 649$

In classification tree, variables used in tree construction for the data are absences, activities, address, Fedu, internet, studytime, traveltime. The root node error is 0.34823.

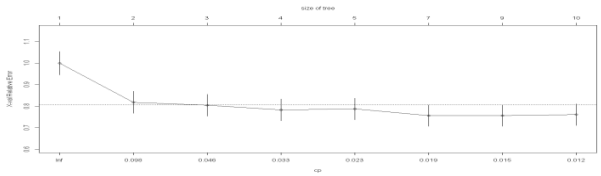


Figure 7. Classification Tree for Complexity Parameter (CP)

The above plot representing the size of the tree and complexity parameter (CP) value is 0.046. (Figure 7)

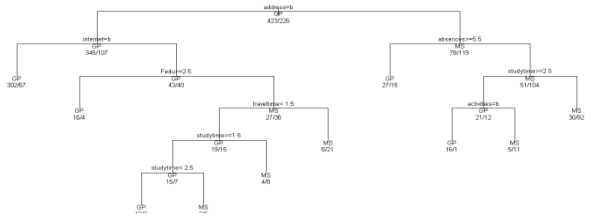


Figure 8. Classification tree for Dataset1

Dataset1 based on classification when considering decision tree, the tree construction represents the address, with internet and absences. On further classification, it has been grouped based on father education, travel time, study time and activities (Figure 8).

Table 2 Comparison of Data Mining Models

Model	Sensitivity	Specificity	Accuracy
Decision Tree	91.25%	46.02%	75.5%
Random Forest	87.47%	51.33%	74.88%
Naive Bayes	81.32%	63.27%	75.04%
SVM	93.14%	56.19%	80.28%
LDA	87.71%	56.64%	76.89%

In this dataset, the researcher compared five data mining classifier based on their sensitivity, specificity and accuracy. It shows that SVM classifier has better classification precision compared with other classifier.

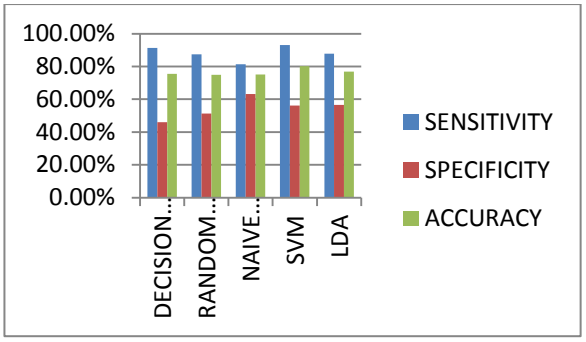


Figure 9. Graphical representations of sensitivity, specificity and accuracy

Table 3 and Figure 9, shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis.

Table 3. True positive rate and false positive rate

Models	True Positive Rate	False Positive Rate
Decision Tree	0.9125	0.0875
Random Forest	0.8747	0.1253
Naive Bayes	0.8132	0.1868
SVM	0.9314	0.0686
LDA	0.8771	0.1229

The results show that SVM outperforms well than Decision Tree, Random Forest, Naive Bayes models, parameters Sensitivity, Specificity, Accuracy and Error Rates.

6.2 Dataset 2

6.2.1 Classification Tree:

In the dataset result established that the root node error $\frac{621}{1565} = 0.39680, and\ sample\ \ Size\ \ n = 1565$

In classification tree, the variables used in tree construction for the data are age, area, record test iodine and state. The root node error is 0.39936.

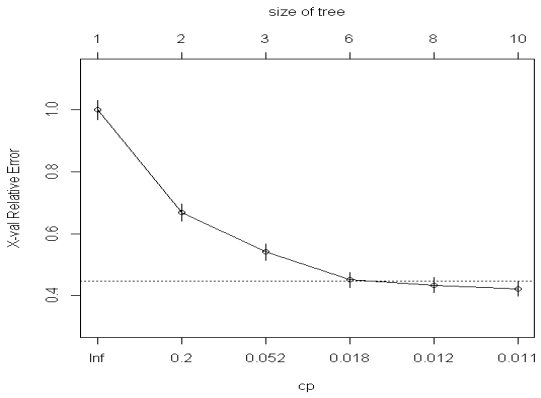


Figure 10. Classification Tree for Complexity Parameter (CP)

The above plot represents size of the tree and cp value is 0.018 (Figure 10.)

