# Feature Selection Based Enhancement Of The Accuracy Of Classificationalgorithms

**A. Adhiselvam[1], K. Umamaheswari[2], J. Ramya[3]**

[1]Assistant Professor and Head, Department of Computer Applications,S.T.E.T. Women's College, Mannargudi, India.

[2]Assistant Professor, Department of Computer Science,Avvaiyar Government College for Women, Karaikal, India.

[3]Assistant Professor, Department of Computer Applications,S.T.E.T. Women's College, Mannargudi, India.

**Article History**: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

***Abstract:*** Feature selection is played vital role for classification algorithmsin the machine learning. In real time data mining process, irrelevant features which are available in the dataset may decrease the accuracy level of the classification algorithms. The selection ofthe appropriateand relevant features of the dataset in classification problems is the important role in data mining. The aim of the research work is to increase accuracy level of the classification algorithmsusing feature selection technique with different domain datasets.This paper also comparesaccuracy of different classification algorithms with and without feature selection method. The classification algorithms such as Bayesian Net, Naïve Bayes, Multi Layer Perception, logistic regression, J48 and Random Forestare used with feature selection methodusing different domain datasetssuch as Breast Cancer, Glass, Iris and Weather for comparison.This experiment is done with the help of Weka tool and datasetsin the machine learning repository. Feature selectiontechniques are effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving result comprehensibility. However, the recent increase of dimensionality of data poses severe challenge to many existing feature selection techniques with respect to efficiency and effectiveness.

***Keywords***: Feature selection, Dimensionality reduction, Classification algorithms, Supervised Learning, machine learning

## 1. INTRODUCTION

This research work conducted a comparison of data mining classification algorithms under Bayesian classification, Function classification, and Decision Tree classification using different datasets. The accuracy of the classification algorithmsis improved using feature selection method.

### 1.1 Classification Algorithms

Classification is a data mining technique which is used to predict group membership for data instances.Three types of classification models are considered for this research study.

Bayesian Classification is statistical classifiers. They can predict class membership probabilities that a given tuple belong to a particular class. It is based on baye's theorem. Types of this classification are Naïve Bayesian Classification and Bayesian Belief Networks. In our work, the Bayesian Net and Naïve Bayes classification methods are considered.

Function classification is based on constructing function taking input feature vector X and predicting it outcome Y. In this work Multi Layer Perception (MLP)and logistics classifier methods are implemented using Weka tool. A MLP is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Their current output depends only on the current input instance. It trains using back propagation. Logistics is a binary classification model.

Decision tree (DT) learning algorithm work based on processing and deciding upon attributes of the data, Random Forest (RF)and J48 were used in our experiments. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. It is a predictive modeling technique used in classification, clustering and prediction tasks. Decision tree classification technique is performed in two phases namely tree building and tree pruning. Tree building is done in top-down manner. Tree Pruning is done in a bottom–up fashion.

### 1.2.Feature selection

Feature selection is one of the vital roles in the field of machine learning. It is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving result comprehensibility. However, the recent increase of dimensionality of data poses severe challenge to many existing feature selection methods with respect to efficiency and effectiveness.

### 1.3. Objectives

The main aim of this research work is to enhance accuracy of the classification algorithms with feature selection. To attain the task to enhance the accuracy of classification algorithms, the following objectivesare framed.

- To build a classifier using different dataset. To create an Ensemble of Techniques that will address the short comings of the existing approaches technique that is not affected by the feature selection
- To use a feature selection technique for reduce the number of attributes and improve the level of classification accuracy.
- To develop the framework based on different classification algorithms and compare the performance of different classifiers with feature selection and without feature selection and analyze the results.

## 2. REVIEW OF LITERATURE

Many researchers studied classification algorithms with feature selection method to enhance the accuracy of the algorithms in various data sets. Some of the important research works are identified and reviewed in this section. These research articles help us to propose a feature based method to enhance the accuracy of classification algorithms.

