International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 1, 2020, PP 1-11
ISSN No. (Online) 2347-3142
DOI: http://dx.doi.org/10.20431/2347-3142.0801001
www.arcjournals.org

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

P. Arumugam[1], Poompavai A[2], Manimannan G[3*], Poongothai[4]

[1]*Associate Professor, Department of Statistics, Annamalai University, Chidhambaram,*

[2]*Assistant Professor, Department of Statistics, Apollo Arts and Science College, Chennai.*

[3]*Assistant Professor, Department of Mathematics, TMG College of Arts and Sciences, Chennai.*

[4]*Assistant Professor, Department of Statistics, Apollo Arts and Science College, Chennai*

Corresponding Author:** ***Manimannan G, *Assistant Professor, Department of Mathematics, TMG College of Arts and Sciences, Chennai.*

Abstract: *In this paper, an attempt is made to identify and cross validate with three 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 programming 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 three 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 the datasets used. Naive Bayes Classifier 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 three 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: *Random Forest, Naive Bayes Classifier, 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 areused 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'.

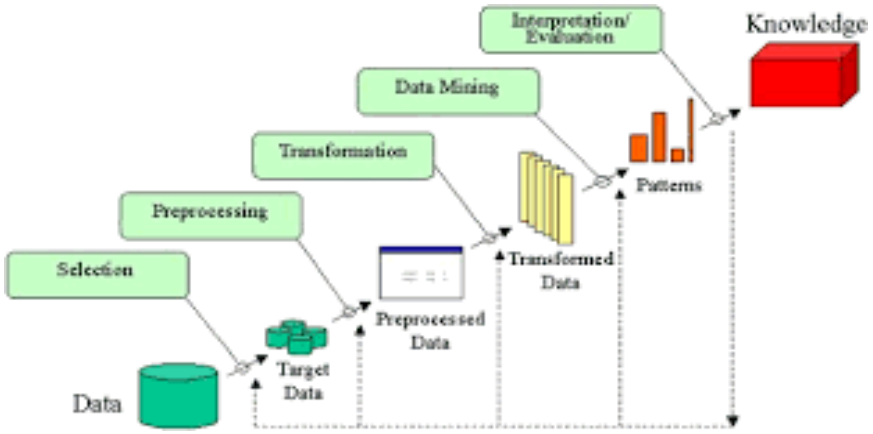


Figure1. *KDD of Data Mining Processes*

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 most of the rest of statistics, has been focused on data sets far smaller than most that are 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 rapidly. Brijesh Kumar Bhardwaj, Saurabh Pal used Bayes classification prediction model to identify the difference between high learners and slow learners student. U. Rajendra Acharya, P. Subbanna Bhat, S.S. Iyengar , Ashok Rao, Sumeet Dua 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. The feedforward architecture ANN classifier is seen *to be* correct in about 85% of the test cases, and the fuzzy classifier yields correct classification in over 90% of the cases.

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich" but „knowledge poor". Wealth of data is available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having

similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. (Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni)

Text classification is the process of classifying documents into predefined categories based on their content. It is an automated assignment of natural language texts to predefined categories. 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. Existing supervised learning algorithms to automatically classify text need sufficient documents to learn accurately. A new algorithm for text classification using data mining that requires fewer documents for training. Instead of using words and word relation association rules from these words is used to derive feature set from pre-classified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of Genetic Algorithm has been added for final classification. A system based on the proposed algorithm has been implemented and tested. The experimental results show that proposed system works as a successful text classifier. (S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan)

Another research paper describes about the performance analysis of different data mining classifiers before and after feature selection on binomial data set. Three data mining classifiers Logistic Regression, SVM and Neural Network classifiers are considered for classification. 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 the 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% accuracy and neural network provides 98.13 % 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 the classifiers with reduced number of features. (Pushpalata Pujari)

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 were deleted from dataset. After deleting records with missing values, 649 were left out. On these 649 records data mining classification such as techniques, Random Forest, Support Vector Machine (SVM) and Naive Bayes Analysis were applied.

4.2. Dataset 2

The data is a secondary data and 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 were 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 Random Forest, Support Vector Machine (SVM) and Naive Bayes 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 were 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 Random Forest, Support Vector Machine (SVM) and Naive Bayes Analysis were applied.

