Vol. 4, No. 1 | January - June 2020

Identifying the Machine Learning Techniques for Classification of Target Datasets

Abdul Ahad Abro[1], Mohammed Abebe Yimer[2], Zeeshan Bhatti[3]

Abstract:

Given the dynamic and convoluted nature of numerous datasets, the necessity of enhancing performance outcomes and handling multiple datasets has become more challenging. To handle these issues effectively and improve the quality of multiple approaches, the capabilities of various Machine Learning techniques such as K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes(NB) and Support Vector Machine (SVM) have been utilized in this study. In this paper, the binary classification method using five different datasets, and many predictor variables have been utilized. Moreover, this research has mainly focused on determining the classification of data into the subsets that share the standard designs. In this regard, many approaches had been studied extensively and used to achieve better yields from the existing literature; however, they were inadequate to provide efficient outcomes. By applying four Supervised ML classification algorithms along with the UCI Datasets of ML Repository, the robustness of the method is progressed. The proposed mechanism is assessed by adopting five performance criteria concerning the accuracy, AUC (Area Under Curve), precision, recall, and F-measure values. The current study experimental results revealed that there is a significant improvement in the confusion matrix rate compared with a similar study and this method can also be used for machine learning problems such as binary classification.

Keywords: *Machine Learning, Data Mining, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Support Vector Machine*

1. Introduction

The combination of classifiers is now an active research area in the ML and Pattern Recognition [1][2][3]. Many theoretical and empirical studies have been published which show the advantages of the combination paradigm over the individual classifier models [4][5]. A significant number of researches have been conducted to design multiple classifier systems based on the same classifier models trained on different data or feature subsets. ML has been widely used in a variety of industries, such as Remote Sensing, Image Classification, and Pattern Recognition.

ML can learn and improve automatically from experience, without explicit programming. It is the primary aim to automate learning without human intervention. ML algorithms use statistics to find patterns in massive amounts of data [6]. Whereas the algorithms which are used in this research are briefly described below:

[1]Department of Computer Engineering, Ege University, Turkey.
[2]Department of Computer Engineering, Dokuz Elyul University, Turkey.
[3]Department of Information Technology, University of Sindh, Pakistan
Corresponding Author: zeeshan.bhatti@usindh.edu.pk

Firstly, KNN: a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. It is used for statistical estimation and pattern recognition[7][8].

Secondly, LR: a standard statistical approach that is ideal for performing regression analysis where the dependent variable is binary. It is used to describe the data and to explain the relationship between one dependent binary variable with one or more independent nominal, ordinal, interval, or ratio-level variables [9].

Thirdly, the NB classifier: combines the Bayes paradigm with the decision rules like the hypothesis, which provides satisfactory results. It applies Bayes theorem, with the naive assumption of conditional independence between each pair of features given the value of the class variable. In [10], proposed the NB learning framework for large-scale computational efficiency and multi-domain platform classification.

Fourthly, SVM: is a paradigm that uses classification algorithms for two-group problems. It is accuracy and predictive performance on the survival of traumatic brain injuries performed significantly better than LR [11].

On the other hand, this paper has structured with several sections. In section 2, previous related work is described briefly. The methodology adopted for performing different experiments is explained in Section 3. Section 4, provides experimental work, datasets detail, evaluation of experiments is performed to obtain different results. Lastly, certain conclusions are drawn based on the outcomes and future work is suggested in Section 5.

2. Related Work

Classifications based on KNN, LR, NB, and SVM has recently witnessed a surge of research efforts. In this paper, we have used the classification of supervised learning. Moreover, in the literature, classification algorithms could be affected significantly or negatively by some features [1][2]. The goal of classification is to accurately forecast the target class for each case in the data. Whereas in the model build training process, a classification algorithm co-ordinates between the values of the predictors and the values of the target. Different classification algorithms execute different procedures for discovery associations. These associations are model, which can function to a different dataset in which the class is unidentified [12] [13] [14].

In [15], KNN is the slowest classification technique because the classification time is directly proportional to the number of data. When the data size is more prominent, more extensive distance calculation should be performed to make it extremely slow. Moreover, it uses the number of nearest neighbors "k" as one of the parameters in classifying an object, and the value of k influences the classifier performance [16].

In [17], Cubic SVM, Quadratic SVM, and Linear SVM have better performances in predicting the outcome of traumatic brain injury as compared to LR.

In [18], NB is the most popular data mining algorithms. Empirical results indicate that the selective NB demonstrates superior classification performance while retaining the simplicity and flexibility at the same time.

SVM is a useful method for solving classification and regression problems. In [19], the SVM approach can substantially improve prediction accuracy and would help to mitigate the adverse impact on urban expansion.

3. Methodology

This section presents an overview of the proposed method, which describes the pre-processing stage of data and classification algorithms.

3.1. Overview of the Proposed System

An overview of the proposed system is given in Fig. 1. This system consists of numerous phases: datasets, base learners, and comparative analysis of the results. Besides, the generalization performance of the system, 10-fold cross-validation is used for all classifier learners and datasets.

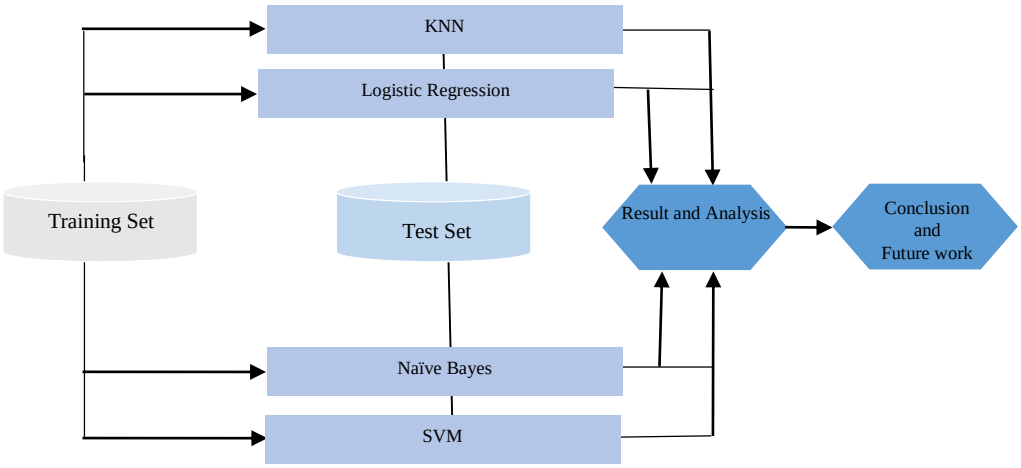

Fig.1. The framework of the method.

3.2. Data Preprocessing

In this phase, the ranges of the values of the data from different ML datasets may be high. In such a scenario, certain features can significantly or negatively affect algorithms for classification accuracy. Therefore, the data values are normalized to [0,1] range using min-max normalization technique [20].

3.3. Classification of Algorithms

In this study, four base learners, including KNN, logistic regression, NB, and SVM, are employed.

There are numerous phases of methods related to the datasets and classifiers focused on ML. In this work, four ML classifiers, along with several datasets, are experienced for binary classification.

LR classifier relies on feature extraction. Typically, it delivers more authentic results than KNN, NB, and SVM. The primary aim of this analysis is to establish the classification accuracy and performance evaluation of multiple datasets.

The KNN classifier does not have a specialized training phase and uses all the data for training during classification and it does not assume anything about the underlying data [15].

LR classifier is another method borrowed by ML from the field of statistics. It is a statistical model and used when the dependent variable is categorical.

NB is a probabilistic ML model. It requires linear parameters in the number of functions of the variables and highly scalable [18].

SVM is an ML algorithm that can be used for classification problems as well as for regression. It is segregated in two classes and co-ordinates the individual observation.

4. Experimental Design

In these subsections, we describe and present the experimental process, evaluation measures, and experimental results.

4.1. Experimental Process

In the experimental process, five datasets have been used from the UCI ML Repository [21]. All experiments are performed on a total of 4 ML classifiers by using WEKA (Waikato Environment for Knowledge Analysis) ML toolkit and JAVA programming language

[22]. On the one hand, we have utilized default parameter values for all the classifiers in WEKA.

On the other hand, we have carried out 10-fold cross-validation to all datasets to yield reliable results. The 10-fold cross-validation is imposed on the original dataset randomly partitioned into 10 equally sized sets, one of which is used as test validation, while the remaining sets are used for training operations. The process is repeated 10 times and calculates the averages of the results.

Dataset characteristics are evaluated concerning the attributes and the number of instances. These datasets are typically used to solve ML related issues. There are various numerical attribute descriptions illustrated in Table 1. The number of instances, attributes, and classes for each dataset are presented in Table 1. The datasets are selected from the UCI ML Repository according to their distinct parameters. It is determined by investigating the appropriate data or datasets which are being utilized for binary classification problems.

Table 1. Characteristics of the five Datasets Used in This Study

Datasets	Instances	Attributes	Classes
Annealing	898	39	6
Breast Cancer	286	10	2
Hepatitis	155	20	2
Vertebral	240	7	2
Yeast	1484	9	10

In this work, four different ML approaches have been carried out along with the five datasets, which are considered suitable for the classification. However, the performance metrics are calculated according to the binary classification problems based on the confusion matrix.

4.2. Assessment of Measures

This section describes the five performance evaluation measures of the proposed method, consisting of accuracy, AUC, precision, recall and F-measure.

Accuracy reflects how close an agreed number is to a measurement. It is specified further in Eq.1.

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

In equation 1, TN, FN, FP and TP show the number of True Negatives, False Negatives, False Positives and True Positives.

AUC represents the area under the ROC Curve. AUC calculates the whole two-dimensional area beneath the whole ROC curve from (0,0) to (1,1).

Precision is a positive analytical value [1][23]. Precision defines how reliable measurements are, although they are farther from the accepted value.

The equation of precision is shown in Eq.2.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The recall is the hit rate [1][23]. The recall is the reverse of precision; it calculates false negatives against true positives. The equation is illustrated in Eq. 3.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F-measure can be defined as the weighted average [1][24] of precision and recall. This rating considers both false positives and false negatives. The equation is illustrated in Eq. 4.

$$F - measure = \frac{2}{1/precision + 1/recall} \quad (4)$$

These criteria are adjusted proportionally in the data by the reference class prevalence in the weighting operation.

4.3. Experimental Results

Tables 2-6 for all datasets present accuracy, AUC, precision, recall, and F-measurement weighted values with ML algorithms. In Table 2-6, high Acc, AUC, Precision, Recall, and F-measure are shown in Bold, while the greyed shows insufficient results.

To sum up, Tables 2-6, has been designed in terms of different specifications according to

the multiple datasets relating to the numerous approaches of ML. In Table 2, LR has better outcomes, which provides 99.1091% Acc when comparing with others.

Probably, in Table 3, KNN indicates 72.3776% Acc adequate consequences. Similarly, in Table 4, the NB presents 84.5161% Acc effective results. Whereas, in Table 5, the SVM illustrates the 92.9167% Acc productive outcomes. However, in the end, LR shows a 58.6253% Acc result in Table 4.

The annealing, hepatitis, and vertebral datasets have significant outputs concerning the accuracy, AUC, precision, recall, and F-measure parameters in Table 2, 4, and 5; however, breast cancer has somehow satisfactory output in Table 3 and yeast shows lower outcomes in Table 6.

Furthermore, it is analyzed that LR for annealing dataset in Table 2, provides a more accurate outcome. Likewise, KNN in breast cancer dataset concerning Table 3, indicates adequate consequences and in Table 4, NB presents effective results in the Hepatitis dataset. In addition, in Table 5, Vertebral dataset SVM provides positive findings. Finally, LR indicates the progressive result in Table 6, yeast dataset.

Table 2: Weighted Values for Annealing Dataset

Annealing					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
KNN	99.1090	0.985	**0.991**	**0.991**	**0.991**
LR	**99.1091**	**0.992**	0.991	0.991	0.991
NB	86.3029	0.957	0.933	0.863	0.882
SVM	39.3096	0.646	0.703	0.393	0.433

Table 3: Weighted Values for Breast Cancer Dataset

Breast Cancer					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
KNN	**72.3776**	0.628	0.699	**0.724**	0.697
LR	68.8811	0.646	0.668	0.689	0.675
NB	71.6783	**0.701**	**0.704**	0.717	**0.708**
SVM	66.0839	0.596	0.662	0.661	0.661

Table 4: Weighted Values for Hepatitis Dataset

Hepatitis					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
KNN	80.6452	0.653	0.794	0.806	0.799
LR	82.5806	0.802	0.814	0.826	0.818
NB	**84.5161**	**0.860**	**0.853**	**0.845**	**0.848**
SVM	79.3548	0.731	0.814	0.794	0.802

Table 5: Weighted Values for Vertebral Dataset

Vertebral					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
KNN	85.4167	0.660	0.852	0.854	0.853
LR	92.5	**0.930**	0.919	0.925	0.920
NB	77.9167	0.854	0.886	0.779	0.812
SVM	**92.9167**	0.788	**0.924**	**0.929**	**0.925**

Table 6: Weighted Values for Yeast Dataset

Yeast					
Methods	**Acc (%)**	**AUC**	**Precision**	**Recall**	**F-Measure**
KNN	52.2911	0.685	0.524	0.523	0.522
LR	**58.6253**	**0.825**	**0.585**	**0.586**	0.577
NB	57.6146	0.816	0.585	0.576	0.566
SVM	58.2884	0.785	0.489	0.583	**0.602**

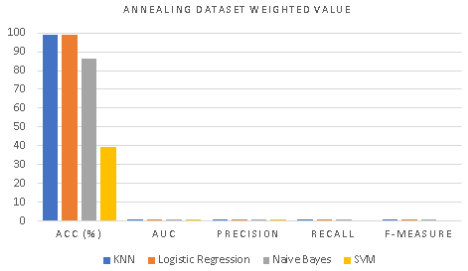

Fig.2. The chart is showing the effects of the Annealing dataset

In Fig. 2-6, indicates the enhanced classification and performance evaluation based on the datasets provided in the following mentioned charts. The LR, Annealing dataset has higher accuracy followed by KNN, NB, and SVM, in Fig. 2. Moreover, in Fig. 3, KNN, Breast Cancer

dataset, provides better outcomes after LR, NB, and SVM. Likewise, in Fig. 4, NB efficiency, Hepatitis dataset, yields efficient outputs as compared to LR, KNN, and SVM sequentially. Whereas, SVM, vertebral dataset, has higher accuracy in contrast to LR, KNN, and NB in Fig.5. Lastly, in Fig. 6, the LR, Yeast dataset, has outperformed than SVM, NB, and KNN.

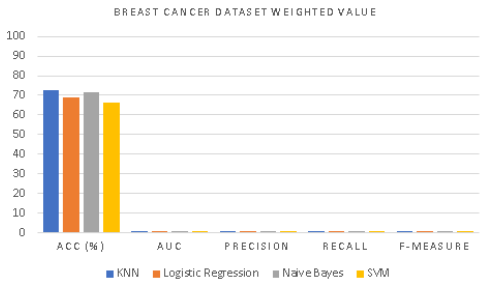

Fig.3. The chart is showing the effects of the Breast Cancer dataset.

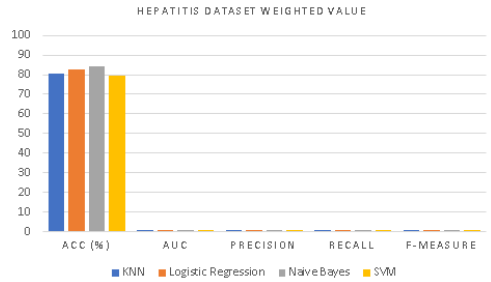

Fig.4. The chart is showing the effects of the Hepatitis dataset.

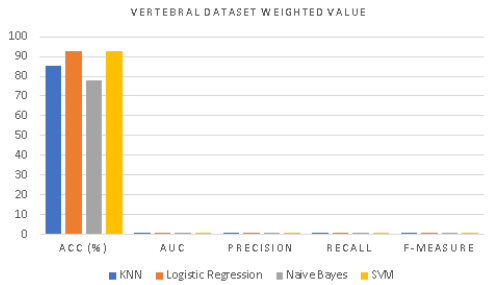

Fig.5. The chart is showing the effects of the Vertebral dataset.

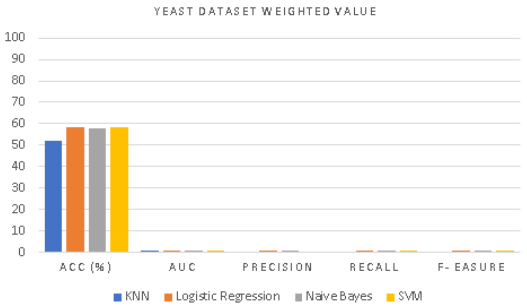

Fig.6. The chart is showing the effects of the Yeast dataset.

5. Conclusions And Future Work

Based on the experimental and numerical results, the main findings of this research work can be summarized as follows:

In this paper, we have examined the implementation of four ML algorithms which are named as k-nearest neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM) to classify multiple datasets. The efficiency of algorithms is further demonstrated in terms of precision, recall/sensitivity, accuracy, and F - score. Whereas many ML algorithms are unable to provide satisfactory results as they are dependent on the datasets. The sensitivity of the same algorithm can be severely affected by analyzed varying sizes of training and test sets.

Generally, LR has more successive consequences than KNN; whereas, in most datasets, the NB delivers more effective outputs than SVM. There is no winner outright in terms of the performance outcomes; it depends on the characteristics of the datasets, the simulation, and the circumstances.

In the future, we plan to reform our study of classification models by introducing the Intelligent ML algorithms, which are more useful to an extensive collection of real-life datasets.

REFERENCES

[1] A. A. ABRO, E. TAŞCI, and A. UGUR, "A Stacking-based Ensemble Learning Method for Outlier Detection," Balk. J. Electr. Comput. Eng., vol. 8, no. 2, pp. 181–185, 2020.

[2] A. A. Abro, M. Alci, and F. Hassan, "Theoretical Approach of Predictive Analytics on Big Data with Scope of Machine Learning."

[3] A. A. Aburomman, M. Bin, and I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," vol. 38, pp. 360–372, 2016.

[4] S. S. Tabrizi and N. Cavus, "A hybrid KNN-SVM model for Iranian license plate recognition," Procedia - Procedia Comput. Sci., vol. 102, no. August, pp. 588–594, 2016.

[5] E. Ayyad, "ScienceDirect Procedia Procedia Computer Procedia Procedia Computer Science Computer Computer Science Science Science Performance evaluation of intrusion detection based on machine Performance evaluation of using intrusion detection on machine learning Ap," Procedia Comput. Sci., vol. 127, pp. 1–6, 2018.

[6] H. Mohammad and D. Science, "Performance Evaluation of Machine Learning Algorithms in Ecological Dataset OF MACHINE LEARNING ALGORITHMS," no. March, 2019.

[7] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," Phys. A Stat. Mech. its Appl., vol. 517, pp. 29–35, 2019.

[8] R. Kalyan, D. Mishra, and A. Kumar, "A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices," Appl. Soft Comput. J., vol. 35, pp. 670–680, 2015.

[9] S. S. Panesar, R. N. D. Souza, F. Yeh, and J. C. Fernandez-miranda, "Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database," World Neurosurg. X, vol. 2, p. 100012, 2019.

[10] F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework," Inf. Process. Manag., no. February, p. 102221, 2020.

[11] J. Feng, Y. Wang, J. Peng, M. Sun, J. Zeng, and H. Jiang, "Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries," J. Crit. Care, vol. 54, pp. 110–116, 2019.

[12] C. Technology and J. Vol, "Design And Develop Network Storage Virtualization By Using GNS3," vol. 13, pp. 20–24.

[13] S. Sabahi and M. Mellat, "International Journal of Production Economics The impact of entrepreneurship orientation on project performance : A machine learning approach," Int. J. Prod. Econ., no. April 2019, p. 107621, 2020.

[14] I. Kayijuka, "computational efficiency of singular and oscillatory integrals with algebraic singularities," no. November, 2018.

[15] A. Ashari, "Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," vol. 4, no. 11, pp. 33–39, 2013.

[16] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naı ̈ ve Bayesian, Decision Tree and KNN classification techniques," J. King Saud Univ. - Comput. Inf. Sci., vol. 28, no. 3, pp. 330–344, 2016.

[17] J. zhou Feng, Y. Wang, J. Peng, M. wei Sun, J. Zeng, and H. Jiang, "Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries," J. Crit. Care, vol. 54, pp. 110–116, 2019.

[18] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," Knowledge-Based Syst., vol. 192, p. 105361, 2020.

[19] F. Karimi, S. Sultana, A. S. Babakan, and S. Suthaharan, "Computers , Environment and Urban Systems An enhanced support vector machine model for urban expansion prediction," Comput. Environ. Urban Syst., vol. 75, no. January, pp. 61–75, 2019.

[20] T. Classification and B. K. Singh, "Investigations on Impact of Feature Normalization Techniques on Investigations on Impact of Feature Normalization Techniques on Classifier ' s Performance in Breast Tumor Classification," no. April 2015, pp. 10–15, 2017.

[21] UCI Machine Learning Repository, 2018, https://archive.ics.uci.edu/ml/index.php.

[22] T. A. Engel, A. S. Charão, M. Kirsch-Pinheiro and L. A. Steffenel, "Performance improvement of data mining in weka through

GPU acceleration," Procedia Comput. Sci., vol. 32, pp. 93–100, 2014..

[23] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[24] L. A. Bull, K. Worden, R. Fuentes, G. Manson, E. J. Cross, and N. Dervilis, "Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data," J. Sound Vib., vol. 453, pp. 126–150, 2019.