M. Krishnaveni et al. summarized the feature selection process, different types of feature selection algorithms such as Filter, Wrapper and Hybrid and their importance. Moreover, it analyzed some of the existing popular feature selection algorithms through a literature survey and also addresses the strengths and challenges of feature selection algorithms[1]. N. Sai Sragvi Vibhushan et al. discussed about predicting the performance of the students. A feature selection algorithm removes the extraneous data and helps in increasing the accuracy of the classifier. An ensemble method produces different models and combines them to produce improvised results and compare with various feature selection algorithms which concluded the best feature selection algorithm [2].

Fifie Francis et al. concluded that J48 and Navie Bayes are most commonly used classification algorithms in the area of prediction analysis and Gain Ratio Attribute Evaluator and Ranker Algorithm are used for feature selection [3].Maryam Zafar et al. proved that Feature Selection (FS) improved the quality of prediction models for datasets. FS algorithms eliminate unrelated data from the dataset and increase the performance of classifier accuracy Best features can produce better results [4].

Aleyani et al. proposed the clustering based feature selectionand summarized the various classification techniques for feature selection [5].Domeniconi et al. explained the local feature selection techniques and computational methods [6].Brodley et al. summarized the techniques for unsupervised learning for the feature selection technique[7].Guyon et al. discussed the various methods for feature selection rather information gainand compared the different approach for the feature selection methods based on filter or wrapper methods[8]. Liu et al. summarized the different techniques supportable for the feature selection and knowledge discovery [9].Mitra et al. proposed the techniques for unsupervised feature selection based on different domain and compare the different similarity measures for the clustering [11].Abdullah et al. discussed the feature selection and data mining classifiers and proposed the ensemble techniques for classification and helpful for decisions making [12].

Gupta et al.summarized the results based on the classification techniques that different feature selection methods are used as Search method in feature selection and apply the best features and find the accuracy of classifier and he concluded that Feature selection played an important role in classification [13].

Nikhil et al. proved that the accuracy of the classifier improved while applying the feature selection technique, after applying the feature selection the classifier performed better prediction [14].Rohit et al. [15] done the comparative analysis of different classification algorithms and find the accuracy and performance measures of each classifier and showed that best classification.

## 3. PROPOSED METHODOLOGY

The proposed work is designed for the analysis of data mining classification algorithms and to enhance the accuracy of the classifier using feature selection methods. This work has two phases such as to find the accuracy of classifier withoutfeature selection from the dataset and to find the accuracy of the same classifier with features selection from the dataset. Finally the results of with and without feature selection of the different classifiers are compared.The architecture of the proposed methodology is depicted in the Figure 1. The proposed method involves the following steps:

Step 1: Take Datasets from the machine learning repository
Step 2: Apply the feature selection methods in the data sets
Step 3: Select the best feature among all the features using Information gain method
Step 4: Remove the irrelevant features
Step 5: Applythe classification algorithms such as Naïve Bayes, Bayes Net, MLP, Logistic, J48 and Random Forest

Step 6: Find the accuracy of each classification algorithm without feature selection and with feature selection using precision,
recall and Fmeasure
Step 7: Predict the best classification algorithm for given data set

