DOI: 10.15849/IJASCA.231130.15

Int. J. Advance Soft Compu. Appl, Vol. 15, No. 3, November 2023
Print ISSN: 2710-1274, Online ISSN: 2074-8523

SFM: A Sequential Fitting Method to Address the Overfitting Problem of Logistic Regression

Abdallah Bashir Musa

Department of Basic Science, Deanship of Preparatory Year and Supporting Studies, Imam Abdulrahman Bin Faisal University, P.O. Box1982, Dammam 34212, SAUDI ARABIA
abhamad@iau.edu.sa

Abstract

Applying logistic regression (LR) when the number of features exceeds the number of instances is one of the great challenges that attracted the researchers' attention. This paper proposes a sequential fitting method (SFM) to address the overfitting problem of logistic regression. The proposed method is based on the fact that logistic regression features should be uncorrelated, and the number of features must be relatively less than the number of instances. Typically, only a few of these features are significant in building the model. In addition, the paper provides a comprehensive comparison of logistic regression (LR), naïve Bayes (NB), and random Forest (RF) in terms of the number of training data, number of features, and balanced or unbalanced data sets. Machine learning metrics such as accuracy, specificity, sensitivity, and area under the Roc curve are used to evaluate the algorithm's performance. The results of the three classifiers on these metrics have been validated and compared using some statistical analysis including the area under the ROC curve, and Wilcoxon signed-rank tests. The study concluded that the proposed method (SFM) is successful in applying logistic regression with overfitting data sets, and the proposed method can compete with Naïve Bayes and Random Forest.

Keywords: *Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Sequential Fitting Method (SFM), Machine Learning.*

1 Introduction

Logistic Regression (LR) [1-3] is the most famous statistical technique for performing classification tasks, that has been extensively utilized in many disciplines, including machine learning [4], and medical studies [5-7]. It has some advantages, such as generating a predicted probability vector for class labels. In addition, the LR model can easily be interpreted. LR is mostly used as a binary classifier. Also, there is multinomial logistic regression which is used for multi-class classification. However, the implementation of logistic regression for classifying a dataset with too many numbers of features with relatively few instances remains a challenge for the researchers. From a statistical point of view, the application of logistic regression requires the number of instances of the data set to be relatively larger than the number of the features [8]; this restrictive the application of

Received 6 July 2022; Accepted 11 September 2023

logistic regression when the number of instances is relatively less than the number of features.

This paper proposes a new method called sequential fitting method (SFM)for implementing the logistic regression using the fact that the logistic regression features are supposed to be uncorrelated and usually only a few features are significant in building the model. The new method makes a big contribution to solving the overfitting problem of logistic regression.

The classification task is the backbone of the machine-learning community. The classification performance of machine learning algorithms on datasets with varying data characteristics is not well covered as various algorithms are continuously being developed. Most of the studies compare the performance of algorithms on a particular application with a single dataset.

To assess the performance of the proposed methods a comprehensive comparison among logistic regression, naïve Bases, and random forest has been conducted, variety of machine learning data sets with different characteristics in terms of training dataset size and number and types of features are been used.

Naïve Bayes [9-10] classifier is a well-known classification algorithm that uses the Bayesian rule with the assumption that the attributes are conditionally independent for a given class. Although this assumption is not satisfied in practice, the naïve Bayes classifier often yields a competitive classification accuracy. In addition to this, its simplest and computational efficiency with many other desirable features makes the naïve Bayes classifier have many uses in many applications. The naïve Bayes classifier has various applications in several areas including medical data classification [11-14], text and signal classification [15-17], and fraud detection [18,19].

Random Forest (RF) developed by Leo Breiman in 2001 [20,21] is a supervised learning algorithm that combines the bagging method with the randomization of features by producing a random subset of features, which yields a low correlation among decision trees. Random Forest has recently become a popular classification algorithm used in disciplines including medical sciences [22,23] and machine learning [24,25].

There are some papers proposed a comparative study between these methods, but most of these studies are built in a particular application with a single data set and few machine learning metrics used in assessing the method's performance; for instance; Bansal [26] evaluated the performance of naïve Bayes, Random Forest, and other machine learning algorithms in detecting dementia using single data sets with 461 instances. The study showed that random Forest performed better than naïve Bayes for the original data while naïve Bayes is better than random Forest for the reduction dimension data in terms of accuracy, Devika [27] examined the performance of Naïve Bayes, K-Nearest Neighbour (KNN) and Random Forest for detecting kidney disease based on its accuracy and preciseness. Random Forest classifier was better than naïve Bayes.

