iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4785

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

A Comparative Study of Classification Algorithms on Spam Detection

G V Gayathri[1], B.Siva Jyothi[2]
[1, 2]*Assistant Professor, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam-531162, India*

Abstract: ***We are going to explore on how the elaboration of text messages in all communication and transactions which are all taking place through web based tool known as email. Spam is an alternative way of an electronic messaging system to send a large number of messages to the user inbox. Here we are going to experiment many data mining techniques to the dataset of spam in an attempt to search the most suitable classifier to text message classification as spam and non-spam. Here we are going to check the performance of many classifiers with the use of feature selection algorithm and we found that in the result analysis part the Random Forest classifier provides finer accuracy with respect to other four classifiers such as Naive Bayes, Support vector machine(SVM), Logistic Model and Decision Tree***
Keywords: Classifier, Random Forest, Logistic Model, Feature selection, text message, Spam mails.

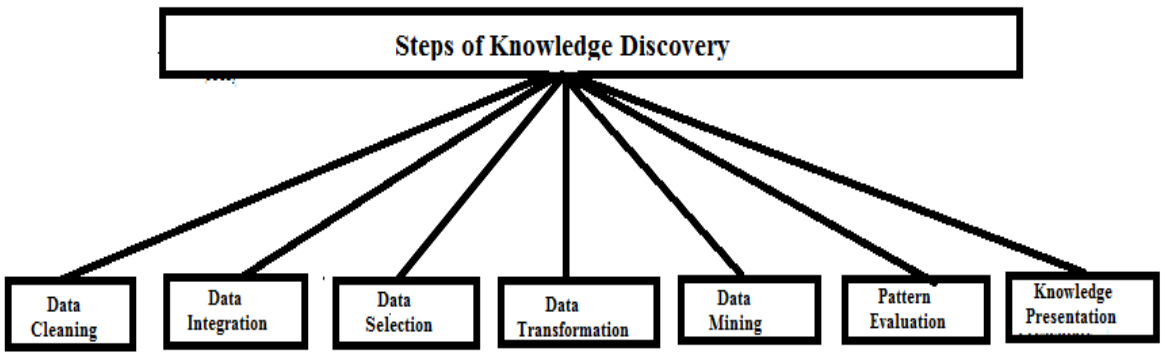
I. INTRODUCTION

A. *Data Mining*

Data mining is described as the technique of fetching the data from very large data sets or it is the art of mining or collecting the knowledge from the data the fetched data can be used for exploration of science and control of production, retention of customer and detection of fraud and for analysis of market .It is combination of fields such as database systems, statics, machine learning and artificial intelligence. The main motto behind data mining is to fetch all related facts from the very big data set and converting it to an intended meaning such that it can be used further. We analyze the data by different perspectives using data mining and is combined or grouped to the helpful information.

Data mining is collection of the information such as business transactions data related to scientific calculations, medical data, details, personal data, satellite sensing digital media, virtual worlds, and text related all information and email messages, which may also contain video or surveillance information and picture related data. Data mining consists of checking of data which has been stored in the data warehouse .The important methods of data mining involve regression, classification and clustering.

Data mining is referred as knowledge discovery in database, and methods of data mining includes:



1) *Cleaning of Data:* It is the step in which the data which is not relevant and noisy data is discarded by the collection.
2) *Integration of Data:* In this step the data from the multiple sources such as heterogeneous aggregated to a single source.
3) *Selection of Data:* In this process the data which is suitable for the analysis is taken into consideration and extracted from the data collection.
4) *Transformation of Data:* This step is referred as consolidation of data and in which the data is converted to most suitable structure for the mining process.

5) *Data mining:* It is critical process in which intelligent methods are used in an order to fetch patterns which are useful.
6) *Evaluation in Pattern:* In this process the patterns which are interesting they have been based on the given measures identified.
7) *Representation of knowledge:* It is the last step and in which the knowledge which has been discovered illustrated before the user.

