IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 7, Issue 3, June – July, 2019
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

FEATURE SELECTION METHODS FOR PREDICTION OF THE INDIVIDUAL'S STATUS OF HIV/AIDS FROM EDHS DATASET - A FILTER APPROACH

Daniel Mesafint[1], Dr. Manjaiah D H[2]
1) Department of Computer Science, Mangalore University, Katakana, India.
2) Department of Computer Science, Mangalore University, Katakana, India,.

Abstract - Lately, data is widely available in information systems and researchers have devoted much attention to data mining in transforming such data into useful knowledge. It implies the presence of low-quality, inaccurate, repeated and noisy data which has a negative effect on the method and meaningful pattern of observing knowledge. Selection of features is the mechanism that recognizes the most appropriate attributes and elimination of the redundant and insignificant attributes. In this research, a feature selection approach was conducted using filter-based feature selection methods to predict the individual status/test outcome of the Ethiopian Demographic and Health Survey (EDHS-HIV/AIDS) dataset for HIV / AIDS. The study uses three widely employed filter-based feature selection methods to validate the efficacy of the proposed feature selection methods namely: univariate, feature importance and correlation coefficient. We used seven classification algorithms to test the performance of selected features, and each classifier output is evaluated using accuracy, precision, recall, f1-score and ROC.

Among the algorithms, the classifiers namely Random Forest, K-Nearest neighbours and Gradient Boosting classifiers achieve higher accuracy levels on the EDHS-HIV/AIDS dataset than others after applying the filter-based feature selection methods. In our research, we have proved that the importance of the specified feature selection methods is improving the performance of learning algorithms.

Index Terms- Feature Selection, Filter Methods, EDHS, HIV/AIDS Status.

I. INTRODUCTION:

Data Mining is the non-trivial extraction from data archive of hidden, previously unknown and potentially valuable information [1]. Because of the availability of computers, a vast amount of data is collected in the fields of demographics and health care and it needs to be mining the useful and potential information from it.

Such vast amounts of data cannot be analyzed by health experts in a short time to take decisions and policies on the epidemic occurrence of the disease. Extracting useful knowledge from repositories for the diagnosis and treatment of diseases is becoming increasingly important. Health data mining has enormous potential to discover the secret trends in data sets for health domains. Pre-processing of data is an essential phase in the cycle of information development, since quality assessments will be made on quality results.

Data preprocessing involves data cleaning, filling out missing values, data creation, data transformation and data reduction [1].

Data quality in the demographics and wellbeing report increases health outcomes. The goal of data reduction or subset selection of features is to find minimum subset of variables so that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. The following benefits require mining on the lesser set of attributes.

- It reduces the number of attributes that occur in the patterns found, thereby helping to promote pattern comprehension.
- It reduces the time required to learn classifiers.
- It enhances precision or accuracy during classification.

In recent years there has been a considerable increase in the requirement to use feature selection methods in health datasets. This is because most health datasets have a significant number in high-dimensional characteristic samples.
This makes inefficient, expensive in computation and produces less classification accuracy when using a whole set of inputs. Therefore, relevant features required for the classification purpose should be obtained by applying on appropriate feature selection method.

The selection of features used as the primary source of knowledge in model creation for any learning algorithm is extremely critical in choosing an optimal subset that will be reflective of the original set. The selection of an optimal subset of appropriate and non-redundant features is a challenging task. If there is a trend off and if too many features are picked, the classifier may have a heavy workload and can reduce the accuracy of the classification. In the other side, if very few

features are chosen, there is a risk that features would be omitted which might have improved the accuracy of the classification. An optimum subset of appropriate and non-redundant features will therefore be collected, which will offer an ideal solution without decreasing the precision of the classification. Although several methods of selection of features are available, no known successful methods been developed for selecting the optimal subset.
Selection of features helps to understand data, reduce computational requirements, reduce the dimensionality curse and improve the performance of predictions [2].