In this study, three 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 description of these databases is given below (Table 1).

Table1. *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

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 programming tool under different classification approaches.

R programming 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. 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 Geo-informatics 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 2).

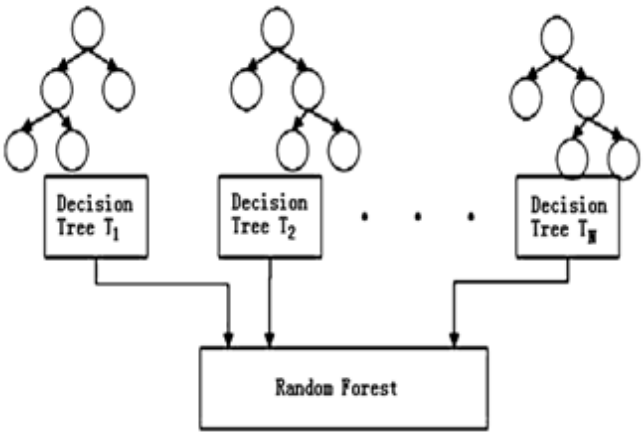


Figure 2. *Sample Random Forest Tree*

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 the expense of a some loss of interpretability, but generally greatly boosts the performance of final model.

5.2. 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 3).

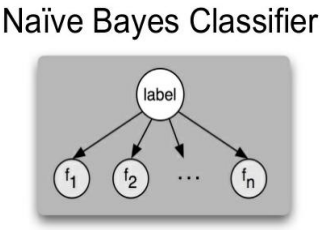
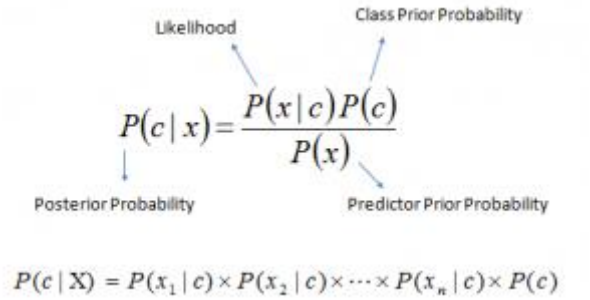


Figure3. .*Naive Bayes Classsifer*

5.2.1. 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:



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.

5.3. 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 4).

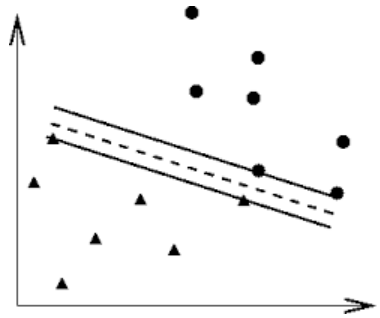


Figure4. *Support Vector Machine*

6. RESULT AND DISCUSSION

6.1. Dataset 1

In the data set result established that the root node error: $\frac{226}{649} = 0.34823, and\ sample\ \ Size\ \ n = 649$. In classification techniques, the following parameter are used, it has been group based on father education, travel time, study time and activities

Table2. *Comparison of Data Mining Models*

.Model	Sensitivity	Specificity	Accuracy
Random Forest	87.47%	51.33%	74.88%
Naive Bayes	81.32%	63.27%	75.04%
SVM	93.14%	56.19%	80.28%

In this dataset, the researcher compared three data mining classifier based on their sensitivity, specificity and accuracy. It shows that SVM classifier has better classification precision compared with other classifier (Figure 5, Table 2).

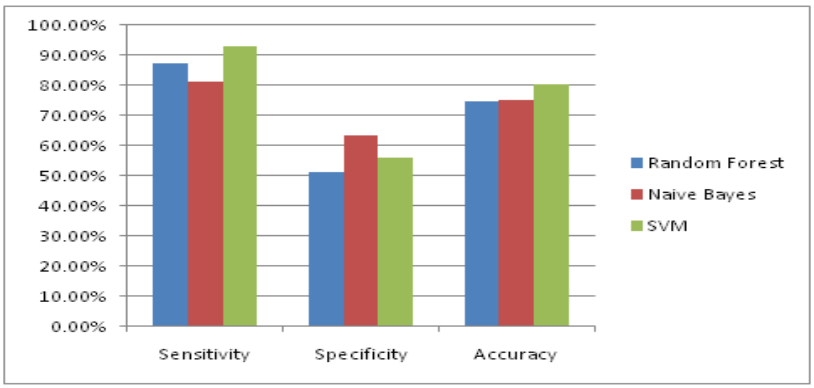


Figure5. *Graphical representations of sensitivity, specificity and accuracy*

Table 3 shows True Positive Rate and False Positive Rate for Random Forest, Support Vector Machine (SVM) and Naive Bayes Classifier.

Table3. *true positive rate and false positive rate*

Models	True Positive Rate	False Positive Rate
Random Forest	0.8747	0.1253
Naive Bayes	0.8132	0.1868
SVM	0.9314	0.0686

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

6.2. Dataset 2

In this dataset result established that the root node error $\frac{621}{1565} = 0.39680, and\ sample\ \ Size\ \ n = 1565$ In classification techniques, the variables used in tree

construction for the data are age, area, record test iodine and state. When considering decision tree, tree construction represents age, with state. On further classification, it is been grouped based on area and record test iodine.

Table4. *Comparison of Data Mining Models*

Model	Sensitivity	Specificity	Accuracy
Random Forest	91.38%	82.24%	87.73%
Naive Bayes	78.51%	84.96%	81.09%
SVM	91.60%	76.96%	85.75%

In this dataset, three data mining classifier based on their sensitivity, specificity and accuracy was compared and found that SVM classifier has better classification precision than other classifier (Table 4, Figure 6).

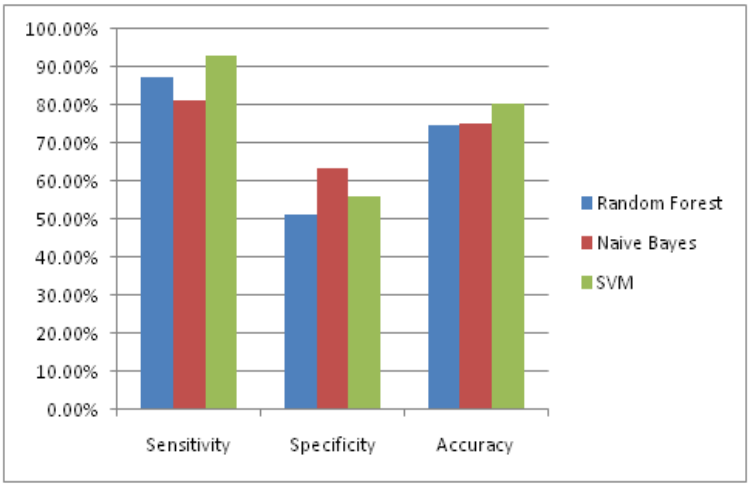


Figure6. *Graphical representations of sensitivity, specificity, and accuracy*

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

Table5. *True Positive Rate and False Positive Rate*

Models	True Positive Rate	False Positive Rate
Random Forest	0.9138	0.0862
Naive Bayes	0.7851	0.2149
SVM	0.9160	0.084

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

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. 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.

Table6. *Comparison of Data Mining Models*

Model	Sensitivity	Specificity	Accuracy
Random Forest	96.55%	40.69%	90.11%
Naive Bayes	91.50%	51.44%	86.88%
SVM	98.95%	23.03%	90.20%

In this dataset, three 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.