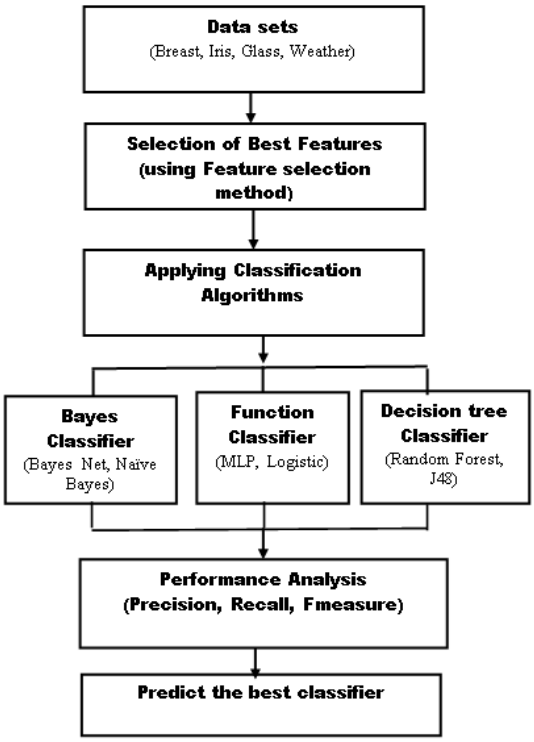


**Figure 1:** Architecture of the proposed methodology

In the first step, Breast Cancer, Iris, Glass, and Weather data setsfrom the machine learning repository are considered. Best features are selected using feature selection method in the second step. In the third step, select the best feature among all the features using Information gain method. In the step 4, irrelevant features of the data sets are removed to apply classification algorithms. The classification algorithms Bayesian Net, Naïve Bayes, Multi Layer Perception, logistic regression, J48 and Random Forest are applied in the next step. The step 6 finds the accuracy of each classification algorithm with and without feature selection. Finally best classification algorithm is predicted for given data set in the step 7.

## 4. METHODS AND MATERIALS

In order to implement the proposed work, first and foremost suitable data mining software tool is required. Weka tool is used. Apart from the Weka tool, four data sets are considered to implement classification algorithms with feature selection method.

### 4.1. Weka tool

This experiment is done with help of data mining tool Weka (Waikato Environment and Knowledge Analysis) to perform the analysis.This software provides a set of methods and algorithms that help in better utilization of data and information available to users, including feature selection methods and classification algorithms for data analysis.

### 4.2. Data set Description

In this research study, different domain datasetssuch as Breast Cancer, Glass, Iris and Weather are used. Breast Cancer data set contains 9 attributes and 286 instances. There are 4 attributes and 150 instances in Iris data set. Glass contains 10 attributes and 214 instances.The fourth datasethas 5 attributes and 14 instances. The description of data set is given in the Table 1.

Table 1: Data set description

| Data set name | Number of Attributes | Number of Instances | Description |
|---|---|---|---|
| Breast Cancer | 9 | 286 | The instances are described by 9 attributes, some of which are linear and some are nominal. |

| | | | |
|---|---|---|---|
| Iris | 4 | 150 | The data set contains 3 classes of 50 instances each, where each class refers to a type of Iris plant. |
| Glass | 10 | 214 | The study of classification of types of Glass |
| Weather | 5 | 14 | The dataset contains classification of prediction whether play or not |

**4.3. Performance Measures**

The performance measures play vital role in finding accuracy of classification algorithms. The following performance measures such as precision, recall and accuracy are used to find the accuracy of the classifiers.

**Precision** is the fraction of relevant instances among the retrieved instances.

$$\text{Precision}=\frac{TruePositive}{TruePositive+FalsePositive}$$

**Recall** is the fraction of relevant instances that were retrieved.

$$\text{Recall}=\frac{TruePositive}{TruePositive+TrueNegative}$$

**Fmeasure** is a measure of the test accuracy.

$$\text{Fmeasure}=\frac{(2*TruePostive*\text{Precision})}{(TruePositive+\text{Precision})}$$

**4.4. Classification algorithms with feature selection**

The classification algorithms such as Bayes Net, Naïve Bayes, Multi Layer Perception, logistics, J48 and Random Forest(RF) are implemented in the four different data set such as Breast Cancer, Glass, Iris and Weather using with and without feature selection methods. For feature selection method, InfogainAttributeEval is used as Attribute selection Evaluator method and Ranker is used as search method. It evaluates the worth of an attribute by measuring the information gain with respect to the class.

InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute)

**5. RESULTS AND DISCUSSION**

The classification algorithms are evaluated with datasets and find the accuracy of each classifier using performance measures after applying feature selection method.The results of the classifiers with and without feature selection method for Breast Cancer data set are given in the Table 2. The results of the classifiers with and without feature selection method for Iris data set are given in the Table 3. The results of the classifiers with and without feature selection method for Glass data set are given in the Table 4. The results of the classifiers with and without feature selection method for Weather data set are given in the Table 5.