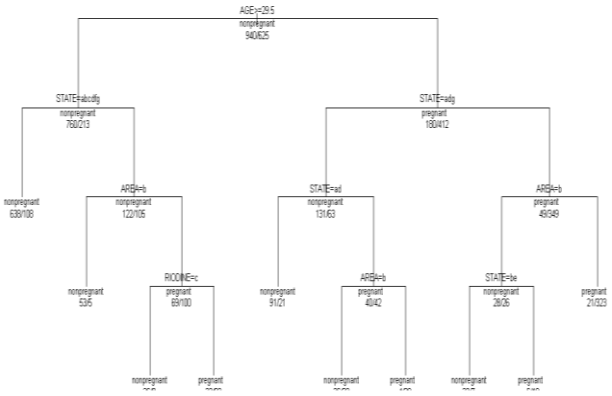


Figure 11. Classification tree for Dataset2

When considering decision tree, tree construction represents age, with state. On further classification, it is been grouped based on area and record test iodine (Figure 11.).

Table 4 Comparison of Data Mining Models

Model	Sensitivity	Specificity	Accuracy
Decision Tree	89.79%	81.28%	86.39%
Random Forest	91.38%	82.24%	87.73%
Naive Bayes	78.51%	84.96%	81.09%
SVM	91.60%	76.96%	85.75%
LDA	86.70%	80.16%	84.09%

In this dataset, five data mining classifier based on their sensitivity, specificity and accuracy was compared and found that SVM classifier has better classification precision than other classifier. (Figure 12, Table 4.)

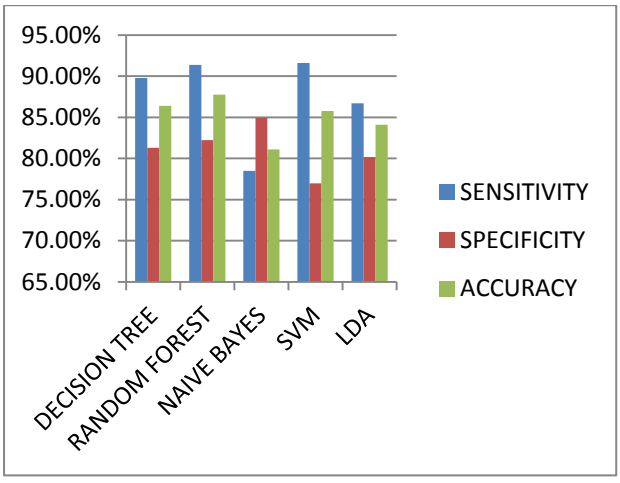


Figure 12. Graphical representations of sensitivity, specificity, and accuracy

Table 5. shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis.

Table 5. True Positive Rate and False Positive Rate

Models	True Positive Rate	False Positive Rate
Decision Tree	0.8979	0.1021
Random Forest	0.9138	0.0862
Naive Bayes	0.7851	0.2149
SVM	0.9160	0.084
LDA	0.8670	0.133

The results shows that SVM outperforms well than Decision Tree, Random Forest, Naive Bayes, LDA models, parameters Sensitivity, Specificity, Accuracy and Error Rates.

6.3 Dataset 3

6.3.1 Classification Tree:

In the dataset result established that the root node error $\frac{521}{4521} = 0.11524, and\ sample\ Size\ \ n = 4521$ .In classification tree, the variables used in tree construction for the data are day, duration, job, marital status, month, pdays and poutcome. The root node error is 0.11524.

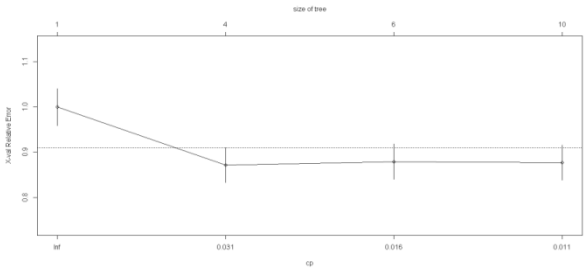


Figure 13. Classification Tree for Complexity Parameter (CP)

The above plot represents the size of the tree and cp value is 0.031 (Figure 13.).

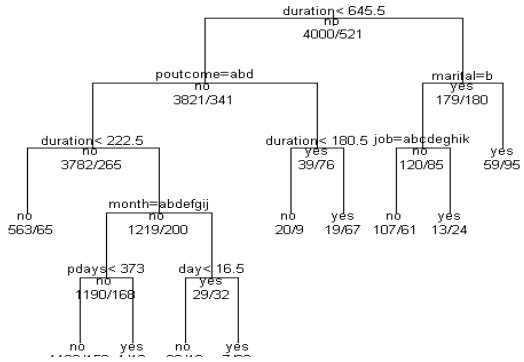


Figure 14. Classification tree for Dataset3

Data based on classification when considering decision tree, the tree construction represents the duration, with poutcome and marital status. On further classification, it is been grouped based on month, pdays, and job (Figure 14.).

Table 6. Comparison of Data Mining Models

Model	Sensitivity	Specificity	Accuracy
Decision Tree	96.88%	46.45%	91.15%
Random Forest	96.55%	40.69%	90.11%
Naive Bayes	91.50%	51.44%	86.88%
SVM	98.95%	23.03%	90.20%
LDA	96.53%	42.99%	90.36%

In this dataset, five data mining classifier based on their sensitivity, specificity and accuracy were compared. The experiment proved that SVM classifier has better classification precision than other classifiers Table 6, Figure 15.)

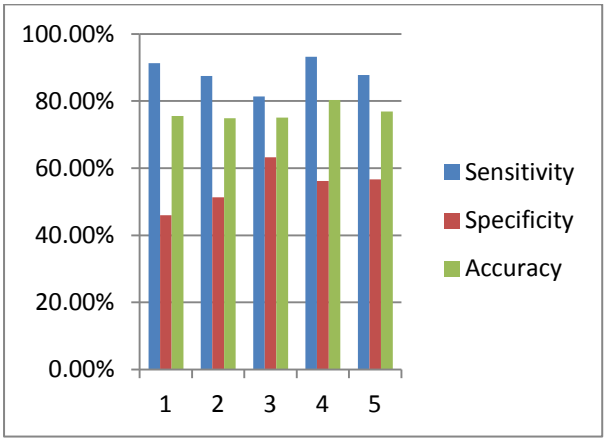


Figure 165 Graphical Representations of Sensitivity, Specificity, and Accuracy

Table 7. shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis.

Table 7. True Positive Rate and False Positive Rate

Models	True Positive Rate	False Positive Rate
Decision Tree	0.9688	0.0312
Random Forest	0.9655	0.0345
Naive Bayes	0.9150	0.085
SVM	0.9895	0.0105
LDA	0.9653	0.0347

The results shows that out of Decision Tree, Random Forest, Naive Bayes, SVM and LDA models, parameters Sensitivity, Specificity, Accuracy and Error Rates, SVM outperforms well

6.4 Dataset 4

6.4.1 Classification Tree:

In the dataset result established that the root node error $\frac{2449}{991} = 0.24712, and\ sample\ \ Size\ \ n = 9910$ In classification tree, the variables used in tree construction for the data are age, education, hours per week, marital status, and occupation. The root node error is 0.24712. (Figure 16)

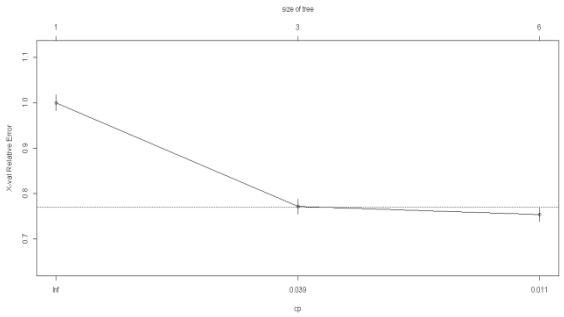


Figure 16. Classification Tree for Complexity Parameter (CP)

The above plot represents the size of the tree and cp value is 0.039. (Figure 16.)

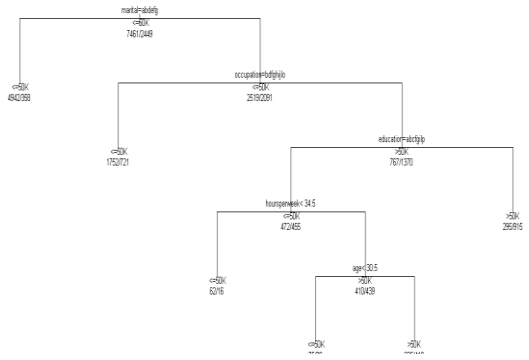


Figure 17. Classification tree for Dataset4

Data based on classification when considering decision tree, the tree construction represents the marital, with occupation. On further classification, it is been grouped based on education, hour per week, and age (Figure 17.)

Table 8. Comparison of Data Mining Models

Model	Sensitivity	Specificity	Accuracy
Decision Tree	91.48%	57.37%	83.05%
Random Forest	91.52%	56.76%	82.93%
Naive Bayes	86.57%	65.66%	81.40%
SVM	92.88%	53.33%	83.11%
LDA	91.93%	54.68%	82.72%

In this dataset, five data mining classifier based on their sensitivity, specificity and accuracy was compared and found that SVM classifier has better classification precision than other classifiers (Table 8, Figure 18.).

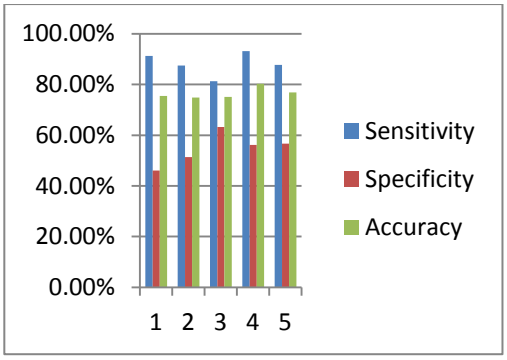


Figure 18. Graphical representation of different classification

Table 9. shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and Linear Discriminant Analysis.

Table 9. true positive rate and false positive rate

Models	True Positive Rate	False Positive Rate
Decision Tree	0.9148	0.0852
Random Forest	0.9152	0.0848
Naive Bayes	0.86.57	0.1343
SVM	0.9288	0.0712
LDA	0.9193	0.0807

The results shows that out of Decision Tree, Random Forest, Naive Bayes, SVM and LDA models, parameters Sensitivity, Specificity, Accuracy and Error Rates SVM outperforms well. A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results. The table below shows the confusion matrix (Table 10.).

Table 10. Classification Matrix

Actual/predicted	**0**	1
0	TP	FN
1	FP	TN

The upper left cell denote the number of samples classified as true while they were true (i.e., TP), and the lower right cell denotes the number of samples classified as false while they were actually false (i.e., TN). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the upper right cell denotes the number of samples classified as false while they were actually true (i.e., FN), and the lower left cell denotes the number of samples classified as true while they are actually false (i.e., FP).

6.5 Sensitivity, Specificity and Accuracy

Below formulae were used to calculate sensitivity, specificity and accuracy:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificit\,y = \frac{TN}{(TN + FP)}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Performance analysis was carried out on five different data mining classifier for four different datasets. Datasets considered are from survey domain. The present work has been implemented in R language environment and the results have been taken under different parameters: the sensitivity, accuracy and Kappa Statistic. The results obtained from these different models have been defined in the form of tables as well as graph (Table 11, Figure 17.).

6.5 Comparison of Sensitivity, Specificity and Accuracy for four Databases

Table 11. Comparison for Sensitivity, Specificity and Accuracy for four databases

Data	Model	Sensitivity	Specificity	Accuracy
Data 1	Decision Tree	91.25%	46.02%	75.50%
	Random Forest	87.47%	51.33%	74.88%
	Naive Bayes	81.32%	63.27%	75.04%
	SVM	93.14%	56.19%	80.28%
	LDA	87.71%	56.64%	76.89%
Data 2	Decision Tree	89.79%	81.28%	86.39%
	Random Forest	91.38%	82.24%	87.73%
	Naive Bayes	78.51%	84.96%	81.09%
	SVM	91.60%	76.96%	85.75%
	LDA	86.70%	80.16%	84.09%
Data 3	Decision Tree	96.88%	46.45%	91.15%
	Random Forest	96.55%	40.69%	90.11%
	Naive Bayes	91.50%	51.44%	86.88%
	SVM	98.95%	23.03%	90.20%
	LDA	96.53%	42.99%	90.36%
Data 4	Decision Tree	91.48%	57.37%	83.05%
	Random Forest	91.52%	56.76%	82.93%
	Naive Bayes	86.57%	65.66%	81.40%
	SVM	92.88%	53.33%	83.11%
	LDA	91.93%	54.68%	82.72%

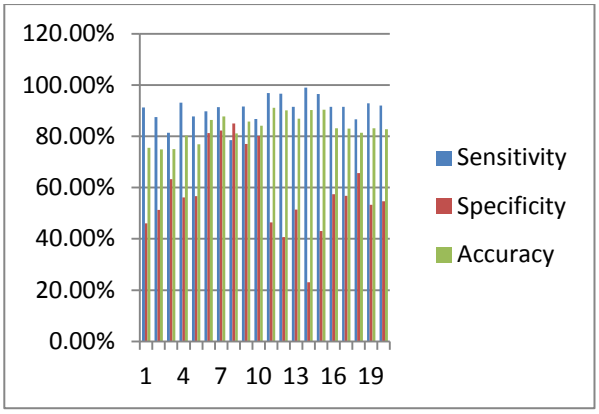


Figure 19. Comparison of Sensitivity, Specificity, and Accuracy for four Databases

It is clear that figure 19 shows the accuracy based comparison of different classification. It shows that SVM is most robust, effective, and consistent classifier for different datasets. SVM provides higher accuracy among all classification where as Naive Bayes is the least effective classification in terms of accuracy analysis.

6.6 Comparison of True Positive and False Positive Rate for All Databases

Table 12. Comparison of True Positive and False Positive Rate for Four Databases

Data	Models	True Positive Rate	False Positive Rate
Data 1	Decision Tree	0.9125	0.0875
	Random Forest	0.8747	0.1253
	Naive Bayes	0.8132	0.1868
	SVM	0.9314	0.0686
	LDA	0.8771	0.1229
Data 2	Decision Tree	0.8979	0.1021
	Random Forest	0.9138	0.0862
	Naive Bayes	0.7851	0.2149
	SVM	0.916	0.0840
	LDA	0.867	0.1330
Data 3	Decision Tree	0.9688	0.0312
	Random Forest	0.9655	0.0345
	Naive Bayes	0.915	0.0850
	SVM	0.9895	0.0105
	LDA	0.9653	0.0347
Data 4	Decision Tree	0.9148	0.0852
	Random Forest	0.9152	0.0848
	Naive Bayes	0.86.57	0.1343
	SVM	0.9288	0.0712
	LDA	0.9193	0.0807

Table 12 shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Naive Bayes, Linear Discriminant Analysis and Support Vector Machine (SVM).

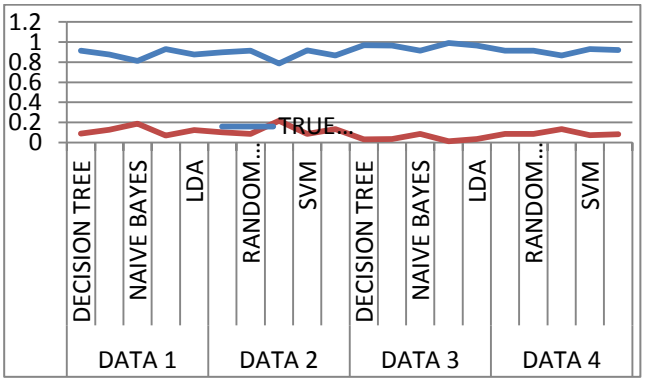


Figure 20. Comparison of True Positive and False Positive Rate of four Databases

Fig. 20 shows that True Positive Rate and False Positive Rate for Decision Tree, Random Forest, Naive Bayes, Linear Discriminant Analysis and Support Vector Machine (SVM). It represents above 80% True Positive Rate and less than 20% False Positive Rate for all four datasets.

6.7 Comparison of Kappa Statistic for Different Datasets

Table 13. Comparison of Kappa Statistics for four Databases using Various Data Mining Tools

Data	Models	Kappa Value
Data 1	Decision Tree	0.4085
	Random Forest	0.4122
	Naïve Bayes	0.4478
	SVM	0.6518
	LDA	0.4655
Data 2	Decision Tree	0.7147
	Random Forest	0.7422
	Naïve Bayes	0.6168
	SVM	0.6977
	LDA	0.6684
Data 3	Decision Tree	0.5002
	Random Forest	0.4344
	Naïve Bayes	0.4003
	SVM	0.3139
	LDA	0.4552
Data 4	Decision Tree	0.5174

	Random Forest	0.5127
	Naïve Bayes	0.511
	SVM	0.5045
	LDA	0.501

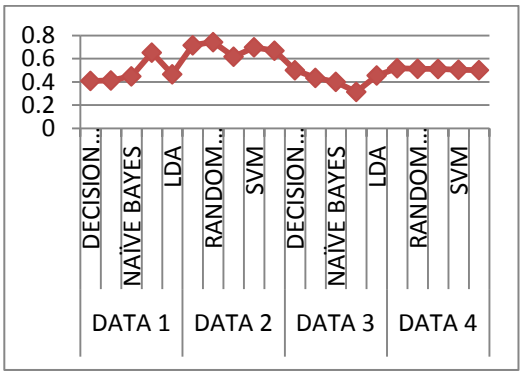


Figure 21. Comparisons Charts for four Databases using Kappa Statistics

Kappa Statistics is a statistical analysis based on inter-ratter agreement for qualitative data. It basically performs the analysis between different classes (Table 13.). Higher Value of kappa statistic is considered as good. Figure 21 shows the comparative analysis of different classification under the kappa statistics.

7. CONCLUSION

This paper focuses on various classification techniques used in data mining and a study on each of them. Data mining can be used in a wide area that integrates techniques from various fields including machine learning, Network intrusion detection, spam filtering, artificial intelligence, statistics and pattern recognition for analysis of large volumes of data. Classification methods are typically strong in modeling communications. Classification is the preliminary stage of data mining which is used to categorize dataset in smaller groups where each group contains similar data items. The classification basically deals with two main parameters in which one is the number of classes and another is the criteria for deciding the class members. The accuracy of classification algorithm also decides the effectiveness of its use in other mining applications. The present work is about to analyze the effectiveness of most popular classification techniques. In this research paper, analysis has been performed for five different classification methods in terms of precision, accuracy, and kappa statistics under four datasets, collected from different domain. The work has been implemented in R language environment and obtained results show that SVM is the most robust classification method. Due to the nature of some data sets, the result reveals that all data mining techniques accomplish their goals perfectly, but each technique has its own characteristics and specification that demonstrate their precision, accuracy, proficiency and preference.

In this research paper, performances of data mining classifiers are analyzed and evaluated. Accuracy can be estimated by calculating error rate between predicated value and actual value. Accuracy of decision tree is better than other data mining techniques, cross validation method, but each of the technique has its own characteristics and specification that demonstrate their accuracy, proficiency and preference. In this study, Support Vector Machines, Naïve Bayes, Decision Trees, Random Forest and Linear Discriminant Analysis have been implemented on 4 datasets. The goal of the research was to evaluate the performance of classification using a variety of performance metrics: classification accuracy, precision, and specificity.

Based on the experimental results, the SVM model proved to have the best performance. It gives better results, when compare to other data mining techniques for all datasets were used. Decision tree and random forest also performed well. The results show that performance of each classification depends on what type of problem is being considered. The performance of classification also depends on performance matrix and the characteristics dataset. The relationships between dataset characteristics and model accuracy were not discussed in this study. It is known that dataset characteristics influence the accuracy of classification and therefore this may influence the conclusion of the findings. Another limiting factor is the sizes of dataset in which two out of the four dataset has less than 2,000 instances.

8. REFERENCES

[1] Brijesh Kumar Bhardwaj, Saurabh Pal (2011), Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011, pp 136-140.

[2] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni (2011), Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, pp. 43-48.

[3] S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan (2010), Text Classification Using Data Mining, *ICTM*, pp.1-9.

[4] Pushpalata Pujari (2013), Classification And Comparative Study of Data Mining Classifiers with Feature Selection on Binomial Data Set, Journal of Global Research in computer Science, Vol.5, No.3, pp.39-45.

[5] Dahinden, C., 2009. An improved Random Forests Approach with Application to the Performance Prediction Challenge Datasets. Hands on Pattern Recognition. Microtome. Seminar fur¨ Statistik CH-8092 Zurich, ¨ Switzerland, pp.1-6.

[6] Jiawei Han, Micheline Kamber Jian Pei (2012), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers is an Imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA.

[7] U. Rajendra Acharya, P. Subbanna Bhat, S.S. Iyengar , Ashok Rao, Sumeet Dua (2003), Classification of heart rate data using artificial neural network and fuzzy equivalence relation, Pattern Recognition, Volume 36, issue.1. pp.61-68.

[8] https://blog.floydhub.com/naive-bayes-for-machine-learning/

[9] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

[10] Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annuals of Eugenics. 1936;7:179–188. Annals of Human Genetics, UCL and Blackwell Publishing Ltd.

9. AUTHOR'S PROFILE

Dr. P. Arumugam received his M.Sc. M. Phil. Ph.D from Annamalai University. He is working as a Associate Professor, Department of Statistics, Annamalai University, Chidambaram, India. He has good research experience by working for many Project Guidance and consultation work in Theoretical and Application of Statistics. He has published more than twenty five research papers in National and International Journals. He was produced more than five Ph. D Degree in the area of Statistics. His area of interest are Bayesian Inference, Stochastic Processes, Data Mining and other related Statistics and Computer Science field. He has published a book entitled, A Bayesian Analysis of Changing Time series models, 99,66123 Saurbrucken, Germany: Lap Lambert Academic Publishing Gmbh & Co, KG, 978-3-8433-9367-6, pp. 1-136. Jan 2011. He has been a resource person for various academic events.

Prof. A. Poompavai. Received her M. Sc. M. Phil. in Statistics from University of Madras, Chennai, India. She is Working as Assistant Professor in Statistics, Department of Statistics, Apollo Arts and Science College, Chennai. She has good research experience. in Application of Statistics and Data Mining. She has published more than five research papers in various national and International journals. Her area of interest are Applied Statistics, Data Mining, Bio-Statistics and other related Statistics.. She has good knowledge in programming languages like, C, C++, VB, etc. and SPSS. R-Studio.

Dr. G. Manimannan received his M. Sc. M. Phil. Ph. D in Statistics from University of Madras, Chennai, India. He received PGDCA (Post Graduate Diploma in Computer Application) from Pondicherry University, Pondicherry; He has also received MCA Degree from Bharathidasan University, Tiruchirappalli, India. He has good research experience by working for many Project Guidance and consultation work in Application of Statistics and Data Mining. He has published more than sixty five research papers in various national and International journals. His area of interest are Quantitative Logistics, Applied Statistics, Data Mining, Text Mining, Neural Networks, Bio-Statistics and other related Statistics and Computer Science field. He is good in computer programming languages, Statistical Packages like, SPSS, SYSSTAT, STATISTICA, MINITAB, MATLAB, Data Mining Software WEKA, Orange Data Mining, R-Studio and working knowledge in SAS.

Dr. R. Lakshmi Priya received her M. Sc. M. Phil. Ph.D in Statistics from University of Madras, Chennai, India. She is Working as Assistant Professor of Statistics, Department of Statistics, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai. She has good research experience by working for many Project Guidance and consultation work in Application of Statistics and Data Mining, Neural Networks, etc.. She has published more than thirty research papers in various national and International journals. Her areas of interest are Applied Statistics, Data Mining, Bio-Statistics and other related Statistics. She has good knowledge in programming languages like, C, C++, VB and SPSS, R-Studio.