This paper presents filter based feature selection methods namely (I) Univariate Chi-squared (II) Feature Importance and (III) Correlation coefficient. These are employed with seven classifiers namely Random forest, k- nearest neighbors (KNN), Support vector machines (SVM), Naive Bayes, Logistic Regression, Ada-boost and Gradient Boosting.

The remaining part of the paper is structured as follows: The related works related to this paper are discussed in Section II, Section III discusses the feature selection methods used in this paper, Section IV presents the classification algorithms as an overview, Section V explains the description of the data variables for EDHS-HIV / AIDS data set, and the experimental results of each selection method are discussed in Section VI with the classification algorithms and finally the paper is finalized in Section VII with concluding remarks.

II. RELATED WORKS:

In this section, we have presented the review works related to feature selection and classification algorithms essential in data analytic. The recent literature includes several works incorporating methods for selecting features, including methods for filter methods.

Sarkar et al. [3] presents an empirical study comparing the efficiency of few feature selection techniques that is Chisquared, Information Gain, Mutual Information and Symmetrical Uncertainty used with various classifiers such as Naïve Bayes, SVM, Decision Tree and KNN. Present the results of feature selection methods on text datasets for different classifiers. The analysis further allows the relative output of the classifiers to be correlated with the methods.

Roslina et al. [4] using SVM to forecast hepatitis deceases using wrapper feature selection approach to identify specific characteristics prior to classification. The combination of SVM and wrapper methods produced strong classification results.

Pinar Yildirim. [5] studies the use of filter-based methods for selecting features on hepatitis data set. The researcher contrast methods for filter-based features selection such as Info Gain, Consistency Subset, RelieF and One-R. The efficacy of the algorithms is checked using four classification algorithms and used for classification from the classification algorithms which, Naive Bayes and Decision Table classifiers were chosen due to the higher precision levels.

Asha Gowda et al. [6] proposed a genetic algorithm (GA) filter method and correlation-based feature selection in cascading mode to filter a subset of features from four UCI publicly available medical data sets. The methods are tested by the researcher with five classifiers namely Decision tree, Naive Bayes, Bayesian, Radial basis function, and K-Nearest Neighbor.

Harb et al. [7] the paper proposes the Particle Swarm Optimization (PSO) with filter and wrapper approaches as a feature selection method for the three medical data sets. And compare the output of the proposed methods to another feature selection algorithm focused on a Genetic approach. PSO has shown improved classification accuracy for five classifiers namely k-nearest neighbor, Decision tree, Radial Basis function, Naive Bayes, and Bayesian.

Khan et al. [8] reviews various methods in filter, wrapper and embedded selection methods that help pick optimal subsets of functions. However, the article shows selection effects on various machine learning algorithms such as Random Forest, KNN and Naive Bayes. Results demonstrated numerous influences on the degree of precision when choosing features at specific margins.

Rajit and Amit [9] discusses methods which are select best and select percentile based on function selection. The work also shows how the selection of features works and helps during the classification process. The researchers have attempted to show how accuracy in classification algorithms used in machine learning has been enhanced through these methods of selection of features. They used five classification algorithms such as K-Nearest Neighbor, SVM, Naive Bayes, and Logistic Regression and suggested method to improve the accuracy.

Li et al. [10] describes a sub-set selector function using a heuristic correlation to evaluate the goodness of the sub-set function and to check its effectiveness with three common machine learning algorithms: a Naive Bayes classifier, a Decision tree inducer and an instance-based learner. Experiments use common data sets drawn from real and artificial realms.

III. FEATURE SELECTION METHODS:

Selection of features is a preprocessing strategy used in machine learning to delete unused and redundant attributes in order to increase algorithm accuracy [2].

Selecting features means not only a reduction in cardinality but also the collection of attributes that may be dependent on the existence of interaction between the classification algorithm and the attributes. The learning models appear to become computationally complicated, over-fit, less comprehensible and less reliable in the presence of several trivial features, some of which don't bring any value throughout the learning process.

For machine learning, variable selection is the method of choosing a subset of appropriate features from a wide range of features, without sacrificing performance quality. The feature selection process is depicted in Figure 1.

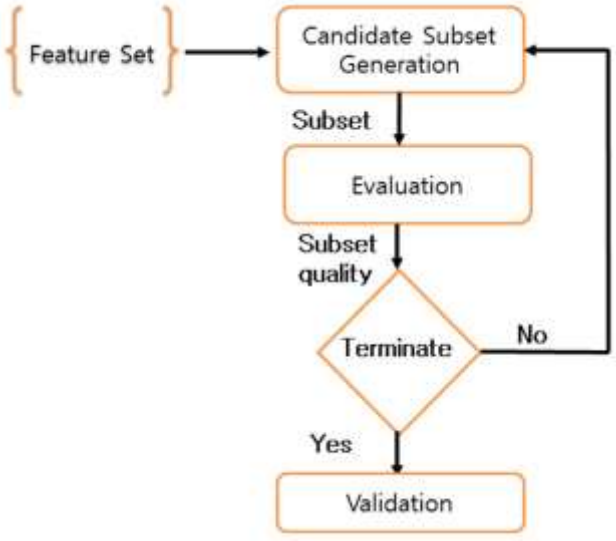

Fig. 1. Feature Selection Process [11]

A. Filter Based Feature Selection Method

The feature selection methods can be classified into Filter, wrapper, embedded and hybrid in the sense of classification. In this paper we focus on the filter based methods of feature selection techniques with Univariate, Feature importance and Correlation coefficient.

The filter based feature selection method selects a subset of features that preserves the relevant information found throughout the whole set of features as much as possible. Methods utilizing the filter method are independent of some single algorithm, as their evaluation feature relies entirely on data properties [12].

The importance of the features is determined by the intrinsic properties of the data being considered. It includes measuring a feature importance score and eliminating less score features and using the remaining feature subset as input to the algorithm.

The solutions to the filter methods are usually independent of the learning induction algorithm. Filters estimate an index of relevance for each feature to determine how important a feature is to the target, then rate features by their relevance indices and conduct rank searches or based on some statistical criteria [6].

- *Univariate Feature Selection*

Univariate feature selection tests each attribute individually to determine the intensity of the relation between the feature and the variable response. These methods are simple to run and understand, and are particularly better in general for gaining a better understanding of data (but not necessarily for optimizing the set of features for better generalization).

The univariate filter methods are the type of methods where specific criteria are used to rank the individual characteristics, and then the top N features are selected. For univariate filter methods, different types of ranking criteria are used such as fishery score, mutual information and feature variance. In this paper we select best features from the total original feature using chi-square methods.

Chi is a statistical test which measures a feature's independence from the class labels. It is a bidirectional metric. Forman [13] noted that this test can act erratically when feature counts are low to be predicted. With class imbalanced data sets this approach is relatively popular.

To pick the desired number of features with best Chi-square scores, We measure the Chi-square between the target and each feature. Table I present the sample of univariate feature selection using chi-square methods and we select best 20 features among the total of 26 features from EDHS-HIV/AIDS data set.

TABLE I
SELECTED FEATURES USING UNIVARIATE FEATURE SELECTION METHODS

RANK	SPECS	SCORE
18	H_STI	4826.220044
23	S_TEST	4653.720019
24	T_IN_LAB	1087.105393

2	Reg	933.170027
22	P_T_HIV	879.389418
17	HIV_Mosq	465.323441
0	Sex	447.536106
11	N_S_Part	345.145407
6	Edu_At	258.215903
5	Edu_Lvl	147.122366
7	M_Sta	121.090588
19	H_O_STI	119.945452
21	E_T_HIV	80.808937
10	R_SeA	57.726482
1	Age	32.821387
14	R_Use_Con	31.845463
12	Had_Sex	22.130725
20	H_AIDS	11.966975
4	Rel	11.696209
8	C_Wor	9.172250

- *Feature Importance*

Feature importance is a selection criterion which determines the significance of a feature by calculating the gain in information relative to the target class. This technique gives you a score for each feature of the results, the higher the score is the feature for your performance variable, the more significant or appropriate.