Table 2: Results for the classifiers with and without feature selection (Breast Cancer)

| **Algorithms** | **Without Feature Selection** | | | | | **With Feature Selection** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **True Positive Rate (%)** | **False Positive Rate (%)** | **Precision** | **Recall** | **Fmeasure** | **True Positive Rate (%)** | **False Positive Rate (%)** | **Precision** | **Recall** | **Fmeasure** |
| Bayes Net | 73 | 27 | 0.70 | 0.72 | 0.71 | 75 | 25 | 0.75 | 0.75 | 0.74 |
| Naïve Bayes | 72 | 28 | 0.70 | 0.70 | 0.70 | 76 | 24 | 0.76 | 0.74 | 0.73 |
| MLP | 72 | 28 | 0.70 | 0.71 | 0.70 | 73 | 27 | 0.74 | 0.73 | 0.71 |
| Logistic Regression | 68 | 32 | 0.66 | 0.68 | 0.67 | 68 | 32 | 0.66 | 0.68 | 0.67 |
| J48 | 75 | 25 | 0.75 | 0.75 | 0.71 | 79 | 21 | 0.78 | 0.78 | 0.72 |
| RF | 69 | 31 | 0.66 | 0.69 | 0.66 | 69 | 31 | 0.69 | 0.69 | 0.66 |

Table 3: Results for the classifiers with and without feature selection (Iris)

| Algorithms | Without Feature Selection | | | | | With Feature Selection | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure |
| Bayes Net | 94 | 6 | 0.95 | 0.95 | 0.95 | 95 | 5 | 0.95 | 0.95 | 0.95 |
| Naïve Bayes | 95 | 5 | 0.96 | 0.96 | 0.96 | 97 | 3 | 0.96 | 0.95 | 0.93 |
| MLP | 94 | 6 | 0.97 | 0.97 | 0.97 | 97 | 3 | 0.97 | 0.97 | 0.97 |
| Logistic Regression | 95 | 5 | 0.96 | 0.96 | 0.96 | 96 | 4 | 0.96 | 0.96 | 0.96 |
| J48 | 94 | 6 | 0.97 | 0.97 | 0.96 | 97 | 3 | 0.97 | 0.97 | 0.96 |
| RF | 93 | 7 | 0.97 | 0.97 | 0.96 | 95 | 5 | 0.97 | 0.97 | 0.96 |

Table 4: Results for the different classifiers with and without feature selection(Glass)

| Algorithms | Without Feature Selection | | | | | With Feature Selection | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure |
| Bayes Net | 70 | 30 | 0.70 | 0.70 | 0.69 | 72 | 28 | 0.72 | 0.71 | 0.61 |
| Naïve Bayes | 68 | 32 | 0.68 | 0.67 | 0.66 | 68 | 32 | 0.68 | 0.67 | 0.66 |
| MLP | 68 | 32 | 0.68 | 0.67 | 0.66 | 70 | 30 | 0.70 | 0.71 | 0.70 |
| Logistic Regression | 65 | 35 | 0.62 | 0.64 | 0.62 | 69 | 31 | 0.69 | 0.68 | 0.67 |
| J48 | 67 | 33 | 0.67 | 0.68 | 0.66 | 69 | 31 | 0.69 | 0.68 | 0.67 |
| RF | 80 | 20 | 0.80 | 0.80 | 0.80 | 82 | 18 | 0.82 | 0.81 | 0.81 |

Table 5: Results for the different classifiers with and without feature selection(Weather)

| Algorithms | Without Feature Selection | | | | | With Feature Selection | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure | True Positive Rate (%) | False Positive Rate (%) | Precision | Recall | Fmeasure |
| Bayes Net | 57 | 43 | 0.57 | 0.57 | 0.56 | 58 | 42 | 0.52 | 0.57 | 0.53 |
| Naïve Bayes | 57 | 42 | 0.57 | 0.57 | 0.56 | 57 | 42 | 0.57 | 0.57 | 0.56 |
| MLP | 77 | 23 | 0.77 | 0.77 | 0.77 | 79 | 21 | 0.80 | 0.79 | 0.79 |
| Logistic Regression | 71 | 29 | 0.71 | 0.71 | 0.71 | 71 | 29 | 0.71 | 0.71 | 0.71 |
| J48 | 70 | 30 | 0.70 | 0.70 | 0.70 | 72 | 28 | 0.72 | 0.72 | 0.72 |
| RF | 57 | 42 | 0.57 | 0.57 | 0.57 | 57 | 42 | 0.57 | 0.57 | 0.57 |

From the results given in Table 2 to Table 5, accuracy of all classification algorithms is improved while feature selection method is applied. The true positive rate of these four classification algorithms with feature selection using four data sets are shown in Figure 2 to Figure 5. The figures compare the four classification algorithms and predict the more accuracy algorithm.Among these four classifiers, J48 is best classifier for Breast Cancer data set, MLP and J48 are best classifier for Iris data set, RF is the best classifier for Glass data set, and MLP is best classifier for Weather data set when feature selection is used.