B. Machine Learning Techniques

Machine learning is a study of neural network and it gives the capacity for computers to study and provides brief description of the program being learned and it concentrates on the progress of computer programs that will result itself to originate and modify when applied to the new data.

The machine learning technique is very much similar to the data mining technique. In machine learning without fetching data from large data set here we are going to utilize data to recognize patterns of data and thereby sets working of program appropriately.

II. LITERATURE SURVEY

A. Introduction to Spam

"Electronic spamming" is the use of electronic messaging systems to send an unwanted message especially by advertising, as well as sending messages repeatedly on the same site. While the most widely known form of spam is email spam, the term is applied to similar abuses in other media: messaging spam or instant messaging spam, Newsgroup spam or Usenet newsgroup spam, spamdexing or Web search engine spam, spam in blogs, wiki spam, online classified ads spam, mobile phone messaging spam, Internet forum spam, junk fax transmissions, social spam.

Spamming remains economically viable because advertisers have no operating costs beyond the management of their mailing lists, servers, infrastructures, IP ranges, and domain names, and it is difficult to hold senders accountable for their mass mailings.

B. Testing a Classification Model

A classification model is tested by applying it to test data with known target values and comparing the predicted values with the known values.

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A huge number of the records are used to build the model; the remaining records are used to test the model as per the requirement or not.

Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data for future predictions.

C. Accuracy

Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data

D. Confusion Matrix

A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is n-by-n, where n is the number of classes.

Figure (a) shows a confusion matrix for a binary classification model. The rows present the number of actual classifications in the test data. The columns present the number of predicted classifications made by the model.

		PREDICTED CLASS	
		affinity_card = 1	affinity_card = 0
ACTUAL CLASS	affinity_card = 1	516	25
	affinity_card = 0	10	725

Figure (a) Confusion Matrix for a Binary Classification Model

E. Naive Bayes Algorithm

Naive Bayes is a technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is a family of algorithms based on a common principle all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

1) Naive Bayes Classifier: Naive Bayes[1][2] is a classifier which uses the Bayes Theorem. It can forecast membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class.

MAP(H)

=max(P(H|E))

=max((P(E|H)*P(H))/P(E))

= max(P(E|H)*P(H))

P(E) is evidence probability, and it is used to normalize the result.. Naive Bayes classifier assumes that all the features are **irrelevant** to each other. A feature does not influence the presence or absence of any other feature.

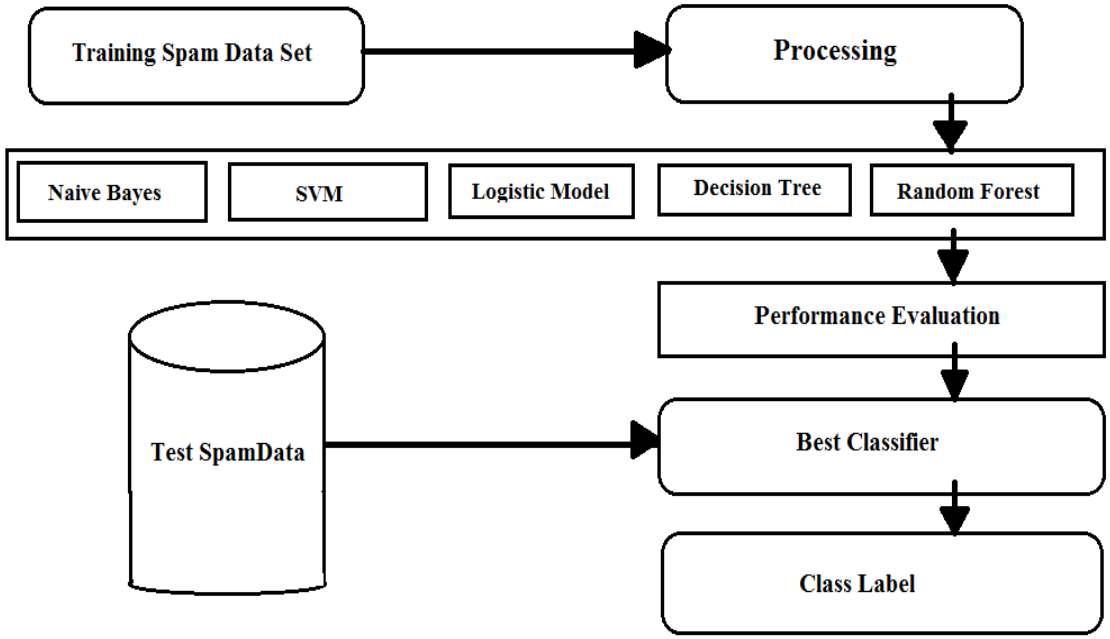
F. Support Vector Machine (Svm)

Support Vector Machine (SVM) is a classifier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1. Thus, a categorical dependent variable consisting of three levels, say (A, B, C), is represented by a set of three dummy variables:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

III. PROPOSED TECHNIQUES

In the present system[3] we used logistic model, decision tree and Random Forest algorithms and improved accuracy. We compared the accuracy of existing classifiers with the proposed classifiers.



Processing Steps

1) *Step 1:* In the first step the unprocessed data is taken and data set is created by processing the using data pre-processing techniques. A number of data sets are trained and rest are used for testing using the trained data sets.
2) *Step 2:* After receiving the data set, it is given for training to trained data set.
3) *Step 3:* Data set is tokenized using word tokenize and part of speech tagging is done.
4) *Step 4:* Data set processed is sent to different classifiers and accuracy is calculated. Data sets are given all the five classifiers and the accuracy is calculated for different data sets. In this paper, data sets of different sizes are tested to calculate the accuracy of each classifier.
5) *Step 5:* Based on the test data input the best classifier is decided.

A. Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. When the outcome variable is categorical, then it comes under a different case of logistic regression, where we are using log of odds as dependent variable. It figures the probability of occurrence of an event by fitting data to a logistic function.

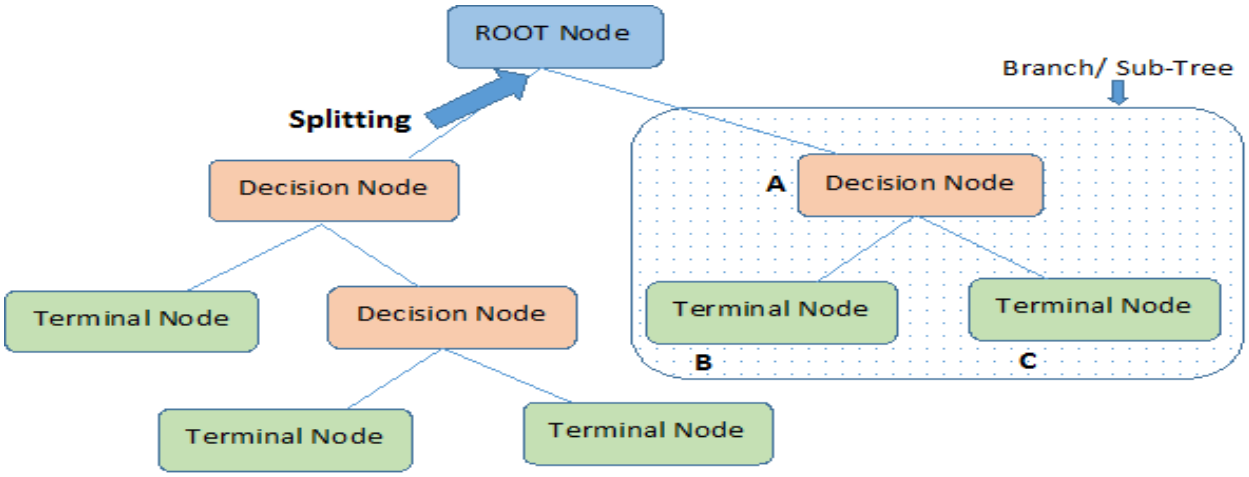
1) Logistic Regression Working

y = e^(a0 + a1*x) / (1 + e^(a0 + a1*x))

y is the predicted output, a0 is the bias or intercept term and a1 is the coefficient for the single input value (x). Each column in your input data will have an associated a coefficient that must be learned from your training data.