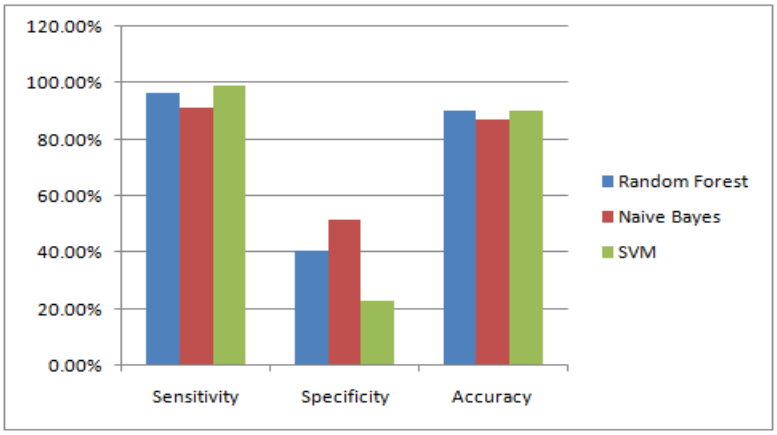


Figure7. *Graphical Representations of Sensitivity, Specificity, and Accuracy*

Table 7 shows True Positive Rate and False Positive Rate for Random Forest, Support Vector Machine (SVM) and Naive Bayes (Table 6, Figure 7).

Table7. *True Positive Rate and False Positive Rate*

Models	True Positive Rate	False Positive Rate
Random Forest	0.9655	0.0345
Naive Bayes	0.9150	0.085
SVM	0.9895	0.0105

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.

The results shows that out of Random Forest, Naive Bayes and SVM 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 8)

Table8. *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 (TP), and the lower right cell denotes the number of samples classified as false while they were actually false (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 (FN), and the lower left cell denotes the number of samples classified as true while they are actually false (FP).

6.4. Sensitivity, Specificity and Accuracy Calculation

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 three different datasets. Datasets considered are from survey domain. The present work has been implemented in R programming 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.

6.5. Comparison of Sensitivity, Specificity and Accuracy for Four Databases

Table9. *Comparison for Sensitivity, Specificity and Accuracy for four databases*

Data	Model	Sensitivity	Specificity	Accuracy
Data Set 1	Random Forest	87.47%	51.33%	74.88%
	Naive Bayes	81.32%	63.27%	75.04%
	SVM	93.14%	56.19%	80.28%
Data Set 2	Random Forest	91.38%	82.24%	87.73%
	Naive Bayes	78.51%	84.96%	81.09%
	SVM	91.60%	76.96%	85.75%
Data Set 3	Random Forest	96.55%	40.69%	90.11%
	Naive Bayes	91.50%	51.44%	86.88%
	SVM	98.95%	23.03%	90.20%

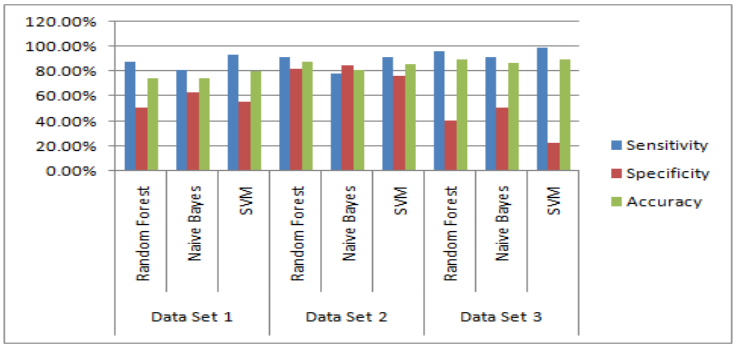


Figure 7 for Comparison of Sensitivity, Specificity, and Accuracy for four Databases

It is clear that figure 7 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 whereas 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

Table10. *Comparison of True Positive and False Positive Rate for Four Databases*

	Models	True Positive Rate	False Positive Rate
Data Set 1	Random Forest	0.8747	0.1253
	Naive Bayes	0.8132	0.1868
	SVM	0.9314	0.0686
Data Set 2	Random Forest	0.9138	0.0862
	Naive Bayes	0.7851	0.2149
	SVM	0.916	0.084
Data Set 3	Random Forest	0.9655	0.0345
	Naive Bayes	0.915	0.085
	SVM	0.9895	0.0105

Table 10 shows True Positive Rate and False Positive Rate for Random Forest, Naive Bayes, and Support Vector Machine (SVM).

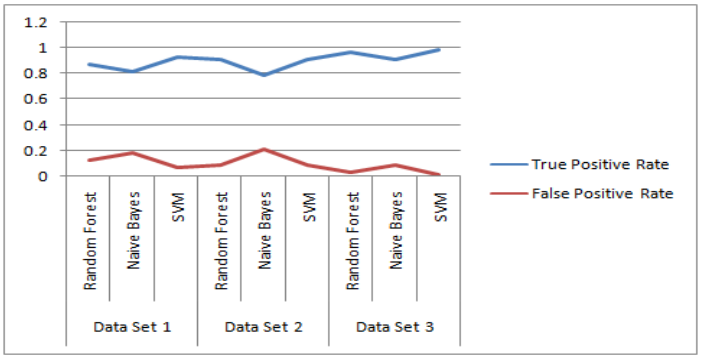


Figure8. *Comparison of True Positive and False Positive Rate of four Databases*

Fig. 8 shows True Positive Rate and False Positive Rate for Random Forest, Naive Bayes 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

Table11. *Comparison of Kappa Statistics for four Databases using Various Data Mining Tools*

	Models	**Kappa Statistic**
Data Set 1	Random Forest	0.4122
	Naive Bayes	0.4478
	SVM	0.6518
Data Set 2	Random Forest	0.7422
	Naive Bayes	0.6168
	SVM	0.6977
Data Set 3	Random Forest	0.4344
	Naive Bayes	0.4003
	SVM	0.3139

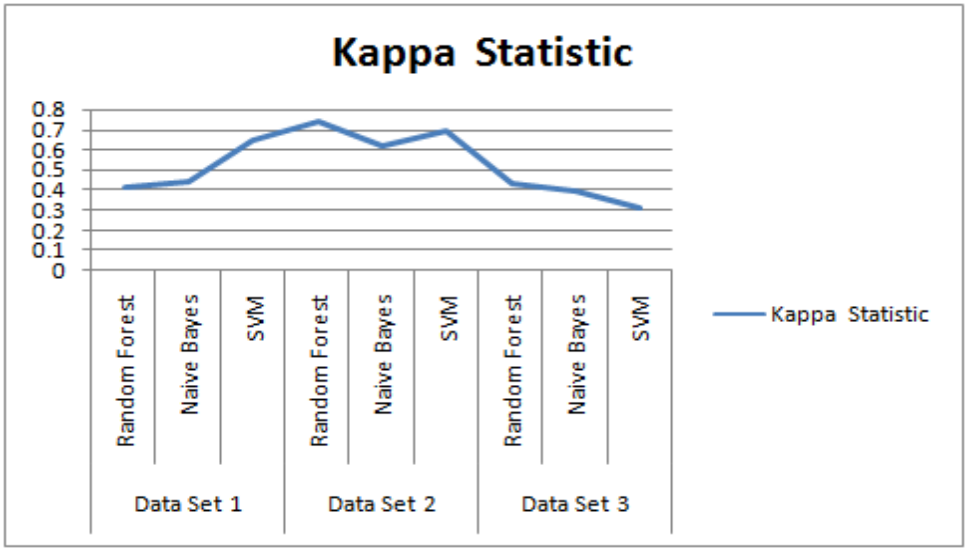


Figure9. *Comparisons Charts for four Databases using Kappa Statistics*

Kappa Statistics is a statistical analysis based on inter-ratter agreement for qualitative data. It basically performs the analysis between different classes. Higher Value of kappa statistic is considered as good (Table 11). Figure 9 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 paper, analysis has been performed for three different classification methods in terms of precision, accuracy, and kappa statistics under three datasets, collected from different domain. The work has been implemented in R programming 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 Addition, Support Vector Machines, Naïve Baye

and Random Forest have been implemented on three datasets. The goal of the research was to evaluate the performance of the classification using a variety of performance metrics: classification accuracy, precision, and specificity.

According to the experimental results, the SVM model proved to have the best performance. It performed better than all of the datasets used. Naive Bayes 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 three dataset has less than 2,000 instances.

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.

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

[3] S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan (2005), Text Classification Using Data Mining, *ICTM.*

[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.

[5] Dahinden, C., 2009. An improved Random Forests Approach with Application to the Performance Prediction Challenge Datasets. Hands on Pattern Recognition. Microtome.

[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.

Citation: *Manimannan G,et.al., (2020). Comparative Study of Data Mining Classifiers with Different Features and Different Databases Domain. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 8(1), pp. 1-11. http://dx.doi.org/ 10.20431/2347 -3142.0801001*

Copyright: *© 2020 Authors, this is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*