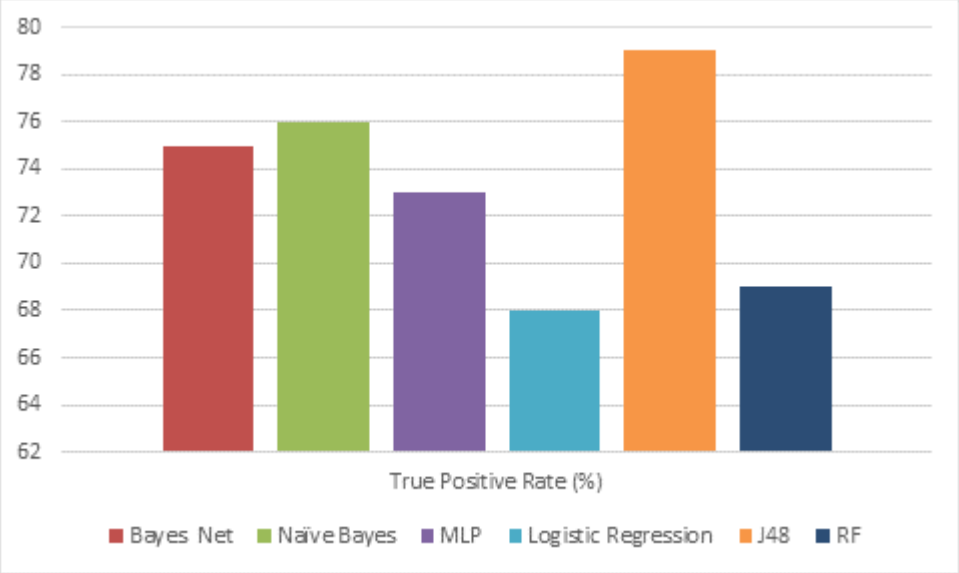


Figure 2: True Positive Rate of classification algorithms with feature selection using Breast Cancer Data set

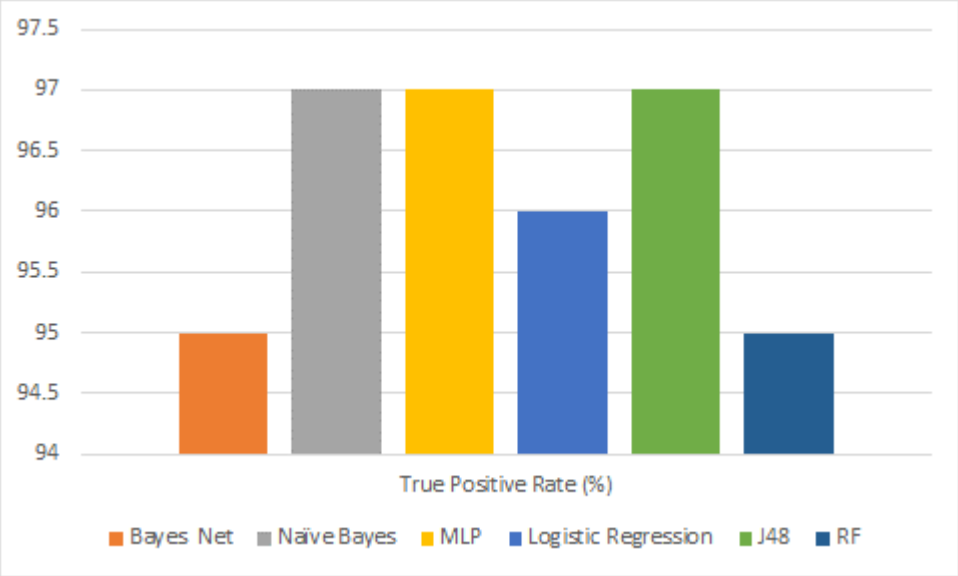


Figure 3: True Positive Rate of classification algorithms with feature selection using Iris Data set

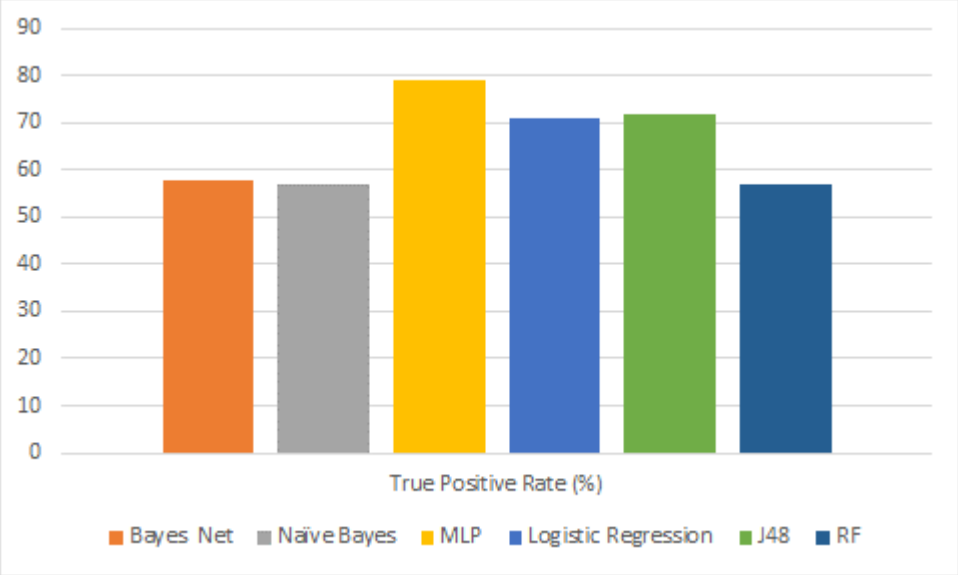


Figure 4: True Positive Rate of classification algorithms with feature selection using Glass Data set

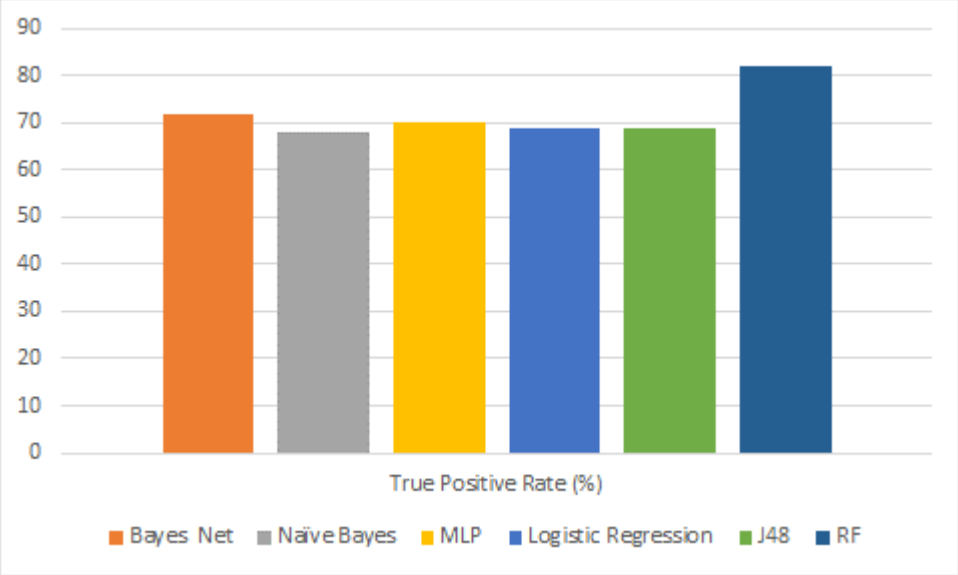


Figure 5: True Positive Rate of classification algorithms with feature selection using Weather Data set

## 6.CONCLUSIONS

Since Feature selection is played vital role for classification in the machine learning algorithms and irrelevant features affect accuracy of the algorithms, relevantfeatures are taken for the classification for better prediction.This research work investigates the performance of the different classifiers with features selection method using different data sets.From the results, it is observed that classifications with feature selection give better accuracy.Moreover, due to nature of the domain, number of attributes and number of instances in the dataset, each classifier produces better accuracy with respect to dataset. In this study, J48 produced more accuracy for Breast Cancer data set, MLP and J48 produced more accuracy for Iris data set, RF produced more accuracy for Glass data set, and MLP produced more accuracy for Weather data set. This paper helps the researcher to identify suitable classification algorithm for their data set. In future, different features selection methods can be applied in the classification algorithms to improve more accuracy.

## REFERENCES

1. N. Krishnaveni and V. Radha, "Feature Selection Algorithms for Data Mining Classification: A Survey",Indian Journal of Science and Technology, vol.12, issue 6, 2019, pp. 1-11.

2. S.N. Sragivi Vibhushan, and Vikas, "Evaluation Of Various Feature Selection Algorithms In Educational Data Mining",International Journal of Computational Engineering Research, vol. 8, issue 6, 2018, pp. 225-230.
3. F. Francis, "Feature Selection and Classifier Accuracy of Data Mining Algorithms", International Research Journal of Engineering and Technology, vol. 5, issue 11, 2018, pp. 1280-1283.
4. Maryam Zaffarm, K.S. Savita and MA Hashmani, "A study of feature selection algorithms for predicting students academic performance", International Journal of Advanced Computer Science and Applications, vol. 9, issue 5, 2018, pp. 541-549.
5. S. Alelyani, J. Tang, and H. Liu. "Feature selection for clustering: A review", DataClustering: Algorithms and Applications book,CRC Press, pp. 29-55, 2013.
6. C. Domeniconi and D. Gunopulos,"Local feature selection for classification", Computational Methods for feature selection, pp. 211-232, 2008.
7. J.G. Dy and C.E. Brodley,"Feature selection for unsupervised learning", The Journalof Machine Learning Research, vol. 5, pp.845–889, 2004.
8. I. Guyon and A. Elisseeff,"An introduction to variable and feature selection", TheJournal of Machine Learning Research, vol. 3, issue 1, 1157–1182, 2003.
9. H. Liu and H. Motoda,"Computational Methods of Feature Selection", Chapman and Hall/CRC Press, 2007.
10. H. Liu and L. Yu,"Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 17, issue 4, pp. 494-502, 2005.
11. P. Mitra, C. A. Murthy, and S. Pal,"Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, issue 3, pp. 301–312, 2002.
12. Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa , "A Comparison Study between Data Mining Tools over some Classification Methods", vol. 8, issue 2, pp. 18-26, 2011.
13. D.L. Gupta, A. K. Malviya and Satyendra Singh, "Performance analysis of Classification Tree Learning Algorithms", International Journal of Computer Applications, vol. 55, issue 6, pp. 39-44, 2012.
14. Nikhil N. Salvithal, Dr. R. B. Kulkarni, "Evaluating Performance of Data Mining Classification Algorithm in Weka",International Journal of Application or Innovation in Engineering & Management,vol. 2, issue 10, pp. 273-281,2013.
15. Rohit Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA",International Journal of Computer Applications, vol. 54, issue13, pp. 21-25, 2012.
16. Tina R. Patil, and Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science And Applications, vol. 6, issue 2, pp. 256-261, 2013.
17. D. R. S. . et. al., "ESTIMATING THE EFFICIENCY OF MACHINE LEARNING IN FORECASTING HARVESTING TIME OF RICE", IJMA, vol. 10, no. 2, pp. 1930 - 1937, Apr. 2021.
18. Asraf Yasmin, B., Latha, R., & Manikandan, R. (2019). Implementation of Affective Knowledge for any Geo Location Based on Emotional Intelligence using GPS. International Journal of Innovative Technology and Exploring Engineering, 8(11S), 764–769. https://doi.org/10.35940/ijitee.k1134.09811s19
19. Muruganantham Ponnusamy, Dr. A. Senthilkumar, & Dr.R.Manikandan. (2021). Detection of Selfish Nodes Through Reputation Model In Mobile Adhoc Network - MANET. Turkish Journal of Computer and Mathematics Education, 12(9), 2404–2410. https://turcomat.org/index.php/turkbilmat/article/view/3720