B. Decision Tree Algorithm

It is a supervised learning algorithm [8] (having a pre-defined target variable) that is used in classification problems. It works for categorical and continuous input and output variables. In this technique, we divide the population or sample into two or more sub-populations based on most significant characteristic in input variables.



C. Random Forest

Random forests [87] 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 the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their . Regression and classification problems can be solved using this classifier . In regression problems, the dependent variable is continuous. In this case, the dependent variable is categorical.

1) *Working of Random Forest:* It uses the Bootstrap algorithm to create random samples. For a given data set DS1 (n rows and m columns), we create a new dataset (DS2) by sampling n cases at random with restoration from the original data. About 1/3 of the rows from DS1 are left out, known as Out of Bag (OOB) samples. Then, the model trains on DS2. OOB sample is used to determine fair estimation of the error. Out of m columns, M << m columns are selected at each node in the data set randomly.

The default choice of M is m/3 for regression tree and P is sqrt(p) for classification tree. Pruning does not takes place in random forest; i.e, each tree is grown fully. In decision trees, pruning is a method to avoid over-fitting. Pruning means selecting a subtree that leads to the lowest test error rate. We can also use cross validation to determine the test error rate of a sub tree. Several trees are grown and the final prediction is obtained by averaging or voting.

IV. PERFORMANCE ANALYSIS AND RESULTS

We used five different classifiers for spam detection. Initially once the preprocessing is done data sets are obtained. The obtained data sets are given as inout to classifiers for spam mail detection. At the initial stage teo different data sets are classified and accuracy is given in the following table :

Data sets	Naïve Bayes	SVM	Logistic Model	Decision Tree	Ramdom Forest
Data set 1	20.46%	96.27%	96.34%	91.46%	97.2%
Data set 2	28.06%	96.12%	95.97%	91.29%	97.27%

Table (1) : Accuracy of classifiers on different data sets

In the next step classifiers are used on different sizes of data sets. Validation is done among different data sets using the classifiers

TEST NAME	NAÏVE BAYES	SVM	DECISION TREE	RANDOM FOREST	LOGISTIC REGRESSION
TEST 1	12.2	98.00	88.4	97.6	90
TEST 2	14	97.6	89.2	97.14	88.8
TEST 3	28.06	96.12	91.29	97.27	95.97
TEST 4	33.71	87.84	93.4	98.45	97.44
AVERAGE	21.99	94.89	90.57	97.615	93.0525

Table (2) : Comparison of classifiers on different data sets

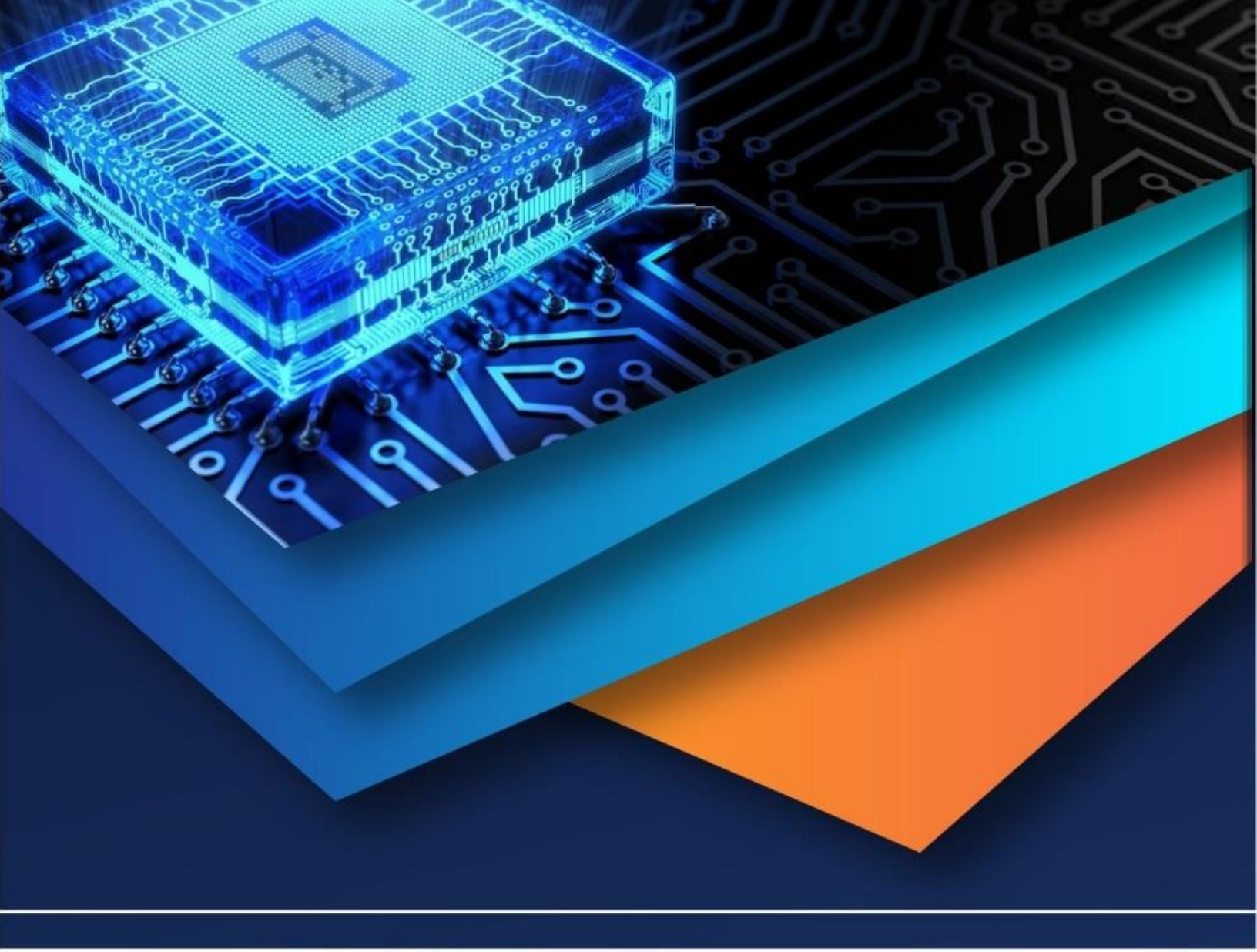
V. CONCLUSION & FUTURE WORK

Using the machine learning algorithms the performance analysis in case of random forest is highest that is 97.27% and in case of logistic model it is 96.12% and in svm it is 96.27% and in Decision tree it is 91.46% and for naïve it is very low with 20.46%. By the above results it signifies that random forest is the best classifier algorithm among other algorithms.

In the future work using many various other algorithms there is a possibility of getting best accurate one. We can also use the classifiers which are much more highly accurate than which we have used in our project to get good performance result in classifying the spam.

REFERENCES

[1] International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 DOI : 10.5121/ijcsit.2011.3112 173 MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION
[2] Spam Filtering based on Naive Bayes Classification Tianhao Sun May 1, 2009.
[3] Combining Classifiers for Spam Detection Fatiha Barigou, Naouel Barigou, and Baghdad Atmani Computer Science Laboratory of Oran Computer Science department, Faculty of Science, University of Oran BP 1524, El M'Naouer, Es Senia, 31000 Oran, Algeria.
[4] International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 11 – No. 11, March 2017 – www.ijais.org 16 A Mapping Study to Investigate Spam Detection on Social Networks
[5] Survey on Web Spam Detection: Principles and Algorithms Nikita Spirin and Jiawei Han.
[6] Study of machine learning classifiers for spam detection, 2016 4th International Symposium on Computational and Business Intelligence (ISCBI).
[7] International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com Random Forest: A Review
[8] A Survey of Decision Tree Classifier Methodology S. Rasoul Safavian and David Landgrebe



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY