iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5652

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: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Study on Feature Selection in Data Mining

Yogita Borole[1], Amruta Raut[2], Tejashree Jadhav[3], Madhavi Pradhan[4]
[1,2,3,4]*Department of Computer Engineering, Savitribai Phule Pune University*

Abstract: Feature selection is an important term in machine learning tasks as it can efficiently improve the performance of the model by eliminating the redundant and irrelevant attributes. Feature selection not only improves the quality of the model, it also makes the process of modeling more efficient. Due to irrelevant and huge dimensions data the quality of the model may degrade. This paper presents the importance of feature selection on various classification algorithms. In this study, the feature selection techniques like attribute evaluator and the best first search is used for reducing the number of features. The dimension is reduced from eight to four. The dataset used is Pima diabetic dataset from UCI repository. The substantial increase is given in terms of accuracy.

Keywords Machine Learning; Feature Selection; Naive Bayes Algorithm; Decision Tree

I. INTRODUCTION

Feature selection is very important term in data mining. The purpose of the feature selection is to decide which features should be included in the final subset..Feature selection not only improves the quality of the model, it also makes the process of modeling more efficient. Due to irrelevant and huge dimensions data the quality of the model may degrade. Feature selection techniques can be classified based on many terms. A feature selection method that is only based on the basic characteristics of the data is called as filter. Whereas the method which evaluates the features based on accuracy is called as wrapper.

Feature selection results in solving the problems: having too long data of little value, or having less data with high value. The goal of feature selection is to identify the minimum number of attributes that are efficient in building the model.

There are various feature selection techniques. Feature selection is further divided into two parts:

A. Attribute Evaluator

B. Search Method

In attribute evaluator each attribute is evaluated based on the class. In search method different combinations of attributes are evaluated in order to get minimum list of selected features. Some attribute evaluator techniques also require some search methods. The methods used to apply feature selection are:

1) *Correlation Based Feature Selection:* In order to select the most relevant attributes, correlation technique is used. The correlation between each attribute and the output variable can be evaluated by selecting only those attributes which have moderate-to-high positive or negative correlation and dropping those attributes with low correlation.
2) *Information Gain Based Feature Selection:* Information gain is used to measure the purity of an attribute. The attribute which give more information will have high information gain value and these attributes are selected, whereas the attributes which give less information will have low information gain value and these attributes are removed.
3) *Learner Based Feature Selection:* In this technique the subsets of attributes that contribute in the best performance are selected. The algorithm used for evaluating the subsets should be quick to train and powerful.

II. NORMALIZATION

Normalization involves scaling all the values of a given attribute so that they fall within a small specified range. It is required to avoid giving the undue significance to the attributes having large range.

III. DISCRETIZATION

Discretization is the process of transforming continuous valued attributes to nominal. It is used as a preprocessing step for the correlation-based approach to feature selection.

IV. ALGORITHMS

For doing comparative study of different machine learning algorithm we have used following algorithm:

A. Decision Tree(J48)

A Decision Tree represents rules and it is very popular tool for classification and prediction. J48 algorithm uses divide and conquer strategy for creation of decision tree using greedy algorithm. It creates tree with the help of recursion. A DT contains leaf node which is class label and decision nodes.

B. Naive Bayes

The naïve bayes classification algorithm is a probabilistic classifier. Naïve Bayes Classifier uses Bayes theorem for prediction of class label of instance.

$$\text{Bayes Theorem: } P(x|Y) = \frac{P(Y|x)*p(x)}{P(Y)}$$

V. EXPERIMENTAL RESULTS

The Pima Indian Diabetes dataset contains total 8 attributes. By applying feature selection techniques it selects 4 most predictive attributes :glucose_concentration, blood_pressure, BMI and age.

Dataset: For our system we have used Pima Indian Diabetes dataset, which is collected from UCI repository. Datasets are analyzed under different classification techniques. All information about the dataset is given in following table.

Dataset	Instance	Attributes
Pima Indian Diabetes	768	9

Table1.Total Instances

Decision tree(J48)	Accuracy %	True Positive Rate %	False Positive Rate %	Error %
Without Pre-processing	77	80.3	29.4	24.21
After Discretization	75	89	58	25
After attribute selection	76.78	90	39.9	24.21

Table2. Performance table for decision tree(J48)

Naive Bayes	Accuracy %	True Positive Rate %	False Positive Rate %	Error %
Without Pre-processing	77	83.3	35.3	23
After Discretization	75.3	82	37	23
After attribute selection	77.5	84.2	38	22

