Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Enhanced Feature Engineering By Discretized Naïve Bayes To Predict Soil Fertility For The Betterment Of Sugarcane Yield

Raynukaazhakarsamy[1] , Dr. J.G.R. Sathiaseelan[2]

[1]Research Scholar, Department of Computer Science Bishop Heber College (Autonomous), Tiruchirappalli-620 017, Tamil Nadu, India. Affiliated to Bharathidasan University, Trichy-24.

[2]Head, Department of Computer Science Bishop Heber College (Autonomous), Tiruchirappalli-620 017, Tamil Nadu, India. Affiliated to Bharathidasan University, Trichy-24.

Abstract— The challenging task of feature engineering in data mining is the dimensionality reduction to extract relevant attributes. Hence, the inherent analysis of data distribution based on its class label is essential for predicting the results. This research work contributes two machine learning models namely Wrapper based Discretized Naïve Bayes (WDN Bayes) and Filter based Discretized Naïve Bayes (FDN Bayes) using the supervised machine learning algorithms such as NB, KNN and DN Bayes that aims for identifying an optimal subclass of features from the collection of primary soil dataset based on chemical nutrients around Theni region to predict soil fertility by improving the classification accuracy. Extensive experiments on four different real-time soil datasets were carried over to demonstrate the effectiveness of KNN embedded wrapper method and CFS+GA combined filter method using Discretized Naïve Bayes (DN Bayes). The model's effectiveness is estimated with all the features and with the significant features obtained by the proposed feature extraction techniques. The experimental results of feature extraction approach profoundly satisfying in terms of error metrics and evaluation metrics in comparison with NB, KNN, SVM, DN Bayes, WDN Bayes and FDN Bayes by producing 91% and 92% of classification accuracy for WDN Bayes and FDN Bayes respectively.

Keywords- Soil fertility, Crop yield Prediction, Soil Classification accuracy, Correlation filter, Wrapper approach and Feature extraction.

1. Introduction:

Machine Learning has ascended with a lot of processing techniques to develop new strategies in multi-disciplinary agricultural innovations. This leads to scale the performance of learning model through feature extraction technique to determine the relevant subset of features using different statistical measures for building a model [1]. An approach that dealt with farm management using information technology is termed as Precision agriculture and also known as “Site-Specific Agriculture”. It assures whether the crops and soil receive their required nutrients for good health and profit [2]. Guaranteed productivity, manageability and environment conservation is the objective of precision agriculture [3]. Currently, more machine learning methods are developed significantly to enhance the agricultural needs efficiently and adjust to various natural changes [4]. In precision agriculture machine learning endows crop management system that assists in soil suitability for the good yield of crops, crop disease management, differentiating crop weeds, acknowledging crop assortments, forecasting of agricultural climate and so on [5]. As there is an enormous increase in the agricultural data, feature space can have irrelevant and redundant features and often generates a classifier that has poor performance and weak robustness [6]. The existing experiments demonstrate that redundant features affect the performance of classifiers and also the instance-based learners are sensitive to irrelevant features [7]. These problems can be solved by removing irrelevant and redundant features from the original feature space using effective feature selection method [8,9].

Feature subset selection is the process of removing irrelevant and redundant features and the identification of a feature subset that contains the most discriminative information from the original feature space [10]. In addition to the dimensionality reduction, feature selection enhances the generalization ability of the classifiers, facilitates data visualization, reduces the training time, improves the performance of the classifiers and help the biologists to identify the hidden biological mechanisms [11,12]. Subset generation module and an evaluator module are the two components of feature selection. The feature subset generation module exploits search strategies to generate candidate subsets, whereas the evaluator module measures the goodness of a subset. Feature selection methods are classified as filter, wrapper and embedded based on the involvement of evaluator in the classifier [13]. Filter method evaluate the quality of a feature or the subset of features by the essential properties of the training samples and flexible in combining with various combination of classifiers. It has better generalization ability and reduces the computational complexity. Wrapper method is specific to a given classifier and evaluate the quality of a candidate subset that tends to obtain better classification performance [14,15]. Embedded method is characterized by a deeper interaction between the feature selection and the construction of the classifier to generate the feature subsets. As wrapper method achieve better classification accuracy it is far more time consuming in actual use of data. It evaluates $O(N^2)$ candidate subset for N-feature dataset using sequential forward selection scheme that creates high time complexity and requires more CPU time [16,17,18]. To pacify this problem and enhance the process of feature selection, in this study, we investigated and proposed two feature engineering methods namely WDN Bayes and FDN Bayes based on wrapper approach and filter approach respectively. The proposed wrapper approach is embedded with KNN and proportional k-interval discretized Naïve Bayes classifiers named as WDN Bayes and the proposed filter approach utilize correlation-based feature selection algorithm (CFS) for feature subset selection and optimized by genetic algorithm together with proportional k-interval discretized Naïve Bayes classifiers named as FDN Bayes. The effectiveness of proposed methods is tested on real-time soil datasets received from Rajashree Sugars and Chemicals Ltd., Periyakulam. The soil samples are collected from the surrounding lands of Theni region. The efficiency of proposed methods is experimentally validated and analyzed by comparing with the state-of-art feature selectors.

The remainder of this paper is organized as follows. Section 2 comprises machine learning algorithm, section 3 & 4 describes the wrapper based proposed feature selection model, section 5 & 6 discuss about the filter based proposed feature selection model, section 7 explains about the experimental results and discussion, finally section 8 concludes this feature engineering research work.

2. K-Nearest Neighbor:

KNN is a non-parametric learning algorithm applied for classification and regression in pattern recognition. It is a simple and efficient algorithm that exhibits time complexity of O(1) [19]. Various distance metrics are used to measure the distance between two instances according to the type of attribute. Consider, two instances based on attribute types x = (x1, x2, …….,xn) and y = (y1, y2, …….,yn) from the experimental samples. The distance between two instances based on attribute xj ($1 \leq j \leq n$) is calculated as follows.

For categorical attributes,

d (xj , yj) = 0, if (yj == xj)
d (xj , yj) = 1, otherwise.

For numerical attributes, Euclidean and Manhattan distances are the commonly used metrics.

Euclidean Distance is $d\ (xj\ ,\ yj) = \sqrt{}\ (xj - yj)^2$
Manhattan Distance is $d\ (xj\ ,\ yj) = |(xj - yj)|$

The distance D (xj , yj) between x and y in terms of Euclidean Distance is recursively defined as,

$$D\ (xj\ ,\ yj)^2 = D\ (x1, x2, \ldots\ldots., xn\text{-}1\ ;\ y1, y2, \ldots\ldots., yn\text{-}1)^2 + d\ (xj\ ,\ yj)^2 \qquad (1)$$

To predict the class label of a new instance, KNN finds the k closest neighbors from the training set according to the distance metric and then assigns the dominant label among the k neighbors to the new instance. If k=1, the label of new instance is determined by its closest neighbor. KNN is commonly used as a standard classifier and integrated into the feature selection framework to evaluate the quality of a feature subset and compare the performance of different feature selection algorithms due to its effective implementation [20].

3. Feature Selection with Wrapper approach:

Wrapper method integrates a classifier in the feature selection process to evaluate the quality of a feature or a feature subset tends to attain better classification accuracy. Enumerating all of the possible combinations of feature subsets and evaluating them one by one is the simplest approach and guarantees optimal feature subset in the N number of features, while the computational complexity grows exponentially at O(N^2) which exhibits high time complexity. This problem can be rectified to generate candidate feature subsets by commonly used search methods such as Step-wise Forward Selection (SFS), Sequential Backward Selection (SBS), bidirectional search, sequential floating search, heuristic search and random search [21]. Even though various search strategies available SFS achieves better results with the quality of obtained feature subset. Starting from an empty set, SFS begins by selecting the feature that is most relevant to the target attribute as evaluated by a classifier and then searches for the next candidate feature that most contributes to the enhancement of the classification accuracy among the remaining features and continues with this process until there is no other candidate feature left over. Through this deterministic search strategy, the wrapper method evaluates only O ((S+1) N) candidate feature subsets with finally selected S features. Algorithm-1 depicts the pseudocode of SFS based on wrapper approach and Algorithm 2 shows the pseudocode of KNN embedded SFS based on wrapper approach.

Algorithm-1: Step-wise Forward Selection (SFS) based on wrapper approach

Input: Denoised Soil Dataset with Feature Set F & class label C

Output: S; //Selected feature subset

```
acc = 0;
S = null;
while(~is empty (F))
 {
   flag = 0;
         for i = 1 to length(F)
          {
           Snew = add(copy(S), Fi);
           accnew = evaluate(classifier, $Data^{SnewU\{C\}}$);
         if(accnew > acc) then
                 x = i;
                 acc = accnew;
                 flag = 1;
         if(flag) then
           {
                 S = add(S; Fx);
                 F = del(F; Fx);
           }
         else
```

```
        break; // stop feature selection }
    return S;
```

4. Enhanced Wrapper-Based SFS with Embedded KNN:

In the process of evaluating a new feature subset a new KNN classifier is constructed each and every time as discussed in section 2. The final feature subsets are obtained incrementally by searching and evaluating the candidate feature subset in the wrapper-based SFS method. Construction of KNN classifier has no explicit training step whereas, all the computation deferred until the classification process gets completed. By comparing the distance between the test instance and all the training instances and then choosing the k nearest neighbors to determine the class label of the test instance. This leads to construct a distance matrix to maintain the distance between any two different instances in the soil dataset utilized of the experiment projected over the selected feature subset. Candidate feature can be evaluated incrementally constructed a new KNN classifier by adding a distance matrix on the candidate feature rather than calculating it with overall features. This works based on two approaches discussed as follows.

Attribute distance matrix denoted as [**Dt(Fi)]** contains the distance between any two different instances in the experimental soil dataset projected over the predictive feature

$$F_i = \{F_1, F_2, \ldots\ldots, F_i \,/\, 1 \leq i \leq n\}.$$

Classifier distance matrix denoted as **(Dt)** contains the distance between any two different instances in the experimental soil dataset projected over the feature subset

$$S=\{S_1, S_2, \ldots\ldots, S_s \,/\, 1 \leq s \leq n\} \text{ in } F_i.$$

Each and every cell of the matrix stores the distance between any two instances and every row or column constitutes distance vector between the consecutive instances. This is represented in Fig-1. The distance between two instances is incrementally updated by the squared Euclidian distance recorded in the matrix and ensures that KNN classifier is constructed incrementally along with the feature selection. In this matrix the k-closest instances of a test instance can be found by using the values stored directly as it is non-negative and monotonically increasing along with the feature selection. The classifier distance matrix Dt performs as a fast KNN classifier to conduct the cross validation for the experimental soil dataset projected over the selected features and also works together with the attribute distance matrix for the incremental construction of a new classifier when evaluating the next candidate feature. This approach enhances the Wrapper- Based Step-wise Forward Selection (SFS) method by embedding KNN classifier described in Algorithm-2.

The quality of feature subset is evaluated by 10-fold cross validation on the classifier distance matrix rather than calculating the distance between the test instance and the remaining training instances. The attribute distance matrix $Dt(F_i)$ is calculated for candidate feature F_i and then obtain a candidate classifier distance matrix (Dt_{new}) by adding $Dt(F_i)$ to (Dt) and performed 10-fold cross validation on (Dt_{new}) to evaluate the quality of the feature subset for each and every execution the feature F_i is added for achieving better accuracy to the selected subset by replacing (Dt) with the corresponding (Dt_{new}) and extends to select the next feature. The stop criterion is that all the features are selected into S or there is no enhancement in the classification accuracy while evaluating the remaining features. This is shown in Fig-2.

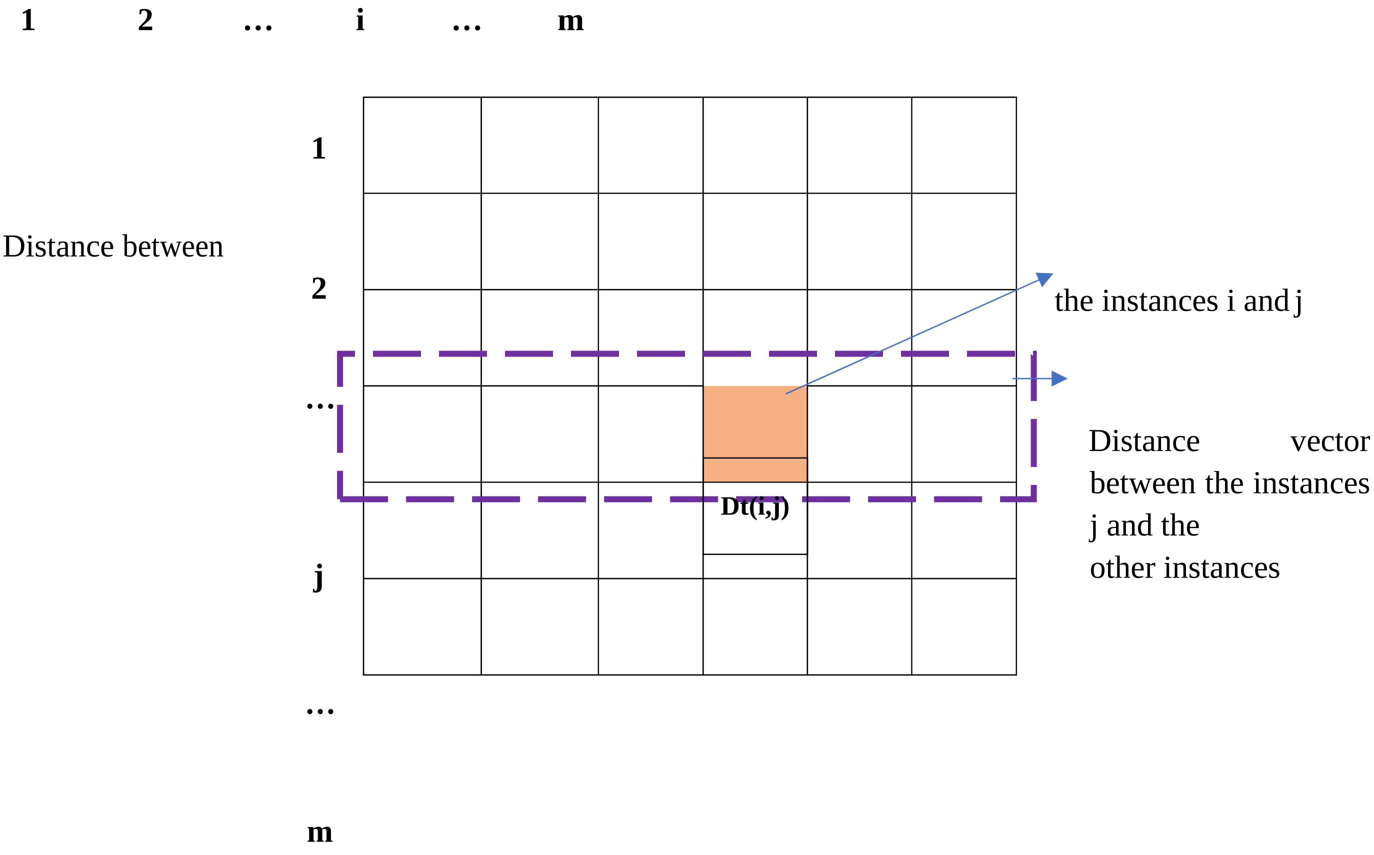
1
2
…
i
…
m
Distance between
1
2
…
j
…
m
the instances i and j
Dt(i,j)
Distance vector between the instances j and the other instances

Fig-1: KNN classifier distance matrix

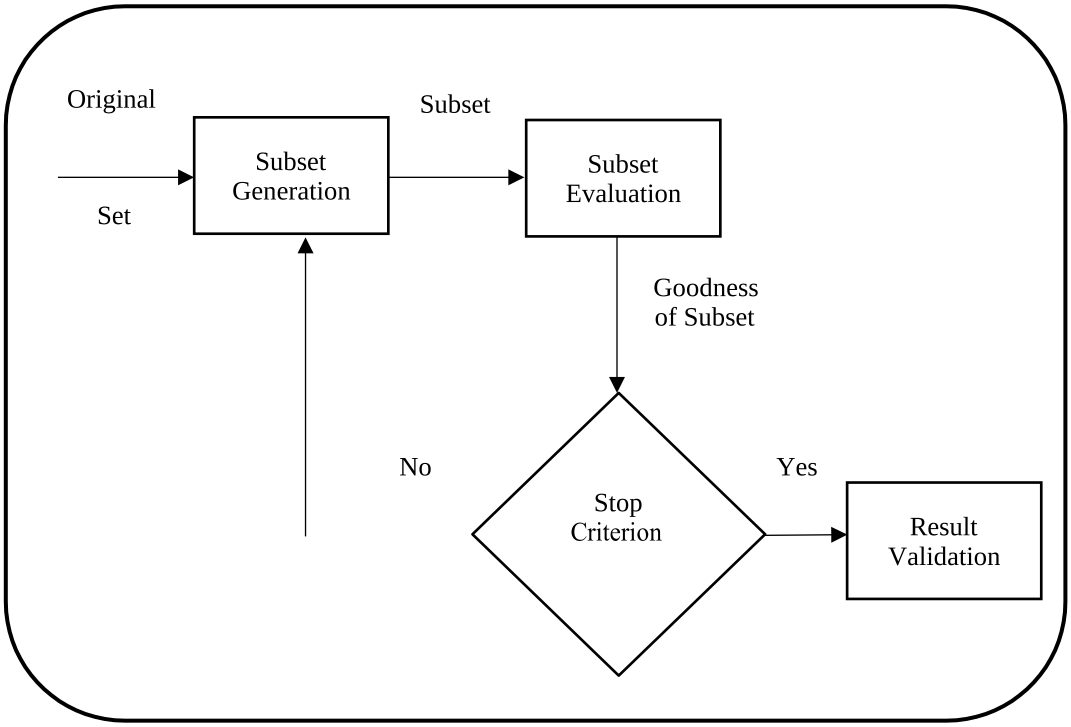

Fig-2: Feature Selection Process

Algorithm-2: KNN Embedded Wrapper-Based Step-wise Forward Selection (SFS)

Input: Denoised Soil Dataset with Feature Set F & class label C

Output: S; //Selected feature subset

```
Dt = null;
acc = 0;
S = null;
while(~is empty (F))
 {
   flag = 0;
        for i = 1 to length(F)
         {
           Compute attribute distance matrix Dt(Fi);
           Dtnew = Dt + Dt(Fi);
           accnew = 10-fold cross validation on Dtnew;
        if (accnew > acc) then
                x = i;
                acc = accnew;
                flag = 1;
                Dtbest = Dtnew;
        if(flag) then
         {
                S = add (S; Fx);
                Dt = Dtnew;
                F = del (F; Fx);
          }
        else
  break; // stop feature selection 25 }
return S;
```

5. Correlation Based Feature Selector (CFS) – Filter Approach:

Quality of feature subsets depends on statistical evaluation criteria in the filter approach for feature extraction. A relevant feature conforms to a class or predicts the class [22]. A characteristic feature (X_i) is observed to be pertinent if and only if there occurs some probability $P(X_i)$ and y such that $P(X_i = x_i) > 0$,

$$P(Y = y \mid X_i = x_i) \neq P(Y = y) \qquad (2)$$

Along with insignificant features, the features which are exceedingly associated with one or more other features also to be removed in feature selection process. The features are specific tests that measure characteristics identified with the variable of importance. If the association among the individual feature and an extrinsic variable is recognize, and the inter relation among every pair of the features is given, then the association among the complicated test comprising of the total features and the extrinsic variable can be calculated as,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where, x_i and $\bar{x}$ defines the observed and average values of the features considered. y_i and $\bar{y}$ defines the observed and average values of the dataset class. The association between a group and an external feature is an operation of the total number of individual characteristic features in the group. The above mention Pearson's Coefficient formula is obtained by standardizing all the variable. It has been utilized in the correlation-based feature selection algorithm enabling the addition or deletion of one feature at a time. The following Algorithm-3 describes the feature selection procedure using CFS filter.

Algorithm-3: Filter-Based Correlation Feature Selection Optimized using GA

```
Input:
$D_{train}$ ← Training dataset
P ← Predictor
n ← No. of selected features
Output:
$F_x$ ← Selected Feature Subset Optimized Feature Subset
Begin:
$F_0 = Ø$;
x = 1;
while ($|F_x| < n$) do
    if ( $|F_x| < n-1$) then
        $F_x$ = CFS ($F_{x-1}$, $D_{train}$, P)
    else
        Add the better feature f to $F_{x-1}$
    endif
    x = x + 1;
```

```
end while
$F_X$ is further optimized using GA
End
```

Predicting relevant features using Correlation based filters defines as, the higher the correlation among the individual and the extrinsic variable, the higher is the correlation among the combination and external variables and the lower the inter-correlation among the individual and the extrinsic variable, the lower is the correlation among the combination and extrinsic variable. The redundant features should be removed from the dataset for an effective prediction.

6. Optimization of Reduced Feature Subset using Genetic Algorithm:

Further, to improvise the prediction performance of the model, the reduced feature set thus obtained through CFS filter is passed on the next step for optimizing using Genetic algorithm. It is adopted to optimize the possible combination of certain number if attributes that best describe the soil dataset while maintaining higher classification rate. The initial population for feature selection is generated based on Information Gain and Information gain threshold values.

The result of a combination of GA optimization and CFS filter approach as an induction algorithm is the set of significant attributes that provides high classification rate. The proportional k-interval discretized Naïve Bayes (DN Bayes) algorithm [23] is considered for the soil classification as fertile or non-fertile. However, with the implementation of loss function, the model is constructed to leverage the efficient of DN Bayes.

7. Experimental Results and Discussion:

Real-Time dataset of soil samples collected from Rajshree Sugars and Chemicals Ltd., Theni is taken for this research work. Sugarcane soil dataset is divided into training and testing samples with the ratio of 60% and 40% respectively. Test dataset with the size of 128 samples is utilized for this research work and experimental results are produced. Following 10 attributes namely, pH, EC, OC, N, P, K, S, Fe, Mn and Zn are considered. The Statistical report for soil attributes and the experimental results of proposed works DN Bayes and DNBQ in the preprocessing stage that are discussed in [23], carried over for the further prediction improvement through feature engineering approaches based on wrapper as WDN Bayes and filter as FDN Bayes.

These proposed models WDN Bayes and FDN Bayes are implemented on 4 different real-time soil datasets and the average of all the four results recorded. The predictive performance of these models is evaluated using various measures of evaluation that examined for our assessment are Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and evaluation metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-Score, Receiver Operator Characteristic (ROC), number of samples classified and classification accuracy.

Table 1: Error Metrics for WDN Bayes

Algorithms & Error Metrics	NB	DN Bayes	DNBQ	WDN Bayes
KS	0.0542	0.6736	0.7343	0.8872
MAE	0.3992	0.1979	0.1924	0.1541
RMSE	0.6265	0.3331	0.3202	0.2267
RAE	0.8222	0.4076	0.3962	0.3142
RRSE	1.2716	0.6761	0.6499	0.4576

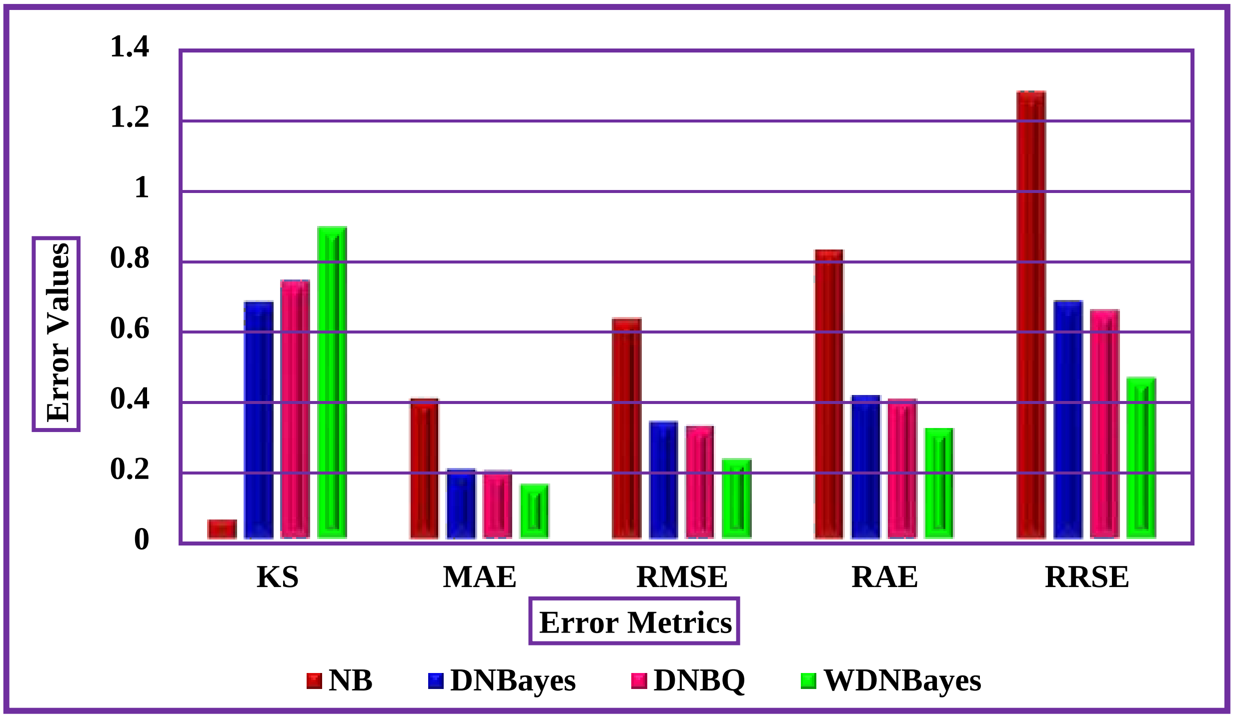

Figure 3: Error Metrics for WDNBayes

Table 2: Evaluation Metrics for WDN Bayes

Algorithms & Evaluation Metrics	NB	DN Bayes	DNBQ	WDN Bayes
TPR	0.602	0.842	0.872	0.945
FPR	0.554	0.171	0.144	0.068
Precision	0.627	0.842	0.872	0.948
Recall	0.602	0.842	0.872	0.945
F1-Score	0.489	0.842	0.872	0.945
ROC	0.731	0.905	0.946	0.965

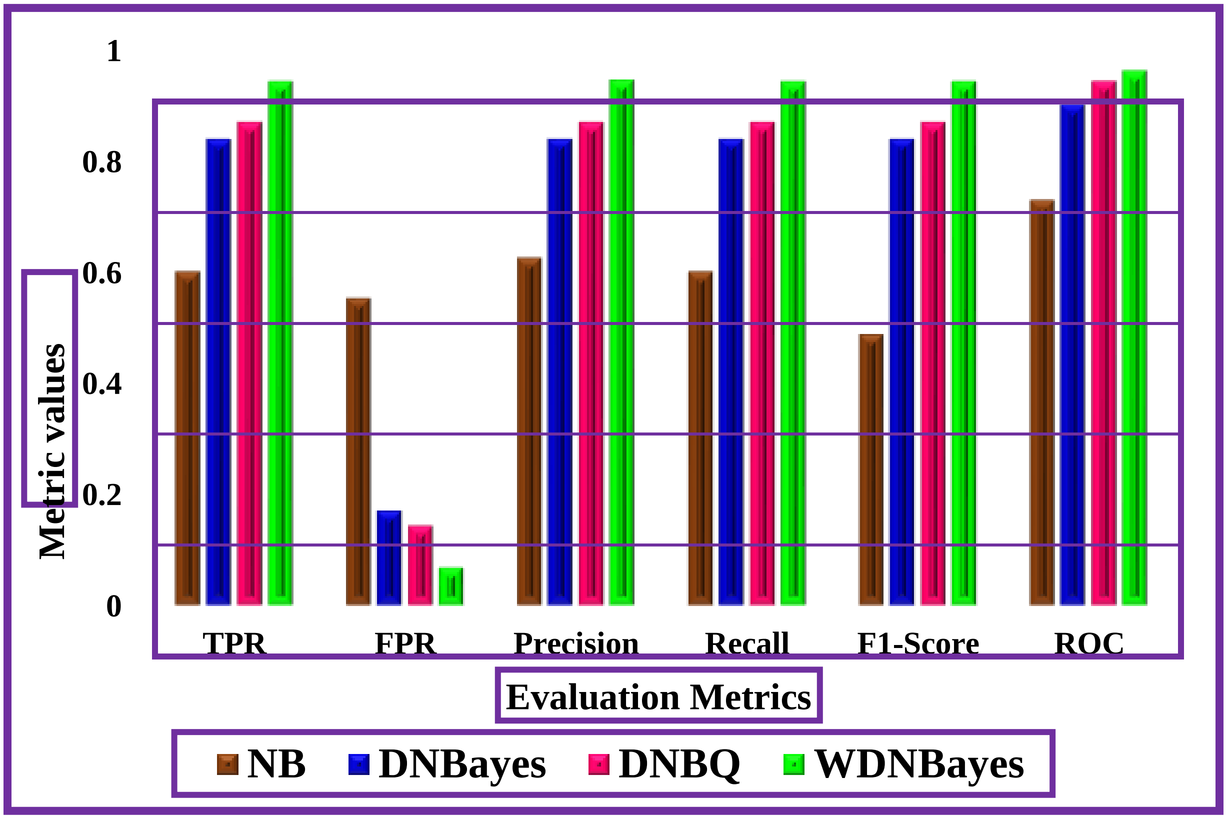

Figure 4: Evaluation Metrics for DN Bayes

Table 3: Classification Accuracy for WDN Bayes

Metrics & Algorithms	No. of Samples		Classification Accuracy (%)
	Correctly Classified	Incorrectly Classified	
NB	78	50	61
DN Bayes	109	19	85
DNBQ	116	12	87
WDN Bayes	121	7	94

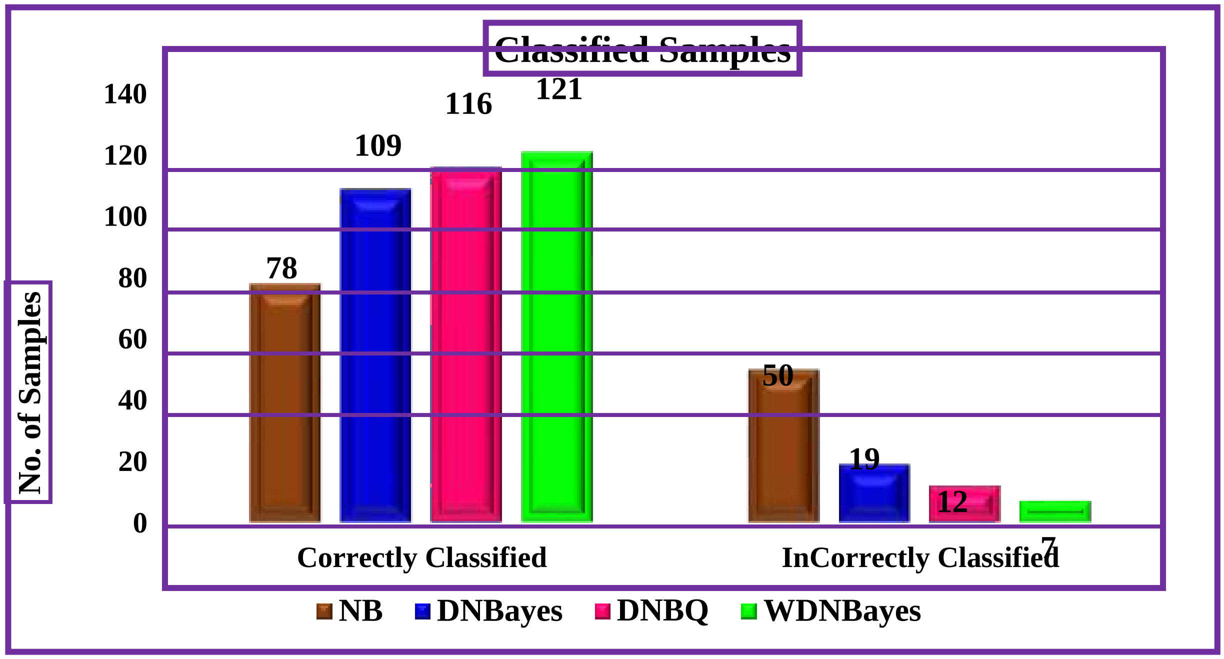

Figure 5: Classified Samples for WDN Bayes

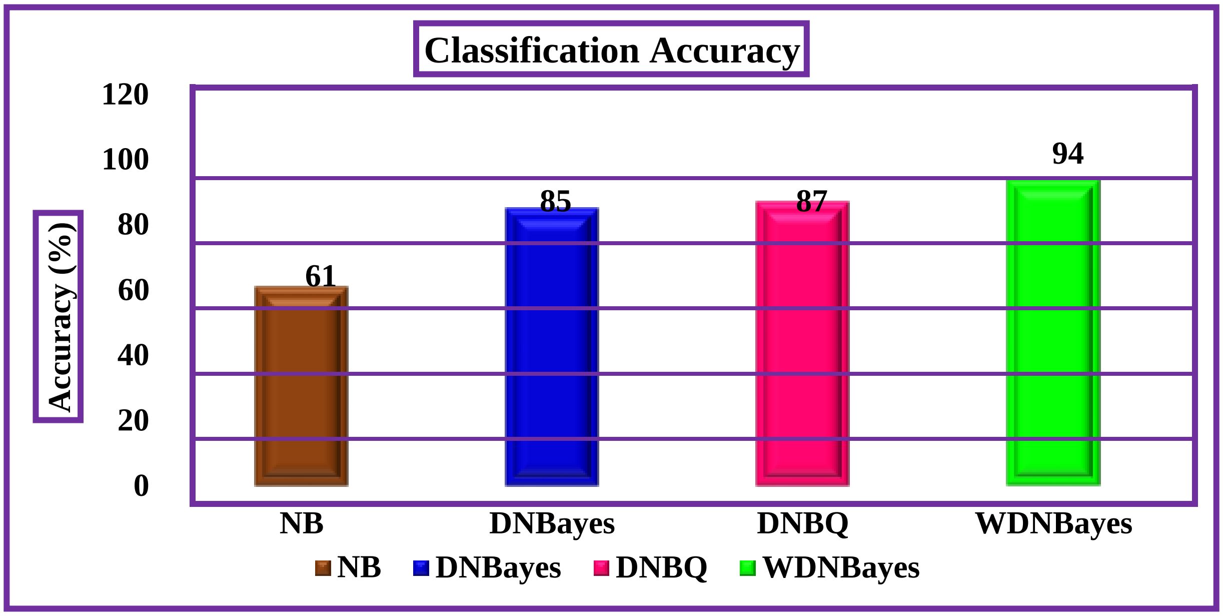

Figure 6: Classification Accuracy for WDN Bayes

Experimental results of Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and evaluation metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-Score, Receiver Operator Characteristic (ROC), number of samples classified and classification accuracy based on WDN Bayes are shown in (Table 1 - 3) also its graphical representation is shown in (Figure 3 – 6). The results reveal that WDN Bayes outperforms other existing algorithm NB and our previously proposed works DN Bayes and DNBQ by producing 94% of classification accuracy in a real-time dataset with 128 samples.

Table 4: Error Metrics for FDN Bayes

Algorithms & Error Metrics	NB	DN Bayes	DNBQ	WDN Bayes	FDN Bayes
KS	0.0542	0.6736	0.7343	0.8872	0.9031
MAE	0.3992	0.1979	0.1924	0.1541	0.1453

RMSE	0.6265	0.3331	0.3202	0.2267	0.2149
RAE	0.8222	0.4076	0.3962	0.3142	0.2962
RRSE	1.2716	0.6761	0.6499	0.4576	0.4338

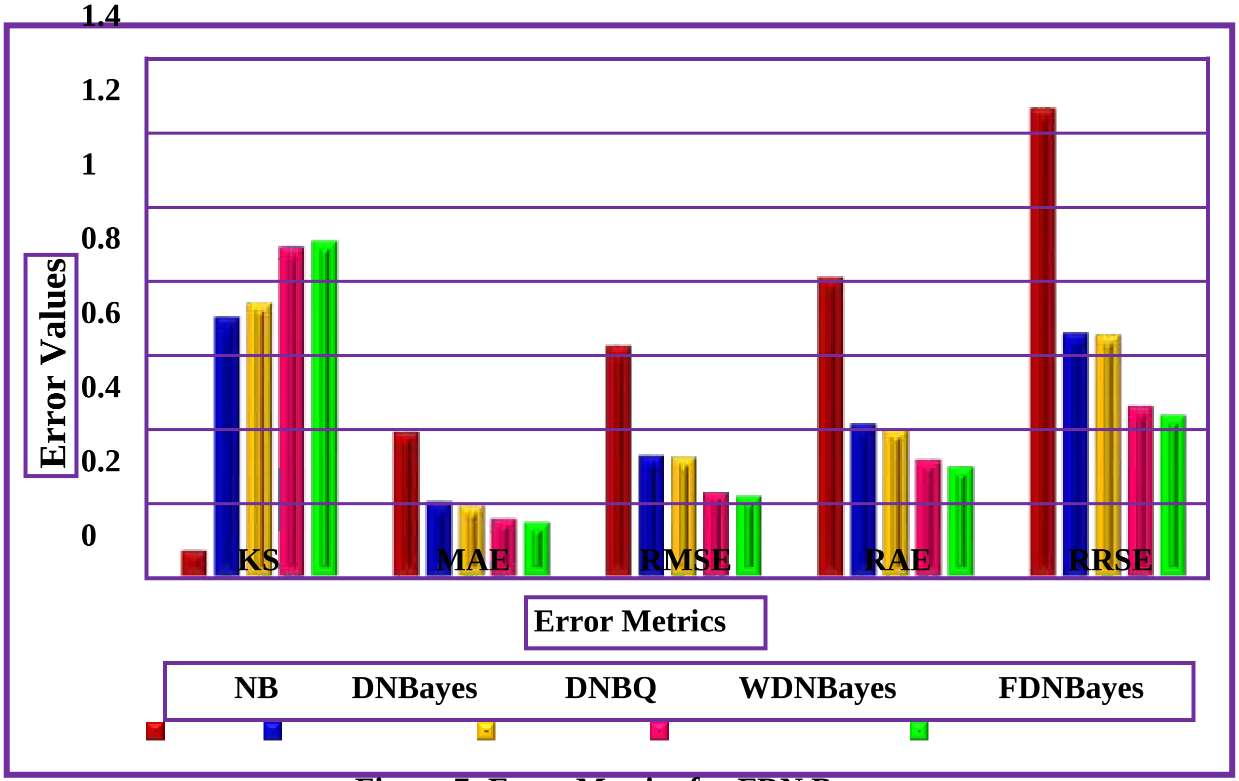

Figure 7: Error Metrics for FDN Bayes

Table 5: Evaluation Metrics for FDN Bayes

Algorithms & Error Metrics	NB	DN Bayes	DNBQ	WDN Bayes	FDN Bayes
TPR	0.602	0.842	0.872	0.945	0.953
FPR	0.554	0.171	0.144	0.068	0.062
Precision	0.627	0.842	0.872	0.948	0.957
Recall	0.602	0.842	0.872	0.945	0.953
F1-Score	0.489	0.842	0.872	0.945	0.953
ROC	0.731	0.905	0.946	0.965	0.971

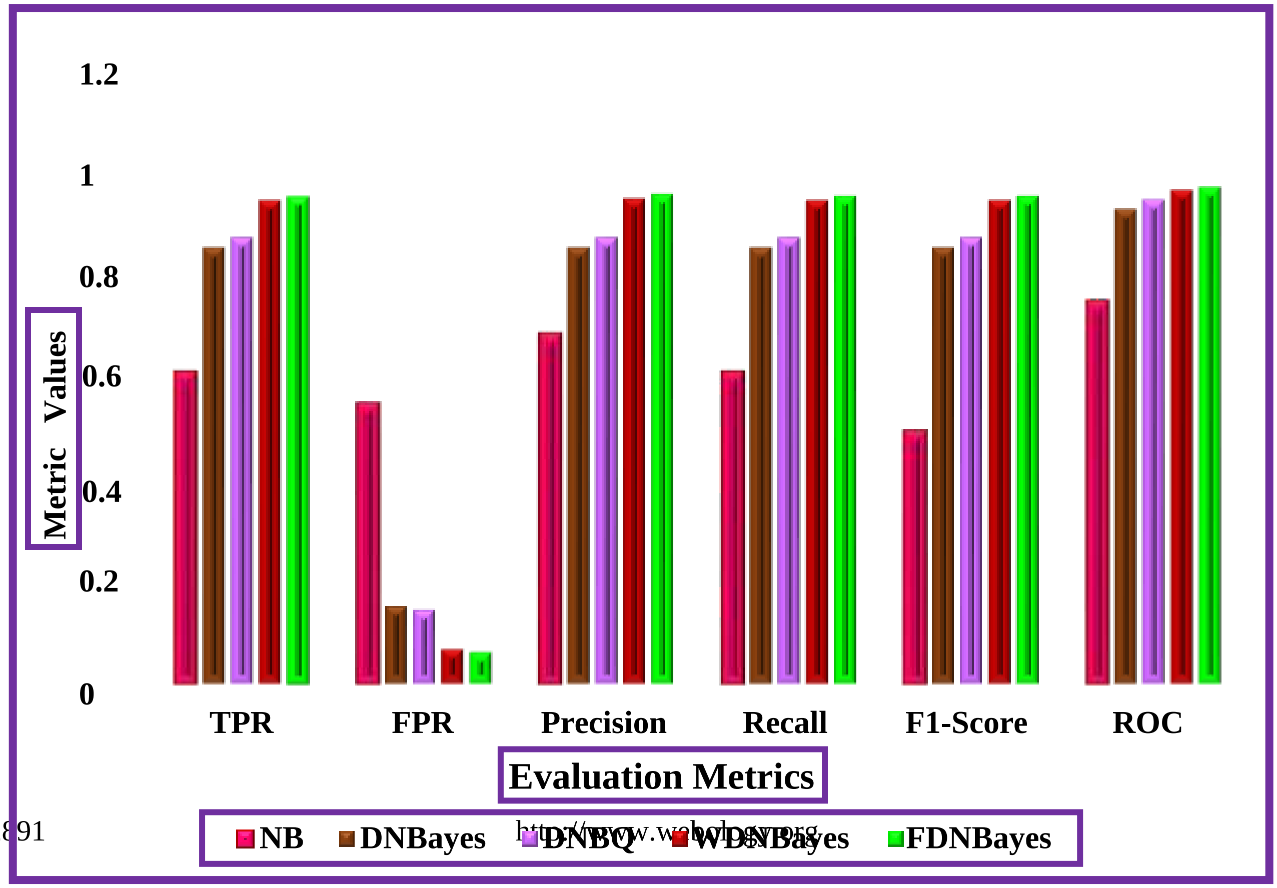

Figure 8: Evaluation Metrics for FDN Bayes

Table 6: Classification Accuracy for FDN Bayes

Metrics & Algorithms	No. of Samples		Classification Accuracy (%)
	Correctly Classified	Incorrectly Classified	
NB	78	50	61
DN Bayes	109	19	85
DNBQ	116	12	87
WDN Bayes	121	7	94
FDN Bayes	122	6	95

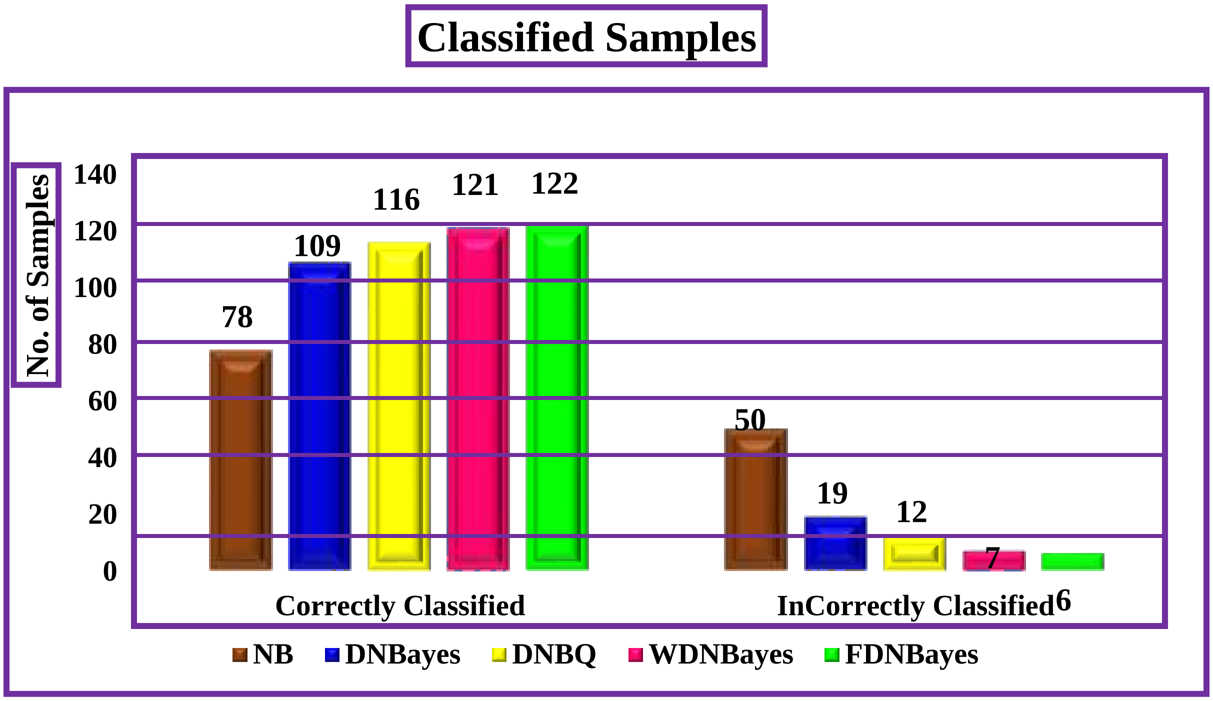

Figure 9: Classified Samples for FDN Bayes

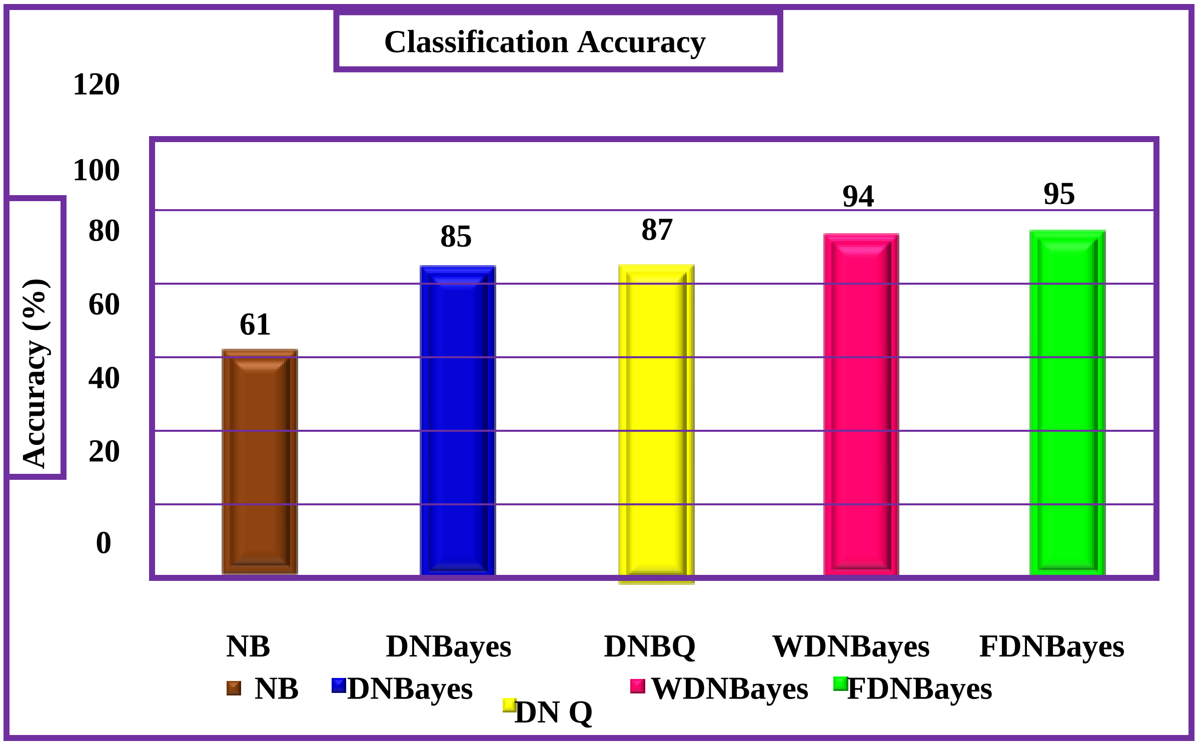

Figure 10: Classification Accuracy for FDN Bayes

Experimental results of Kappa Statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE) and evaluation metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-Score, Receiver Operator Characteristic (ROC), classified samples and classification accuracy based on FDN Bayes are shown in (Table 4 - 6) also its graphical representation is shown in (Figure 7 - 10). The results reveal that FDN Bayes outperforms existing algorithm NB (61%) and our previously proposed works DN Bayes (85%), DNBQ (87%) and WDN Bayes (94%) by producing 95% of classification accuracy in a real-time dataset with 128 samples. In the FDN Bayes model the reduced feature subset obtained through CFS filter is further optimized using Genetic Algorithm and hence shows better improvement in classification accuracy over WDN Bayes model.

8. Conclusion and Future Scope:

Various methods are adopted for selecting relevant features in the soil dataset to predict and classify as fertile and non-fertile soil. The proposed methodologies WDN Bayes and FDN Bayes provide the promising results as 94% and 95% respectively in terms of classification accuracy whereas NB, DN Bayes and DNBQ give 61%, 85% and 87% of classification accuracy with all features. Experimental results obtained for real-time soil dataset of 128 testing samples and have been proved that FDN Bayes outperforms WDN Bayes with respect to the optimized feature subset using various evaluation metrics. A wrapper method is then enforced on the reduced subgroup of features to find the feature set with high predictive accuracy. In spite of it the advantage of the CFS filter among the other filter methods is the significantly shorter computation time and one of the essential highlights is that it does not need any fine-tuning to obtain competing results. Experimental results also confirm that the developed filter-based feature extraction method FDN Bayes is superior to the other proposed model WDN Bayes and existing inbuilt feature selector methods. In addition, the efficiency of the results with fewer error measures shows improved prediction accuracy of the machine learning models. This supports soil scientists for decision making to help farmers in sugarcane cultivation. In future this work can be carried over by combining these models together for extracting relevant attributes to increase the classification accuracy.

References:

1. Hamzeh, S.; Mokarram, M.; Haratian, A.; Bartholomeus, H.; Ligtenberg, A.; Bregt, A.K. Feature selection as a time and cost-saving approach for land suitability classification (Case Study of Shavur Plain, Iran). Agriculture 2016, 6,52. [CrossRef]

2. Monzon, J.P.; Calviño, P.A.; Sadras, V.O.; Zubiaurre, J.B.; Andrade, F.H. Precision agriculture based on crop Physiological principles improves whole-farm yield and profit: A case study. Eur. J. Agron. 2018, 99, 62–71. [CrossRef]

3. Cisternas, I.; Velásquez, I.; Caro, A.; Rodríguez, A. Systematic literature review of implementations of precision agriculture. Comput. Electron. Agric. 2020, 176, 105626. [Cross Ref]

4. Rehman, T.U.; Mahmud, M.S.; Chang, Y.K.; Jin, J.; Shin, J. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. Comput. Electron. Agric. 2019, 156, 585–605. [CrossRef]

5. Saikai, Y.; Patel, V.; Mitchell, P.D. Machine learning for optimizing complex site-specific management. Comput. Electron. Agric. 2020, 174, 105381. [Cross Ref]

6. J. Hua, W.D. Tembe, E.R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recogn. 42 (3) (2009) 409–424.

7. V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray

data classification, Pattern Recogn. 45 (1) (2012) 531–539.

8. A. Jain, R. Duin, J. Mao, Statistical Pattern Recognition: a review, IEEE Transaction Pattern Anal, Mach, Intell, 22(1) (2000) 4 – 37.

9. I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recogn. 43 (1) (2010) 5–13.

10. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

11. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. (2003) 1157–1182.

12. D. Koller, M. Sahami, Toward optimal feature selection, 1996.

13. W.J. You, Z.J. Yang, G.L. Ji, PLS-based recursive feature elimination for high dimensional small sample, Knowl.-Based Syst. 55 (2014) 15–28.

14. R. Kohavi, H. George, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.

15. I. Inza, P. Larrañaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artif. Intell. Med. 31 (2) (2004) 91–103.

16. P. Bermejo, J. Gámez, J. Puerta, Speeding up incremental wrapper feature subset selection with Naive Bayes classifier, Knowl.-Based Syst. 55 (2014) 140–147.

17. P. Bermejo, L. Ossa, J. Gámez, J. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking, Knowl.-Based Syst. 25 (1) (2012) 35–44.

18. M. Gutlein, E. Frank, M. Hall, A. Karwath, Large-scale attribute selection using wrappers, in: IEEE Symposium on Computational Intelligence and Data Mining, CIDM'09, IEEE, 2009.

19. T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theor. 13 (1) (1967) 21–27.

20. E. Xing, M. Jordan, R. Karp, Feature selection for high-dimensional genomic microarray data, in: ICML, vol. 1, 2001, pp. 601–608.

21. M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (3) (1997) 131–156.

22. Kohavi, R.; John, G.H. Wrapper Approach. In Feature Extraction, Construction and Selection; Liu, H., Motoda, H.Eds.; Springer US: New York, NY, USA, 1998; Volume 453.

23. Raynukaazhakarsamy, J. G. R. Sathiaseelan, ” Outlier Detection And Denoising By The Measures Of Dispersion With Naïve Bayes To Predict Soil Fertility”, Webology (ISSN: 1735-188X), Volume 18, Number 6, 2021.