In this work we use the selection criteria for these features to select best features that have higher score. Figure 2 shows how feature choices are made using feature importance selection. The features are picked using Extra Tree Classifier to extract the top 20 features from the EDHS-HIV / AIDS dataset.

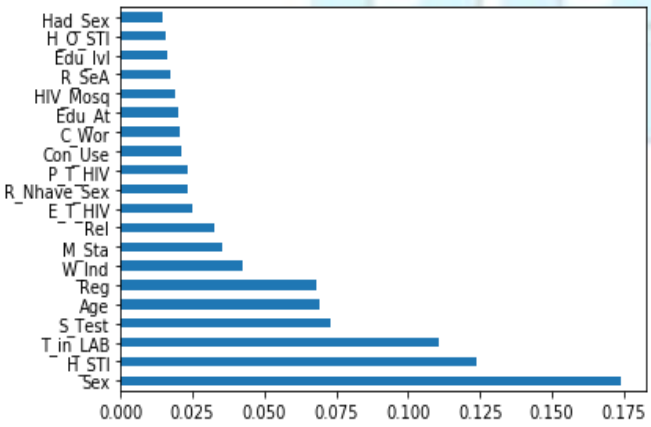

Fig. 2. Selected Features using Feature Importance criteria

- *Correlation Coefficient (CC)*

The correlation-based feature selection approach tests subsets of features by selecting subsets of features containing features that are strongly correlated, but do not correlated with each other. CC assesses a subset by considering individually the potential predictive of each features and also the level of redundancy or similarity thereof. This implies that, provided a function, the algorithm will decide on its next step by choosing the alternative that maximizes this function's output [14].

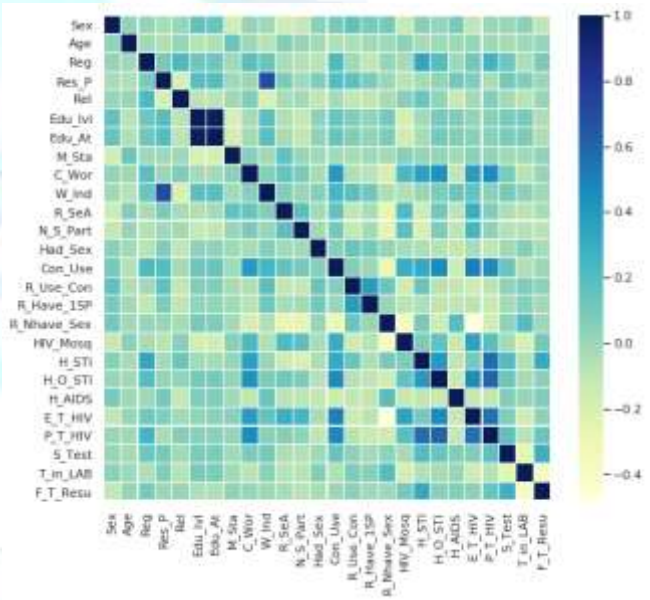

Fig. 3. Correlation Coefficient with heat-map

Correlation can be positive, meaning that the increase in one value of the feature increases the target variable or negative value which is increases in one value of the feature decreases the other variable value. In this paper, we used Correlation coefficient with Heat-map which makes it easy to identify which features are most closely related to the target variable. Figure 3 shows how the target variables correlate the features.

IV. CLASSIFICATION ALGORITHMS:

This section provides a brief description of an algorithm used in the analysis. There are a broad variety of classification algorithms with its strengths and disadvantages. There's no particular learning method that fits well on all supervised

learning problems. For the purpose of selecting features and testing the accuracy of each feature selection, we used some of the classification algorithms for each selected features. The classification algorithms which tested in this work are RF, KNN, SVM, Adaboost, Logistic Regression, Gradient Boosting and Naive Bayes.

- **Random Forest:** RF algorithms construct a family of classification methods that rely on the multiple decision trees combined. These Classifier Ensembles (EoC) have a peculiarity in increasing their tree-based components from certain number of randomness. Based on that concept, RF is characterized as a collection of randomized decision trees ensembles generic principles [15]. RF's core unit the so-called foundation learner is a binary tree constructed using recursive partitioning (RPART).

- **K-Nearest neighbors:** KNN is one of the most important algorithms, a non-parametric model, and a supervised learning algorithm [16]. The classification rules are generated by the training samples themselves, without additional data. The KNN classification algorithm determines the test sample category according to the K training samples which are the nearest neighbors to the test sample, and tests it according to the category with the highest category likelihood.

- **Support Vector Machine:** SVM is a classification algorithm which has multiple kernel options depending on the fashion of the distribution of data. It can classify data in multiple linear ways but SVM gives us the optimal among all the possible options. Types of kernel in SVM are linear, rbf, poly, sigmoid.

- **Naive Bayes:** NB is a simple model of generative probabilistic classification model that assumes independence between characteristics of the objects to be classified [17]. Therefore, the Naive Bayes classifier applies Bayes theorem with the assumption that the presence or absence is unrelated to other features. Despite his assumption of independence, its effectiveness in classification has been proven [18]. Moreover, to approximate the parameters required for the classification, Naive Bayes needs only a small amount of training data.

- **Logistic Regression:** LR is a statistical approach for evaluating a collection of data where there is one or unknown independent variables that determine the outcome. The effect is calculated using dichotomous equation (in which only two outcomes are possible). The dependent variable in logistic regression is dichotomous or binary which contains only data coded as either 1 (TRUE) or 0 (FALSE) [19].

- **AdaBoost:** AB performs the classification by selecting only those discrete features that can best be distinguished between the classes [20]. The most influential algorithm within the Boosting family is AdaBoost. It preserves the distribution of one set of probabilities of training samples, and changes the distribution of probabilities during each iteration for each study. The member-classifier is developed using a specific learning algorithm, and its error rate is calculated on the training samples. AdaBoost uses the error rate to adjust the probability distribution of training samples [21].

- **Gradient Boosting:** GB is an algorithm that iteratively builds and improves a set of decision trees, each one conditioned and pruned on instances that previously learned trees have passed through. The previous trees wrongly labeled instances are re-sampled with higher likelihood to give a new distribution of likelihood for the next iteration. Gradient boosting is a method for regression and classification problems in machine learning.

V. DATA SET DESCRIPTION:

The HIV/AIDS data set was obtained from Ethiopia Demographic and Health Survey (EDHS), data set from Central Statistics Agency (CSA). It includes 78,877 instances, out of which 55,209 instances belongs to one class and 23,668 instances of another class. 26 attributes define cases, some of which are numerical and some nominal, and the performance type to be expected is negative or positive. For preprocessing purposes the nominal attributes are modified or converted into numeric. This data set has a class shows whether the individual are Negative or Positive. Table II presents the detail description of the data set used in our study.

TABLE II
DESCRIPTION OF THE DATA SET FROM EDHS DATASET

No.	Variables	Values	Description
1.	Sex	M, F	Gender of Individual
2.	Age	Continues	Age of Individual
3.	Reg	1,2,3,4,5,6,7,8, 9,10,11	Region of Living
4.	Res_P	1,2	Place of Living Urban/Rural
5.	Rel	1 – 15	Religion
6.	Edu_lvl	0,1,2,3,4	Educational Level
7.	Edu_At	0,1,2,3,4,5	Attainment of Education

8.	M_Sta	0,1,2,3,4,5	Marital Status
9.	C_Wor	0,1	Working Status
10.	W_Ind	1,2,3,4,5	Wealth Index
11.	R_SeA	0,1	Resent Sexual Activity
12.	N_S_Part	0,1	No of Sex Partner
13.	H_Sex	0,1	Had ever Sex
14.	Con_Use	0,1	Usage of Condom
15.	R_Use_Con	0,1	Reducing Usage of Condom
16.	R_Have_1SP	0,1	Reduce Sexual Partner to One
17.	R_NHave_Sex	0,1	Reduce HIV without have Sex
18.	HIV_Mosq	0,1	HIV transfer by Mosquito
19.	H_STI	0,1	Heard Sexual transmit infection
20.	H_O_STI	0,1	Heard Other STI
21.	H_AIDS	0,1	Heard about HIV
22.	E_T_HIV	0,1	Ever tested HIV before
23.	P_T_HIV	0,1	Place of HIV test
24.	S_Test	0,1	Sample Test
25.	T_in_LAB	0,1	Test in Laboratory
26.	F_T_Resu	0,1	Final Test Result

VI. EXPERIMENTAL RESULTS:

EDHS-HIV / AIDS was used to evaluate numerous filter based feature selection methods for predicting individual test status. To test the classification accuracy, seven classification algorithms mentioned above were considered. The methods of classification of the functions implemented in this paper are Univariate feature selection, Feature Importance and Correlation coefficient.

In the Univariate feature selection methods Chi-square is used to find relevant features in the HIV/AIDS data set and then classification algorithms are added to the chosen features to verify the sorting methods of the system. During feature selection task 20, 15, and 10 features were chosen by the feature selection algorithms. After selecting the required features same experiment was repeated for seven classifiers.

In this work evaluation metrics are used to evaluate the performance of the algorithms in the selected features. The most widely used evaluation metrics are accuracy, precision, recall, and confusion matrix. The next part will show each evaluation metrics as follows:

Accuracy: Is the amount of accurate predictions determined by the overall number of predictions multiplied by hundred to give it a percentage.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Recall: the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Negatives (FN). Another way to express is the number of positive predictions divided by the number of positive class values in the test data.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Precision: is calculated based on the number of True Positives (TP) divided by the number of True Positives (TP) and False Positives (FP). In another way the number of positive predictions divided by the total number of positive class values predicted.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

F1-measure: is calculated based on precision and recall

$$F1 - Measure = 2\frac{Precision * Recall}{Precision+Recall} \quad (4)$$

ROC: is commonly employed when determining statistical outcomes. Those are the instances with true positive situations for the false positive figure on the X and Y axes.

Confusion Matrix: is a metric shows correctly classified and Miss-classified samples from a given test data. Table III shows the confusion matrix used in this work.

TABLE III
CONFUSION MATRIX

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True Positive	False Positive
	Negative	False Negative	True Negative

Table IV shows the experimental results of the proposed method. The highest accuracy values are obtained for the HIV/AIDS data set using Univariate Feature selection with Random Forest, Gradient Boosting, K-Nearest neighbors, Support Vector Machine, Ada-boost, Naive Bayes and Logistic Regression classifiers are 0.95, 0.91, 0.90, 0.83, 0.81, 0.76 and 0.76 respectively. The above scores are obtained from the

Classification Algorithm	Feature Selection Methods	No of Selected Features	Evaluation Metrics				
			Accu	Prec	Rec	F1-S	ROC
Random Forest (RF)	Univariate	20	0.952	0.925	0.984	0.954	0.95
		15	0.890	0.842	0.962	0.898	0.89
		10	0.890	0.842	0.962	0.898	0.89
	Feature Importance	20	0.952	0.885	0.970	0.925	0.92
		15	0.922	0.885	0.970	0.925	0.92
		10	0.869	0.817	0.950	0.879	0.87
	Corrolation matrix	9	0.856	0.879	0.827	0.852	0.86
K- nearest Neighbor (KNN)	Univariate	20	0.901	0.871	0.942	0.905	0.90
		15	0.866	0.827	0.927	0.874	0.87
		10	0.828	0.827	0.833	0.830	0.83
	Feature Importance	20	0.900	0.868	0.943	0.904	0.90
		15	0.883	0.848	0.932	0.888	0.88
		10	0.848	0.806	0.917	0.858	0.85
	Corrolation matrix	9	0.844	0.848	0.844	0.846	0.84
Support Vector Machine (SVM)	Univariate	20	0.833	0.776	0.937	0.849	0.83
		15	0.823	0.775	0.911	0.838	0.82
		10	0.821	0.773	0.908	0.835	0.82
	Feature Importance	20	0.791	0.733	0.917	0.814	0.79
		15	0.783	0.716	0.940	0.812	0.78
		10	0.821	0.773	0.909	0.836	0.82
	Corrolation matrix	9	0.830	0.774	0.931	0.845	0.83
Naïve Bayes (NB)	Univariate	20	0.760	0.741	0.804	0.771	0.76
		15	0.732	0.692	0.845	0.761	0.73
		10	0.730	0.690	0.845	0.759	0.73
	Feature Importance	20	0.720	0.676	0.856	0.755	0.72
		15	0.709	0.660	0.872	0.751	0.71
		10	0.709	0.660	0.872	0.751	0.71
	Corrolation matrix	9	0.738	0.707	0.820	.760	0.74
Logistic Regression (LR)	Univariate	20	0.758	0.743	0.801	0.771	0.76
		15	0.732	0.696	0.842	0.762	0.73
		10	0.731	0.690	0.844	0.759	0.73
	Feature Importance	20	0.779	0.805	0.903	0.851	0.78
		15	0.700	0.640	0.913	0.753	0.70
		10	0.700	0.636	0.933	0.757	0.70
	Corrolation matrix	9	0.738	0.711	0.819	0.761	0.74
AdaBoost (AB)	Univariate	20	0.811	0.790	0.854	0.821	0.81
		15	0.782	0.773	0.806	0.789	0.78
		10	0.770	0.770	0.774	0.772	0.77
	Feature Importance	20	0.817	0.793	0.861	0.826	0.82
		15	0.798	0.774	0.841	0.806	0.80
		10	0.776	0.774	0.787	0.780	0.78
	Corrolation matrix	9	0.786	0.783	0.803	0.793	0.79

selected feature with 20 features from the total features compared with selected features of univariate such as feature 15 and 10.

Similarly, in the Feature Importance Feature selection methods, the experimental result is obtained with same classifiers which are used in univariate methods and the result with Random Forest, Gradient boosting, K-Nearest neighbors, Adaboost, Support Vector Machine, Logistic Regression and Naïve Bayes are 0.92, 0.91, 0.90, 0.82, 0.79, 0.78 and 0.72 respectively.

In this Experiment, we also test Recursive feature selection methods with 20, 15 and 10 features and experiments are done with same classifier as taken from FFS and BFS methods. The experimental result are obtained with Random Forest, Gradient Boosting, K-Nearest neighbors, Ada-boost, Support Vector Machine, Logistic Regression and Naive Bayes is 0.948, 0.909, 0.903, 0.821, 0.815, 0.714 and 0.713 respectively.

In addition to the above evaluation metrics, we used ROC as a parameter to evaluate the selected features. As table III shows that the ROC result of each selected features with the same classifiers obtained to compare one to the others.

TABLE IV
CLASSIFICATION RESULT OF FILTER BASED FEATURE SELECTION METHODS ON HIV/AIDS DATASET

Gradient Boosting (GB)	Univariate	20	0.905	0.906	0.907	0.907	0.91
		15	0.880	0.887	0.887	0.882	0.88
		10	0.851	0.876	0.819	0.847	0.85
	Feature Importance	20	0.905	0.904	0.908	0.906	0.91
		15	0.897	0.905	0.890	0.897	0.90
		10	0.876	0.894	0.855	0.874	0.88
	Corrolation matrix	9	0.878	0.889	0.942	0.915	0.88

VII. CONCLUSION AND FUTURE WORK:

Feature selection is an important phase in data mining studies to process data, and in other machine learning algorithms, large volumes of irrelevant features can hardly cope. Therefore, the approaches to feature design became a prerequisite for many experiments.

In this research, a proposed feature selection approach was conducted using filter-based feature selection methods to predict the individual status or test outcome of HIV / AIDS from the EDHS data set. The research uses three methods of selecting features under the filter base to classify the data set and assess their output using Random Forest, KNearest neighbors, Support Vector Machine, Gradient Boosting, AdaBoost, Naive Bayes and Logistic Regression. The performance was measured by five evaluation metrics namely: Accuracy, Precision, Recall, F1-measure and ROC.

From the experiment, Random Forest, K-Nearest neighbors and Gradient Boosting classifiers have higher accuracy levels on EDHS-HIV/AIDS data set than the others after the applying the filter based feature selection methods. This study shows that methods of selecting features are capable of improving the learning algorithms efficiency.

Finally, the output of this study can make significant contributions in the prediction of the status of HIV/AIDS result of individuals in health domain research and provide filter based feature selection methods for machine learning studies. As a future work, a research will be designed as a potential job to explore the other methods of selecting features which to compete with filter based on the efficiency of feature selection methods and classification accuracy.

VIII. ACKNOWLEDGEMENT:

The authors are thankful for all positive reviews and suggestions to the publisher and the anonymous reviewers.

REFERENCES

[1] M. K. Jiawei Han, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2006.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014.

[3] S. D. Sarkar and S. Goswami, "Empirical study on filter base feature selection methods for text classification," International Journal of Computer Applications, vol. 81, no. 6, 2013.

[4] A. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using support vector machines and wrapper method," in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5. IEEE, 2010, pp. 2209–2211.

[5] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease," International Journal of Machine Learning and Computing, vol. 5, no. 4, p. 258, 2015.

[6] A. G. Karegowda, M. Jayaram, and A. Manjunath, "Feature subset selection using cascaded ga & cfs: A filter approach in supervised learning," International Journal of Computer Applications, vol. 23, no. 2, pp. 1–10, 2011.

[7] H. M. Harb and A. S. Desuky, "Feature selection on classification of medical datasets based on particle swarm optimization," International Journal of Computer Applications, vol. 104, no. 5, 2014.

[8] M. Khan and S. Quadri, "Effects of using filter based feature election on the performance of machine learners using different datasets," BVICA M's International Journal of Information Technology, vol. 5, no. 2, p. 597, 2013.

[9] R. Nair and A. Bhagat, "Feature selection method to improve the accuracy of classification algorithm," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 6, pp. 124 – 127, 2017.

[10] B. Li, Q. Wang, and J. Hu, "Feature subset selection: a correlation-based svm filter approach," IEEJ Transactions on Electrical and Electronic Engineering, vol. 6, no. 2, pp. 173–179, 2011.

[11] K. Siddique, Z. Akhtar, H.-g. Lee, W. Kim, and Y. Kim, "Toward ulk synchronous parallel-based machine learning techniques for anomaly detection in high-speed big data networks," Symmetry, vol. 9, no. 9, p. 197, 2017.

[12] G. Brown, A. Pocock, M.-J. Zhao, and M. Luj´an, "Conditional likelihood maximization: a unifying framework for information theoretic feature selection," Journal of machine learning research, vol. 13, no. Jan, pp. 27–66, 2012.

[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of machine learning research, vol. 3, no. Mar, pp. 1289–1305, 2003.

[14] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[15] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[16] B. V. Dasarathy, "Nearest neighbor (nn) norms: Nn pattern classification techniques," IEEE Computer Society Tutorial, 1991.

[17] T. O. Ayodele, "Introduction to machine learning," New Advances in Machine Learning, pp. 1–9, 2010.

[18] I. Rish et al., "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41–46.

.