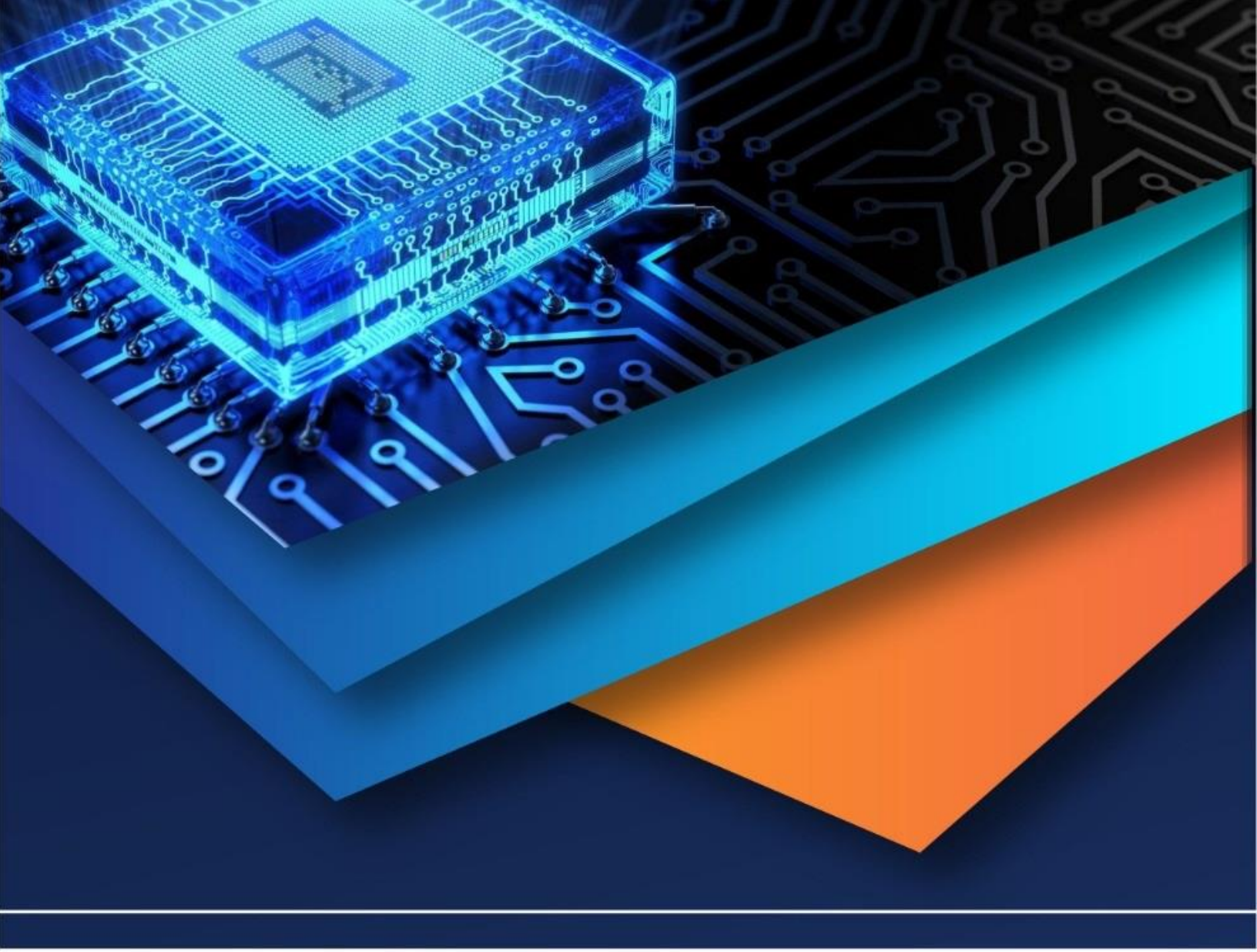
Table3. Performance table for Naïve bayes

VI. CONCLUSION

In this paper, we studied the importance of pre-processing of data. Very small set of non-redundant features are obtained with an increasing accuracy. From the results we can see that because of pre-processing the accuracy values we have obtained is higher. The performance of classification algorithm is improved by the removal of irrelevant features.

REFERENCES

[1] SeemaSharma ,JitendraAgrawal, ShikhaAgarwal.”Machine Learning Techniques for Data Mining: A Survey”978-1-4799-1597-2/13/$31.00 ©2013 IEEE
[2] Huang, Y., McCullagh, P., Black, N., Harper, R.: Feature selection and classification model construction on type 2 diabetic patients’ data. Artificial Intelligencte Medicine Journal 41, 251–262 (2007)
[3] Bolón-Canedoa, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, B.F.: A review of microarray datasets and applied feature selection methods. Information Sciences 282, 111–135 (2014)
[4] MessanKomi, Jun Ki, Y ongxinZhai, Xianguo Zhang, “Application of Data Mining Methods in Diabetes Prediction” 978-1-5090-6238-6/17/$31.00 ©20 17 IEEE
[5] Guo, Yang, GuohuaBai, and Yan Hu."Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 InternationalConferece For, pp. 471-472. IEEE, 2012.
[6] J. Han and M. kambar, Data Mining:Concepts and Techniques,Morgan Kaufman Publishers,(2004)
[7] MadhaviPradhan and G.R. Bamnote."Efficient Binary Classifier for Prediction of Diabetes Using Data Preprocessing and Support Vector Machine."



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY