IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Analysis of Different Classifiers' Performance After Applying Three Different Feature Selection Methods

Kasturi Ghosh[1], Susmita Nandi[2*]

[1]Dept. of IT, University Institute of Technology, The University of Burdwan, Burdwan, India
[2]Dept. of CSE, University Institute of Technology, The University of Burdwan, Burdwan, India

[*]*Corresponding Author: sus1622.nandi@gmail.com, Tel.: +918327371301*

Available online at: www.ijcseonline.org

***Abstract*—**Feature selection (FS) is an important aspect of data mining. Now a days availability of information with hundreds of variables leads to high dimensional, irrelevant and redundant data. Thus FS techniques must be applied on the datasets before classification or rule generation. It basically aims at reducing the number of attributes by removing irrelevant or redundant ones, while trying to reduce computation time and improve performance of classifiers. In this paper three different FS methods are used, Correlation Based, Information Gain Based and Rough set Based FS method. A statistical analysis of three different classifier's performance is also done in order to provide a detailed view.

***Keywords*—**Data Mining (DM), Feature Selection (FS), Rough Set, Degree of Dependency, Decision Tree (J48 algorithm), Naive Bayes Algorithm (NB), K-Nearest Neighbor Algorithm (KNN), Classification, Statistical Analysis.

I. INTRODUCTION

In medical domain researches DM techniques has great impact for exploring the patterns hidden in the datasets. These patterns can be utilized for clinical diagnosis. Preprocessing techniques involve different processes like data cleaning, data integration, data transformation and data reduction. These processes can significantly improve the quality of the information and also it reduces the time required for the mining process. Data preprocessing is a significant step in the knowledge discovery process, to standardize the quality of data. FS is an important step of data preprocessing [1][2]. The aim of FS process is to find out a minimum set of features to reduce dimensionality of datasets. Applying mining process on the reduced set of features has several benefits. It reduces the number of attributes present in the patterns, which reduces the complexity of the patterns. Further it enhances the classification accuracy. In this paper different FS algorithms are applied on seven benchmark datasets. Then a detailed statistical analysis of three different classifiers' performance is done. Here three different feature selection methods are used: Correlation Based, Information Gain Based and Rough Set Based Feature Selection algorithms.

II. CORRELATION BASED FEATURE SELECTION

There exist broadly two types of measures for the correlation between two random variables: linear and non-linear. Of linear correlation, the most well-known measure is linear correlation coefficient.

For a pair of variables (X, Y), the linear correlation coefficient ρ is given by,

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

Where, x_i is the mean of X, and y_i is the mean of Y. The value of ρ lies between -1 and 1.

If X and Y are completely correlated, ρ takes the value of 1 or -1; if X and Y are independent, ρ is zero. It is a symmetrical measure for two variables. Other measures in this category are basically variations of the above formula, such as least square regression error and maximal information compression index. However, linear correlation measures may not be able to find out the correlations that are not linear in nature. It can also be observed that linear correlation coefficient is not suitable for nominal data. A good feature subset is the one that contains features which are highly correlated with the class and less correlated or uncorrelated with each other. Using the linear correlation coefficient formula all features of the datasets are ranked and then 50% features from the whole set of attributes are selected for evaluation.

III. INFORMATION GAIN BASED FEATURE SELECTION

Several non-linear correlation measures are based on the concept of entropy, which is a measure of the uncertainty of a random variable. The entropy of a variable A is defined as

$$H(A) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (2)$$

and the entropy of A after observing values of another variable Y is defined as

$$H(A|Y) = -\sum_i P(y_i)\sum_i P(x_i|y_i)\log_2(P(x_i|y_i)) \quad (3)$$

Where $P(x_i)$ is the prior probability for the values of A, and $P(x_i|y_i)$ is the posterior probability of A given the values of B. The amount by which the entropy of A decreases provides another information about A provided by B and it is called information gain, which is given by

$$IG(A|B) = H(A) - H(A|B) \quad (4)$$

Attribute B is considered more correlated to A than C, if IG(A|B) > IG(C|B). Information gain is a symmetrical measure which means the ordering of the attributes will not affect the final result. Information gain tends to favor multivalued attributes, so it should be normalized. Therefore, symmetrical uncertainty is chosen which can be defined as

$$SU(A,B) = 2\left[\frac{IG(A,B)}{H(A)+H(B)}\right] \quad (5)$$

It compensates for information gain's bias towards multivalued attributes and restricts its values to the range [0, 1]. The value 1 represents that knowing the values of either attribute completely predicts the values of the other; a value of 0 indicates that the features are independent. Information Gain based measures handle nominal or discrete attributes, and therefore continuous features need to be discretized first in order to use this measures.

C-correlation- The correlation between any feature $F_i$and the class C is called C-correlation, denoted by $SU_{i,c}$.

F-correlation- The correlation between any pair of features F_i and F_j ($i \neq j$) is called F-correlation, denoted by $SU_{i,j}$.

Approximate Markov blanket- For two relevant features F_i and F_j \$($i \neq j$), F_j forms an approximate Markov blanket for F_i if $SU_{j,c} \geq SU_{i,c}$ and $SU_{i,j} \geq SU_{i,c}$

Predominant feature- A relevant feature is predominant if it does not have any approximate Markov blanket in the current set.

There are two steps in the algorithm: For a data set S with N features and class C, the algorithm finds a set of predominant features S_{best} . In the first step, it calculates the SU value for each feature moves those features into S'_{list} and orders them in a descending order according to their SU values. In the second step, the ordered list S'_{list} is further processed to select predominant features. A feature F_j that has already been determined to be a predominant feature can always be used to filter out other features for which F_j forms an approximate Markov blanket. Since the attribute with the highest C-correlation does not have any approximate Markov blanket, it must be one of the predominant features. So the iteration starts from the first element in $S'_{list}$and continues as follows. For all the remaining features (from the one right next to F_j to the last one in S'_{list}, if F_j happens to form an approximate Markov blanket for F_i, F_i will be removed from S'_{list}. After one round of filtering features based on F_j , the algorithm will take the remaining feature right next to F_j as the new reference to repeat the filtering process. The algorithm stops when all predominant features are evaluated.

Algorithm: Feature Selection Using Information Gain (FSUIG)

Input: S(F_1 , F_2 ,…, F_n ,C) (A training Dataset)
Output: S_{best} (Selected Subset)

```
begin
        for i=1 to N do begin
                calculate SU_{i,c} for F_i
                append F_i to S'_list;
        end

        order S'_list in descending SU_{i,c} value;
        F_j=getFirstElement(S'_list);

        do begin
                F_i=getNextElement(S'_list, F_j);
                if ( F_i ≠ NULL)
                   do begin
                        if ( SU_{j,c} ≥ SU_{i,c})
                                remove F_i from S'_list
                        F_i=getNextElement(S'_list, F_i);
                   End until ( F_i= = NULL)
                F_j=getNextElement(S'_list, F_j);
        End until ( F_j= = NULL)
S_best= S'_list
end
```


IV. ROUGH SET BASED FEATURE SELECTION

Rough set is a new mathematical approach to deal with imperfect knowledge. The main goal of the rough set theory is induction of approximations. It offers mathematical tools to discover patterns which are hidden within data. Like fuzzy set theory it is not an alternative approach to traditional set theory, it is embedded in it. Imprecision in this approach is expressed by a boundary, and not by the partial membership, like fuzzy set. Rough set concepts can be defined by means of topological operations which are known as approximations [3][4]. Rough set theory has many interesting applications. The rough set approaches are especially used in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, pattern recognition.

Upper and Lower Approximation- Suppose, I^n is an information system, $I^n = (U^n, Y)$,where U^n is a non empty finite set of objects, known as universe and Y is a set of features S.T. a:u$\rightarrow Z_a$for every a, Z_a is the set of values that attribute a may take. For any Q∈Y, there is an associated equivalence relation IND(Q).

$$\mathrm{IND} = \{(S, S') \in U^n / \forall a(s) = a(s')\} \quad (6)$$

The portion of U^ngenerated by this equivalence relation is denoted by U^n/Q

$$U^n/Q = X\{U^n \mathrm{IND}(a) \forall a \in Q\} \quad (7)$$

Where X for set A and B can be defined as

$$AXB = \{(H \cap H')\forall(H \in A)\text{and } (H' \in B), H \cap H' \neq \emptyset\} \quad (8)$$

If (s,s') ∈ IND(Q) then they are indiscernible by the features of Q. $[S]_Q$ is the the equivalence classes from Q-indiscernibility relation.
Let, K⊆ U^n. K can be approximated by Q-Lower and Q-Upper approximations:

$$Q_*K = \{s \in U^n | [S]_Q \subseteq K\} \quad (9)$$
$$Q^*K = \{s \in U^n | [S]_Q \cap K \neq \emptyset\} \quad (10)$$

Degree of dependency- Let, Q and R be sets of attributes inducing equivalent relations over U^n then the positive, negative and boundary regions can be defined as:

$$\mathrm{POS}_Q(R) = \cup_{K \in U^n/R}\ Q_*K \quad (11)$$

$$\mathrm{NEG}_Q(R) = U^n - \cup_{K \in U^n/R}\ Q^*K \quad (12)$$

$$\mathrm{BND}_Q(R) = \mathrm{NEG}_Q(R) - \mathrm{POS}_Q(R) \quad (13)$$

The positive region consists of all objects of the universe that can be classified into classes of U^n/Q.

Discovering dependencies between attributes is very important to determine in data analysis. A set of attributes B depends totally on a set of attributes A, which are denoted A→B, if all attribute values from B can be uniquely determined by the attributes from A. If there exists a functional dependency between B and A, then B depends totally on A. In rough set theory, dependency is explained as follows:
For A,B ⊂ Y, it is said that B depends on A in a degree n (0≤ n ≤ 1) if

$$n = \text{degree of dependency}(\gamma) = \frac{\mathrm{POS}_A(B)}{|U^n|} \quad (14)$$

Algorithm: Feature Selection Using Degree of Dependency (FSUDD)
Input: All Conditional Attributes of Dataset (α), Decision Attribute (β)
Output: Selected Subset of Features (σ→{ })

do $\tau \leftarrow \sigma$
  for each $x_i \in (\alpha - \sigma)$
    if $\gamma_{\sigma \cup x_i}(\beta) > \gamma_\tau(\beta)$
      $\tau \leftarrow \sigma \cup \{x_i\}$
      $\sigma \leftarrow \tau$
  Until $\gamma_\sigma(\beta) = \gamma_\alpha(\beta)$
Return σ

In each iteration a feature is added to the set of features σ and it is checked whether the degree of dependency has been increased after adding that new feature. The iteration stops when adding more features does not increase the degree of dependency.

V. CLASSIFIERS: J48, KNN, NB

Classification is a process which assigns class labels to a set of data in order to achieve more accurate predictions and analysis [5][6]. The goal of classifiers is to create a set of classification rules that will predict the class labels of objects. To start the classification process, a set of training data is developed for which the likely outcome of class labels is already known. The aim of the classification algorithm is to find out how that set of features reaches its conclusion and develop a model accordingly. Based on this model it can predict the class label of objects in future. Here three popular classifiers are used: Decision Tree (J48), Naive Bayes (NB) and k-Nearest Neighbor (KNN).

A. **Decision Tree J48 Classifier-** Decision tree is a popular tool for classification and prediction in DM. It is a tree

structure similar to a flowchart, where the internal nodes represent a test on a feature, and the branches represent outcome of the test. The leaf nodes hold class labels. A deterministic decision tree, can be mapped into a set of rules, with each leaf of the tree corresponding to a rule [7]. A decision tree is constructed by splitting the source set of features into smaller subsets based on a test. This process is then continued on each subset in a recursive manner. This process is known as recursive partitioning. This process is continued until the subset of features at a node has the same class value, or when partitioning process no longer adds value to the predictions. This decision tree construction process does not require any parameter setting, so this process is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. The ID3 algorithm is a very simple and popular decision tree algorithm which uses information gain as splitting criteria. One limitation of this algorithm is that it is biased towards features with large numbers of unique values. J48 is an improved version of ID3 and it uses gain ratio as splitting criteria and overcomes the bias which ID3 has towards attributes with large number of unique values [8] [9] Gain ratio, is defined as follows:

$$\text{GainRatio (S,T)} = \frac{\text{Gain(S,T)}}{\text{SplitInfo(S,T)}} \quad (15)$$

$$\text{SplitInfo (S,test)} = -\sum_{j=1}^{n} S'\left(\frac{j}{k}\right) \times \log\left(S'\left(\frac{j}{k}\right)\right) \quad (16)$$

$S'\left(\frac{j}{k}\right)$ - Part of elements existing at the position k, taking the value of j-th test. At each node, decision tree J48 chooses one feature that most fruitfully splits the set of features into smaller subsets. The feature with the largest ratio is chosen for splitting purpose. Then the algorithm repeats the same procedure on the smaller subset of features.

B. **k-Nearest Neighbors Classifier-** k-nearest neighbors classifier is a simple yet popular classification algorithm that classifies new objects based on a similarity measure [10]. KNN has been used in pattern recognition in 1970's as a non-parametric method. This is a non-parametric, lazy learning algorithm. It uses a database in which the objects are separated into several classes in order to predict the classification of a new object. KNN is non-parametric since it does not make any assumptions on the underlying data distribution. It is also known as a lazy algorithm since it does not use the training data to do any generalization. There is no explicit training phase. The training data is needed during the testing phase for similarity measure. KNN Algorithm is based on attribute similarity that means how closely a sample feature from test set resembles the training set determines how the sample object is classified. In WEKA this k-nearest neighbors classifier is known as IBK. This algorithm usually uses the Euclidean or the Manhattan distance as a measure of similarity. In this experiment, Euclidean distance is used to determine the similarity. Suppose, the sample object has coordinates (p, q) and the coordinate of the object from training set is (r, s) then square Euclidean distance:

$$x^2 = (r-p)^2 + (s-q)^2 \quad (17)$$

C. **Naive Bayes classifier-** Naive Bayes is a probabilistic classifier and is based on the Bayes theorem [11]. It presumes that the features of a dataset are not dependent on each other. This assumption is known as class conditional independence. NB classifier generates probability approximations. For each unique class, the probability of an instance to belong to that class is approximated. Small amount of training data is required to get an approximate idea of the parameters needed for classification. This is an advantage of this classifier.

$$\text{Product rule:} \quad P(S) = P(S|Q)\,P(Q) = P(Q|S)P(S) \quad (18)$$

$$\text{Sum rule:} \quad P(S) = P(S)+P(Q)-P(S) \quad (19)$$

$$\text{Bayes theorem:} \quad P(j|K) = \frac{P(K|j)P(j)}{P(K)} \quad (20)$$

Theorem of total probability, if event Si is mutually exclusive and probability sum to 1.

$$P(Q) = \sum_{i=1}^{n} P(Q|Si)P(Si) \quad (21)$$

Given a hypothesis j and data K which bears on the hypothesis:

P(j): independent probability of j: prior probability
P(K): independent probability of K
P(K|j): conditional probability of K given j: likelihood
P(j|K): conditional probability of h given K: posterior probability.

Discretization- Discretization is a method that can transforms quantitative data into qualitative data. Quantitative data are commonly involved in DM applications. But many learning algorithms are designed primarily to handle qualitative data. Even for those algorithms which can deal with numerical data, learning is often less efficient. So to transform quantitative data into qualitative discretization is necessary. In this process the original data with continuous values are transformed by limiting them into a finite set of intervals and thus it simplifies the original data. The datasets used here contains continuous data. So, for evaluation of the performance of NB algorithm and for Information Gain Based feature selection they are discretized. Discretization techniques can be divided into two parts, unsupervised and supervised. Unsupervised methods use a systematic plan for discretization purpose without using the feature-class information and in the other hand supervised methods considers the feature-class information [12]. An issue of unsupervised methods is that it is hard to determine the number of intervals. Here Equal Interval Binning Method is used to discretize the datasets which is an unsupervised method.

Equal Interval Binning Method- This is the most straight-forward method for discretization. But outliers may dominate presentation. In this technique the range is divided into K intervals, each of equal size. If P and Q are the minimum and maximum values of the feature, the width of intervals will be:

$$W = \frac{Q-P}{K} \quad (22)$$

VI. STATISTICAL ANALYSIS

Statistical tests begin with a Null Hypothesis. In statistics, it is denoted as H_0 . It basically states that there is no pattern or no relationship between two measured phenomena. The null hypothesis is assumed to be true until it is proved to be wrong. Statistics gives precise criteria to reject a null hypothesis. Z Score is one of them and is commonly used to test statistical significance which helps to decide whether or not to reject the Null Hypothesis.

Let, x_1 and x_2 be the number of correctly classified instances from sample1 of size n_1 and sample2 of size n_2 respectively.

$$p_1 = \frac{x_1}{n_1} \text{ And } p_2 = \frac{x_2}{n_2} \quad (23)$$

$$p = \frac{p_1 * n_1 + p_2 * n_2}{n_1 + n_2} \quad (24)$$

$$\text{Standard Error (SE)} = \sqrt{p * (1-p) * (\frac{1}{n_1} + \frac{1}{n_2})} \quad (25)$$

$$Z = \frac{p_1 - p_2}{\text{Standard Error}} \quad (26)$$

Our intention is to prove that global accuracy of method 2 is better than method1. This frames our hypothesis as,
$H_0 : p_1 = p_2$ or $p_1 > p_2$
[Null hypothesis stating both are equal] Against
$H_1 : p_1 < p_2$
[Alternative Hypothesis claiming the later one is better]

The rejection region is given by,
$Z < -Z_\alpha$ [If true reject Null Hypothesis]
Here, Z_α is obtained from a standard normal distribution that pertains to a level of significance α. confidence level = $(1 - \alpha)$. If the test is a two tail test then the rejection region is given by, $Z < -Z_{\alpha/2}$. Also we have to find out cumulative probability (p-value) in order to discard null hypothesis. In this case a Normal Distribution calculator can be used. If p-value is less than the significance level we can't accept the null hypothesis. If the test is a two tail test then p-value is not equal to cumulative probability. In this case p-value = cumulative probability * 2. Here, confidence level = α = 95% so, significance level = 1-0.95 = 0.05. α/2 = 0.025.
$Z_{\alpha/2}$ = -1.96.

VII. DATASET DESCRIPTION

Here seven benchmark datasets are used: Diabetic Retinopathy Debrecen Dataset(DR), EEG Eye State Dataset(EEG), Cardiotocography Dataset(Cardio), Thoracic Surgery Dataset(TS), Pima Indians Diabetes Dataset(PIDD), Indian Liver Patient Dataset(ILPD), Breast Cancer Wisconsin Original Dataset(BCWD).

A. Diabetic Retinopathy Debrecen Data Set [13]

Table 1. Diabetic Retinopathy Debrecen Dataset Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	1151	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	20	**Date Donated**	2014-11-03
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	52254

B. EEG Eye State Data Set [14]

Table 2. EEG Eye State Dataset Description

Data Set Characteristics:	Multivariate, Sequential, Time-Series	**Number of Instances:**	14980	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	15	**Date Donated**	2013-06-10
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	66858

C. Cardiotocography Data Set [15]

Table 3. Cardiotocography Dataset Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	2126	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	23	**Date Donated**	2010-09-07
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	105278

D. Thoracic Surgery Data Set [16]

Table 4. Thoracic Surgery Dataset Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	470	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	17	**Date Donated**	2013-11-13
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	60100

E. Pima Indians Diabetes Data Set [17]

Table 5. Pima Indians Diabetes Dataset Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	1990-05-09
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	320972

F. Indian Liver Patient Data Set [18]

Table 6. Indian Liver Patient Dataset Description

Data Set Characteristics:	Multivariate	Number of Instances:	583	Area:	Life
Attribute Characteristics:	Integer, Real	Number of Attributes:	10	Date Donated	2012-05-21
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	63347

G. Breast Cancer Wisconsin Original Data Set [19]

Table 7. Breast Cancer Wisconsin Original Dataset Description

Data Set Characteristics:	Multivariate	Number of Instances:	699	Area:	Life
Attribute Characteristics:	Integer	Number of Attributes:	10	Date Donated	1992-07-15
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	300292

VIII. WORKFLOW OF THE EXPERIMENT

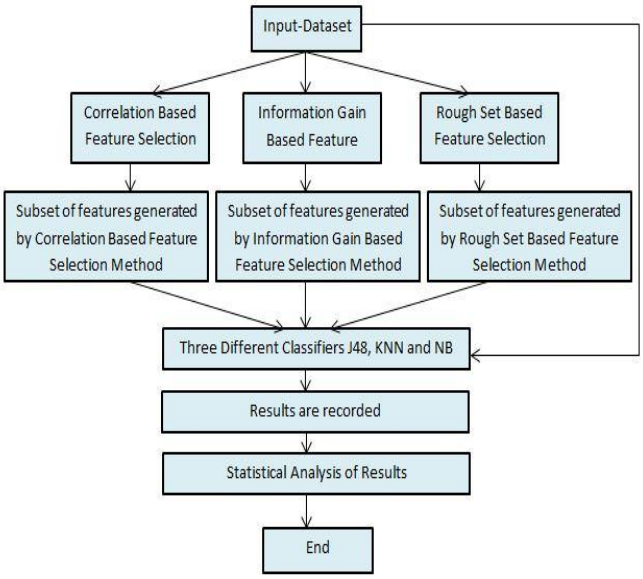

Figure 1. Workflow of the Experiment

Algorithm: Classifiers Performance Analysis Using Feature Selection (CPAUFS)-

Input: 7 Datasets (DR, Cardio, BCWD, EEG, TS, ILPD, PIDD)

Output: Performance of different classifiers based on Without Feature Selection (WFS) method and different feature selection methods (PC, FSUIG and FSUDD)

Step1: Subsets of features are selected from 7 datasets using Correlation Based Feature Selection method, Information Gain Based Feature Selection method and Rough Set Based Feature Selection method. (Equation (1) is used to implement Correlation Based Feature Selection method. Equation (2), (3), (4), (5) are used to implement Information Gain Based Feature Selection method. Equation (6-14) are used to implement Rough Set Based Feature Selection Method.

Step2: Performance of three classifiers (J48, KNN and NB) are enlisted for 7 datasets using WFS method. (Equation (15) and (16) are used to evaluate the performance of J48 classifier. Equation (17) is used to evaluate the performance of KNN classifier. Equation (18), (19), (20) and (21) are used to implement NB classifier. Equation (22) is used to implement Equal Interval Binning method to discretize datasets.)

Step3: Depending the the subset of features getting from step1, performance of the classifiers are again enlisted.

Step4: Output of step2 and step3 are compared for all 7 datasets separately.

Step5: Output of step4 is verified using 'z-test' statistical method. Equation (23), (24), (25) and (26) are used for statistical analysis.

IX. RESULTS

A. **Accuracy of the Classifiers for Different Datasets (Outputs of step4):** Figure(2) to figure(22) are representing the outputs of step4.

Diabetic Retinopathy Debrecen Dataset (DR)

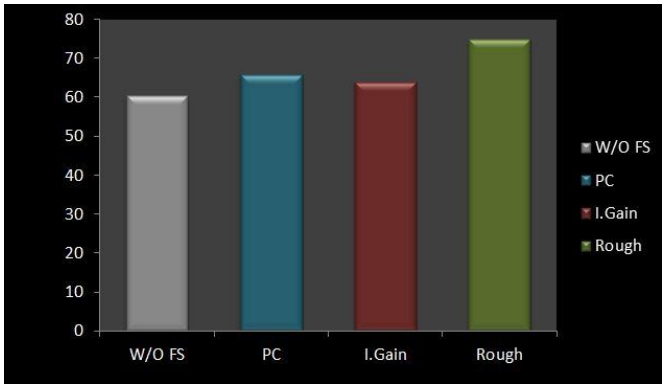

Figure 2. Accuracy of J48 classifier on DR Dataset before and after applying feature selection methods

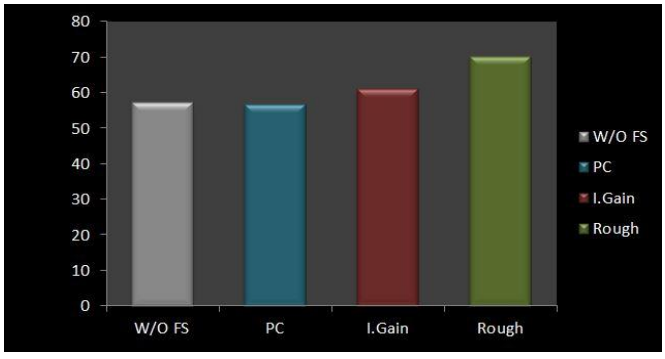

Figure 3. Accuracy of KNN classifier on DR Dataset before and after applying feature selection methods

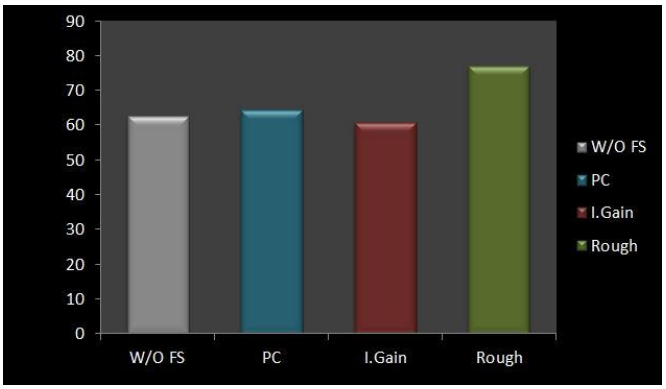

Figure 4. Accuracy of NB classifier on DR Dataset before and after applying feature selection methods

EEG Eye State Dataset (EEG)

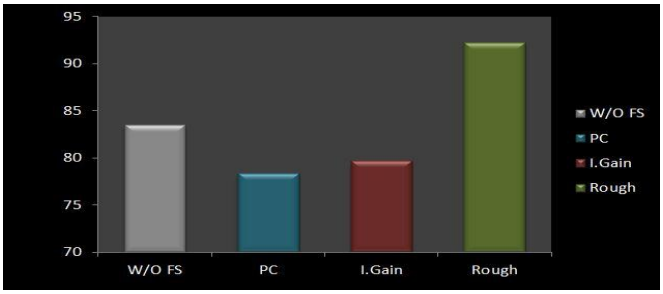

Figure 5. Accuracy of J48 classifier on EEG dataset before and after applying feature selection methods

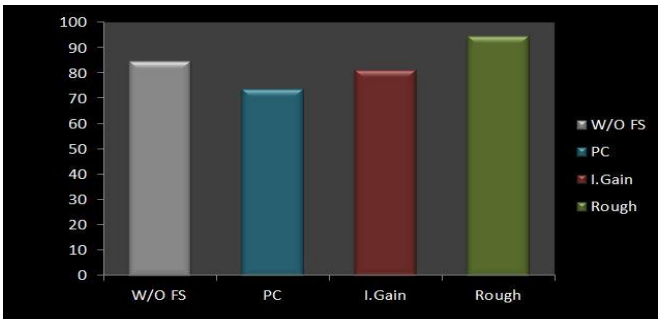

Figure 6. Accuracy of KNN classifier on EEG dataset before and after applying feature selection methods

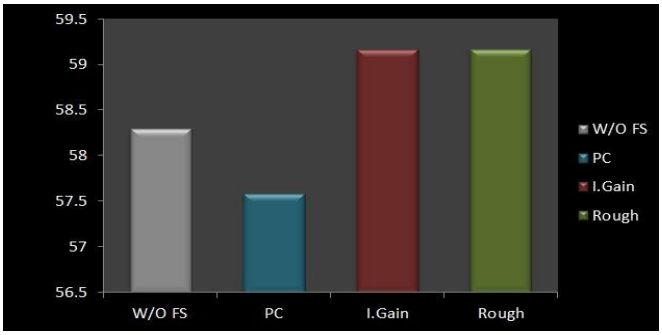

Figure 7. Accuracy of NB classifier on EEG dataset before and after applying feature selection methods

Cardiotocography Dataset (Cardio)

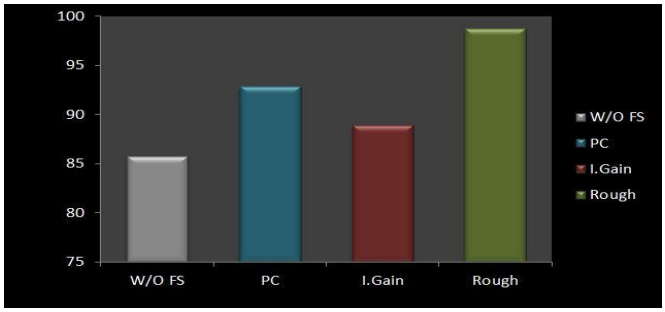

Figure 8. Accuracy of J48 classifier on Cardio dataset before and after applying feature selection methods

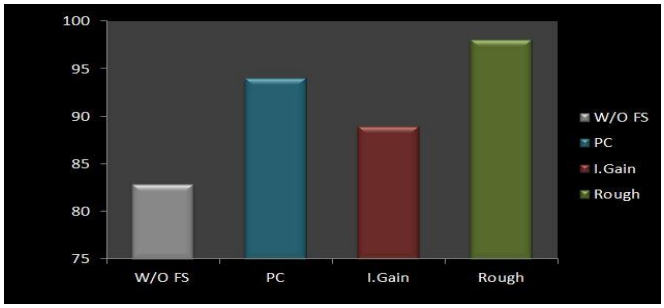

Figure 9. Accuracy of KNN classifier on Cardio dataset before and after applying feature selection methods

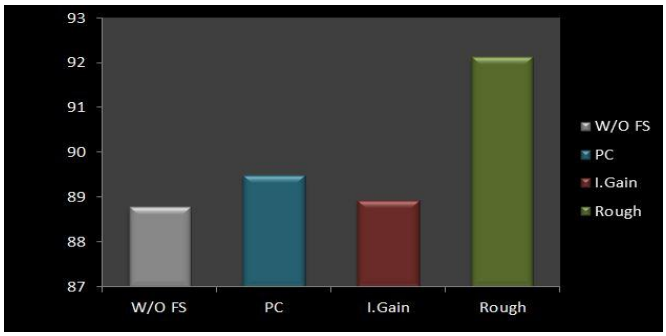

Figure 10. Accuracy of NB classifier on Cardio dataset before and after applying feature selection methods

Thoracic Surgery Dataset (TS)

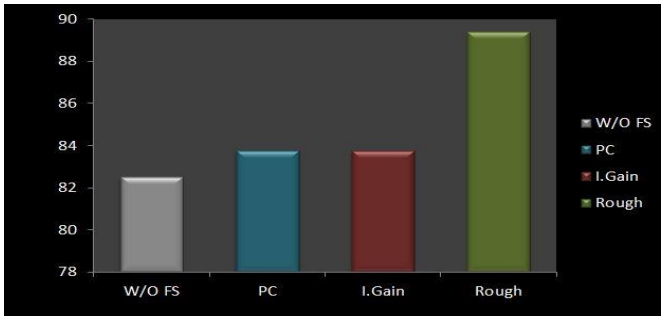

Figure 11. Accuracy of J48 classifier on TS dataset before and after applying feature selection methods

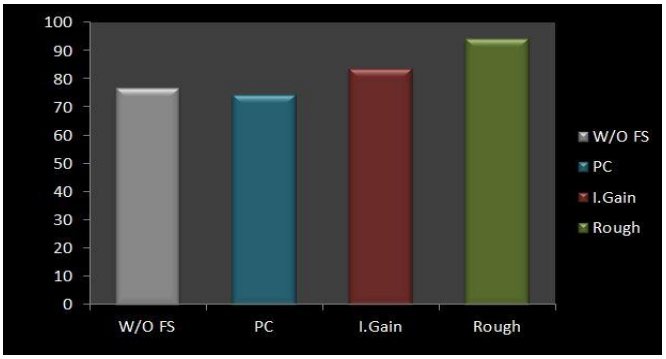

Figure 12. Accuracy of KNN classifier on TS dataset before and after applying feature selection methods

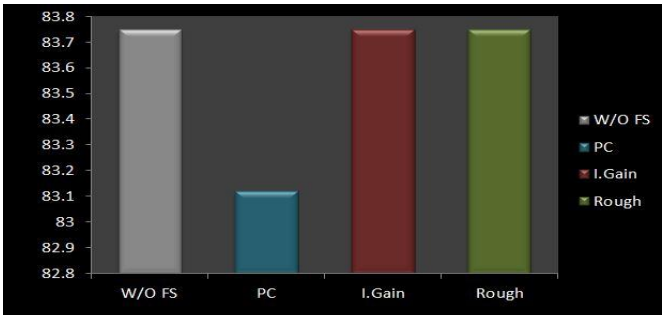

Figure 13. Accuracy of NB classifier on TS dataset before and after applying feature selection methods

Pima Indians Diabetes Dataset (PIDD)

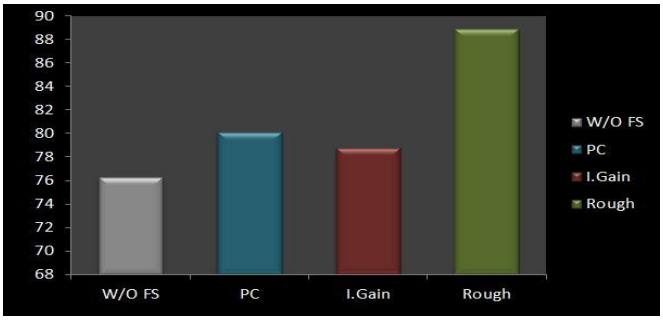

Figure 14. Accuracy of J48 classifier on PIDD dataset before and after applying feature selection methods

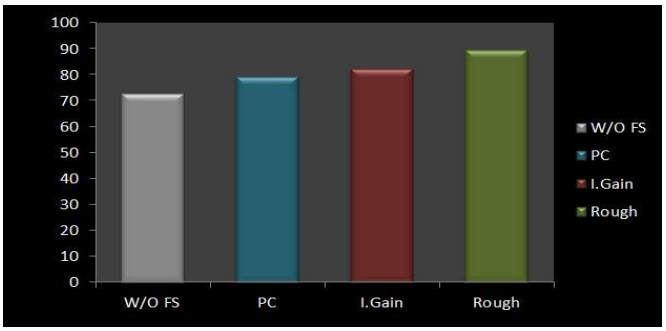

Figure 15. Accuracy of KNN classifier on PIDD dataset before and after applying feature selection methods

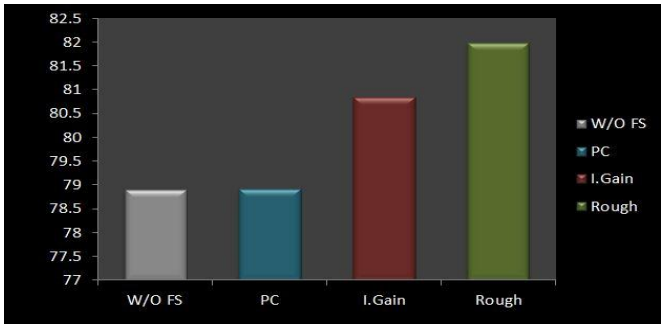

Figure 16. Accuracy of NB classifier on PIDD dataset before and after applying feature selection methods

Indian Liver Patient Dataset (ILPD)

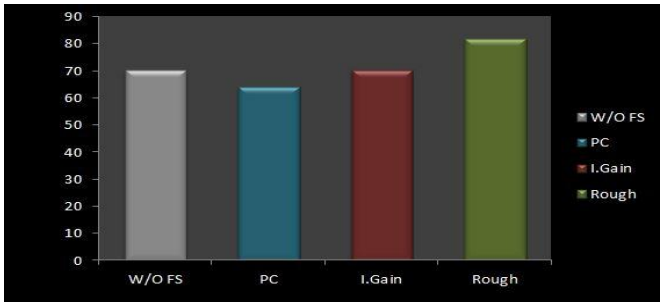

Figure 17. Accuracy of J48 classifier on ILPD dataset before and after applying feature selection methods

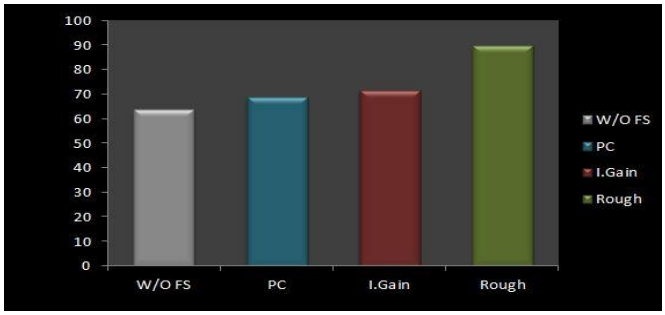

Figure 18. Accuracy of KNN classifier on ILPD dataset before and after applying feature selection methods

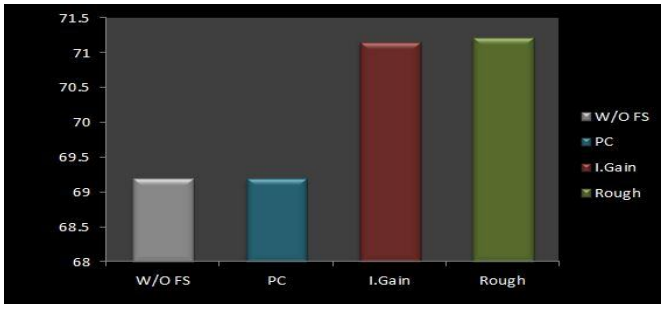

Figure 19. Accuracy of NB classifier on ILPD dataset before and after applying feature selection methods

Breast Cancer Wisconsin Original Dataset (BCWD)

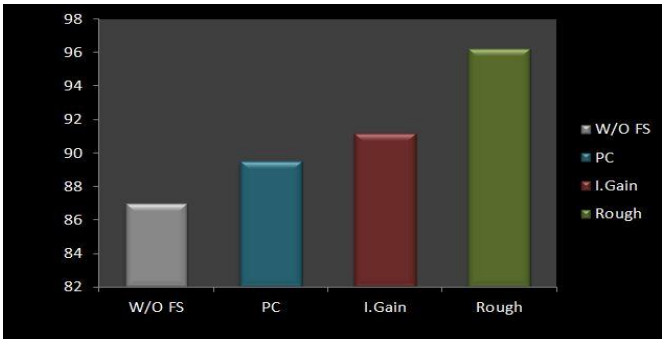

Figure 20. Accuracy of J48 classifier on BCWD dataset before and after applying feature selection methods

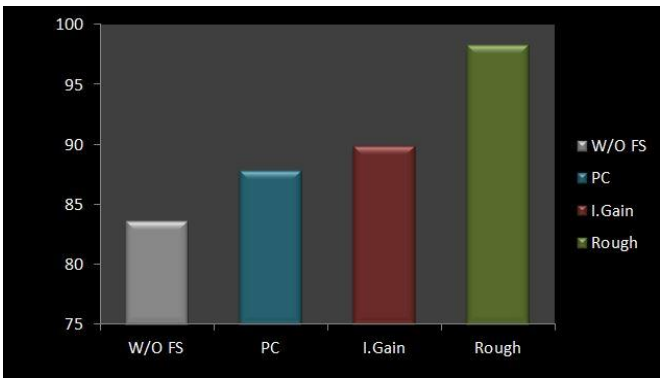

Figure 21. Accuracy of KNN classifier on BCWD dataset before and after applying feature selection methods

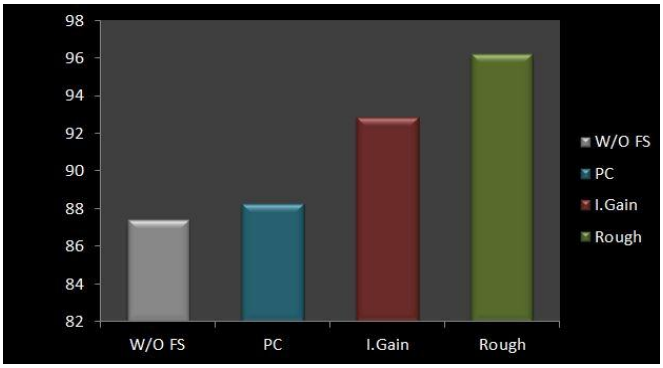

Figure 22. Accuracy of NB classifier on BCWD dataset before and after applying feature selection methods

Output of step5:

From above figures it can be claimed that for all 7 datasets and in case of all three classifiers FSUDD method (Rough Set Based method) is giving better results. To verify these results a statistical analysis, 'z-test' is used. In this test accuracy of three classifiers in Rough Set Based method are compared with the accuracy of classifiers using other three approaches (WFS, PC and I.Gain) in table 8 to table 14.

B. **Statistical Analysis**

Here, WFS-Rough Set represents comparisons between Without Feature Selection results with Rough Set Based results, PC-RoughSet represents comparisons between Correlation Based Feature Selection results with Rough Set Based Feature Selection results and I.Gain-RoughSet represents comparisons between Information Gain Based Feature Selection results with Rough Set Based Feature Selection results.

Diabetic Retinopathy Debrecen Dataset (DR)

Table 8. Statistical Analysis of Different FS Methods on Diabetic Retinopathy Dataset

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-4.16	0	-2.73	0.006	-3.31	0.0009
KNN	-3.77	0.00016	-4.05	0.00	-2.64	0.008
NB	-4.23	0	-3.66	0.0002	-4.79	0.00

EEG Eye State Dataset (EEG)

Table 9. Statistical Analysis of Different FS Methods on EEG Eye State Dataset (EEG)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-13.73	0	-19.78	0.00	-18.63	0.00
KNN	-16.12	0	-28.55	0.00	-19.83	0.00
NB	**-1.02**	**0.3077**	-2.044	0.041	**0.00**	**1**

Cardiotocography Dataset (Cardio)

Table 10. Statistical Analysis of Different FS Methods on Cardiotocography Dataset (Cardio)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-8.86	0	-5.23	0	-7.74	0
KNN	-10.14	0	-3.881	0.0001	-7.74	0
NB	-2.53	0.011	-1.94	0.051	-2.53	0.011

Thoracic Surgery Dataset (TS)

Table 11. Statistical Analysis of Different FS Methods on Thoracic Surgery Dataset (TS)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-1.77	0.07	-1.65	0.09	-1.65	0.09
KNN	-4.50	0	-4.87	0.00	-3.084	0.00208
NB	**0**	**1**	**-0.15**	**0.880**	**0**	**1**

Pima Indians Diabetes Dataset (PIDD)

Table 12. Statistical Analysis of Different FS Methods on Pima Indians Diabetes Dataset (PIDD)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-3.56	0.00036	-2.49	0.012	-3.04	0.00236
KNN	-4.90	0	-3.11	0.0018	-2.27	0.0232
NB	**-0.84**	**0.39**	**-0.84**	**0.39**	-0.28	**0.77**

Indian Liver Patient Dataset (ILPD)

Table 13. Statistical Analysis of Different FS Methods on Indian Liver Patient Dataset (ILPD)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-2.54	0.011	-3.78	0.00016	-2.54	0.01108
KNN	-6.05	0	-5.08	0.00	-4.47	0.00
NB	**-0.43**	**0.66**	**-0.43**	**0.6672**	**-0.0132**	**0.99**

Breast Cancer Wisconsin Original Dataset (BCWD)

Table 14. Statistical Analysis of Different FS Methods on Breast Cancer Wisconsin Original Dataset (BCWD)

FS Method	WFS-Rough Set		PC- Rough Set		I.Gain- Rough Set	
Classifier	Z	P-Value	Z	P-Value	Z	P-value
J48	-3.81	0.00014	-2.89	0.003	-2.2125	0.0271
KNN	-5.58	0	-4.55	0.00	-3.98	6E-05
NB	-3.52	0.00044	-3.21	0.0012	-1.83	0.06

From table 8, 10 and 14 it can be noticed that, in case of DR, Cardio and BCWD datasets, WFS-Rough Set, PC-Rough set and I.Gain-Rough set for all three comparisons and for all three classifiers the conditions to discard null hypothesis are satisfied. From table 9, 11, 12 and 13 it can be noticed that the conditions to discard null hypothesis are satisfied in case of J48 and KNN classifier only.

X. CONCLUSION

By analysing accuracy of classifiers it is noticed that the selected classifiers work better on the features selected by the FSUDD method than the other two methods. By observing the statistical analysis it is further noticed that for DR, Cardio and BCWD dataset the FSUDD method provides significantly better accuracy for all three classifiers. For rest of the four datasets though the FSUDD method provides better accuracy but that is not statistically significant for the naive bayes classifier. But it provides statistically significant accuracy in case of other two classifiers i.e. J48 and KNN. So, we can conclude that the FSUDD method provides better subset of features which in turn provides better accuracy for all classifiers on seven datasets but the results are statistically significant only for J48 and KNN classifier.

REFERENCES

[1] Imran Fareed Nizami, Muhammad Majid, Hammad Afzal and Khawar Khurshi, *"Impact of Feature Selection Algorithms on Blind Image Quality Assessment"*, Arabian Journal for Science and Engineering, pp **1–14**, August **2017**.

[2] Abdullah S. Ghareb, Abdul Razak Hamdan and Azuraliza Abu Bakar, *"Integrating Noun-Based Feature Ranking and Selection Methods with Arabic Text Associative Classification Approach"*, Arabian Journal for Science and Engineering, Vol.**39**, Issue.**11**, pp 7807–7822, November **2014**.

[3] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences, **11**, **341-356**, **1982**

[4] Javad Rahimipour Anaraki, Kerman, Iran, Mahdi Eftekhari, *"Rough Set Based Feature Selection: A Review"*, 5th Conference on Information and Knowledge Technology, IEEE, **2013.**

[5] G. K. Gupta, *"Introduction to Data Mining with Case Studies"*, Prentice Hall of India New Delhi, **2006.**

[6] P-N. Tan, M. Steinbach, V. Kumar, *"Introduction to Data Mining"*, Addison Wesley Publishing, **2006.**

[7] O.Maimon and L.Rokach, *"Data Mining and Knowledge Discovery"*, Springer Science and Business Media, **2005.**

[8] X. Niuniu and L. Yuxun, *"Review of Decision Trees"*, IEEE, **2010.**

[9] Payam Emami Khoonsari and AhmadReza Motie, *"A Comparison of Efficiency and Robustness of ID3 and C4.5 Algorithms Using Dynamic Test and Training Data Sets",* International Journal of Machine Learning and Computing, Vol.**2**, Issue.**5**, October **2012.**

[10] V. Garcia, C. Debreuve, *"Fast k Nearest Neighbor Search using GPU"*, IEEE, **2008**.

[11] A. Ashari I. Paryudi and A Min Tjoa, *"Performance Comparison between Naïve Bayes Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool"*, International Journal of Advanced Computer Science and Applications, Vol.**4**, Issue. **11**, **2013.**

[12] Dougherty, J., R. Kohavi and M. Sahami, *"Supervised and unsupervised discretization of continuous features"*, Proceeding of the 12th International Conference on Machine Learning, **1995.**

[13] https:// archive.ics.uci.edu/ ml/ datasets/ Diabetic + Retinopathy + Debrecen + Data + Set

[14] https:// archive.ics.uci.edu/ ml/ datasets/ EEG + Eye + State

[15] https:// archive.ics.uci.edu/ ml/ datasets/ cardiotocography

[16] https://archive.ics.uci.edu/ ml/ datasets/ Thoracic + Surgery + Data

[17] PIDD Dataset, https:// archive.ics.uci.edu/ ml/ datasets/ pima + indians + diabetes

[18] https:// archive.ics.uci.edu/ ml/ datasets/ ILPD + (Indian + Liver + Patient + Dataset)

[19] https:// archive.ics.uci.edu/ ml/ datasets/ breast + cancer + wisconsin + (original)

Authors Profile

Mrs. Kasturi Ghosh pursed Bachelor of Technology from University of Kalyani In 2003 and Master of Engineering from IIEST, Shibpur in 2009. She is currently pursuing Ph.D from Kalyani University and working as Assistant Professor in Department of Information Technology, The University of Burdwan . She has already published five papers in different journals. Her main research area focuses on Data Mining (Application in Medical Domain), Big data, Rough Set and Statistics.

Ms Susmita Nandi pursued Bachelor of Engineering in Computer Science and Engineering from University Institute of Technology, The University of Burdwan in 2016 and Master of Engineering in Computer Science and Engineering from the same institution in 2018. Her research work focuses on Feature Selection, Data Mining. She has published a paper on Comparison of Performance of Data Mining Classifiers [Published in International Journal of Engineering and Inventions (UGC approved). e-ISSN:2319-6734. p-ISSN:2319-6726]