Fayaz Itoo [28] provided a comparison and analysis of logistic regression, naïve Bayes, and KNN machine learning algorithms for credit card fraud detection using three different proportions of datasets with a resampling technique, logistic regression was found to be better than naïve Bayes in term of accuracy. Anwar [29] compares the performance of logistic regression and naïve Bayes in the classification of Authorship of Tweets using a single data set with 46895 instances, logistic regression was better than naïve Bayes in terms of accuracy. Trigila [30] compares and assesses the performance of Logistic

Regression and Random Forest for shallow landslide susceptibility, random Forest was performed better than logistic regression. Wonsuk Yoo[31] compared four classification methods including logistic regression and random Forest for identifying important genes, the random Forest performed better than the logistic regression and Kanish Shah [32] examined Logistic Regression, Random Forest and KNN for the Text Classification designed a BBC news text classification system using several machine learning metrics, the study has shown that logistic regression' performance was better in term of accuracy and it is performed well in terms of all measures. While most of the studies compare the performance of these algorithms on a particular application with a single dataset, the performance of classifier algorithms on datasets with various data characteristics is not well studied and remains a hot area of research.

The performance of the proposed method A Sequential Fitting Method (SFM) is been compared with the performance of naïve Bayes and random forest using various machine learning datasets with different characteristics in terms of the training dataset, size number, and types of features. Also, the study provides a comprehensive statistical analysis of compassion.

The performance of machine learning algorithms also relies on the characteristics of the dataset. Different evaluation metrics assess different characteristics of machine learning algorithms, and it is plausible for a learning algorithm to exhibit proficiency in one metric while displaying suboptimal performance in others. To tackle this issue, our study incorporates a diverse range of criteria to evaluate the classification performance of the learning methods. Hence, this study differs from the others in that it includes a comprehensive statistical analysis of the performances of the algorithms.

Also, a statistical analysis is performed to compare the proposed methods with the performance of naïve Bayes and random Forest.

The paper provides an overview of the three classifiers in Section 2. The proposed method is presented in Section 3. Section 4 explains the dataset, performance metrics, and experimental setup. Section 5 presents and discusses the performance results, statistical analysis, and ROC analysis. The conclusions are drawn in Section 6.

2 The classification methods

2. 1 Logistic regression

Logistic Regression (LR) [1-3] is a popular statistical technique for classifying binary data. It is based on minimizing an average logistic loss function based on the conditional probabilities of training data whose variables are the parameters of a classifier. Suppo that we have a set of training data set of size m, $\{x_i, y_i\}_{i=1}^{m}$, where $x_i \in R^n$ and $y_i \in \{-1, +1\}$ denote the i-th sample and the associated class label respectively. Assume that the training samples are independent. According to the logistic model, the vector of the conditional probabilities corresponding to these samples are stated as:

$$\text{pr}\ (\alpha, \beta)_i = \text{p}\ (y_i|x_i) = \frac{\exp y_i(\beta^T x_i + \alpha_i)}{1+\exp y_i(\beta^T x_i + \alpha_i)} i = 1, \ldots\ldots, m \qquad (1)$$

The likelihood function associated with the samples is $\prod_{i=1}^{m} \text{pr}\ (\alpha, \beta)_i$, and the log likelihood function is

$$\sum_{i=1}^{m} log\ \text{p r}\ (\alpha, \beta)_i = -\sum_{i=1}^{m} f(\beta^T a_i + \alpha y_i) \qquad (2)$$

where $a_i = x_i y_i \in R^n$ and f is the logistic loss function given by

$$f(z) = log(1 + exp(-z)) \quad (3)$$

Combining (2) and (3) yields the following equation:

$$\sum_{i=1}^{m} log\ \mathrm{p\,r}\ (\alpha, \beta)_i = -\sum_{i=1}^{m} log(1 + exp - (\beta^T a_i + \alpha y_i))) \quad (4)$$

The logistic loss is the negative of the log likelihood function. Hence the average logistic loss is found as

$$l_{avg}(\alpha, \beta) = \frac{1}{m}\sum_{i=1}^{m} log(1 + exp(-(\beta^T a_i + \alpha y_i))) \quad (5)$$

The maximum likelihood estimation method is used to determine the parameters β and α from the training data set, through solving the following convex optimization problem:

$$\min\ l_{avg}(\alpha, \beta) \quad (6)$$

Finally, the solutions of (7) are used to form the following logistic regression classifier.

$$\varphi(x) = sgn(\beta^T x + \alpha) \quad (7)$$

where

$$sgn(z) = \begin{cases} +1 & z > 0 \\ -1 & z \leq 0 \end{cases}$$

2.2. Naïve Bayes

Naïve Bayes [9-10] is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. It is a form of Bayesian Network Classifier based on Bayesian rule data. For the given training data set $x \in R^n$ where x_i is the i^{th} component in x, with an associated target variable $y \in \{-1, +1\}$ as the paper is dealing with only binary classification.

According to Bayes rule, the probability of x_i given that it is classified as $y \in \{-1, +1\}$ is

$$p(x_i|y) = \frac{p((x_i).p(y|x_i)}{p(y)} \quad (8)$$

By assuming all the components are independent, the probability of an instance $E_k = (x_1$, x_2 , …, x_n) where k = 1,2, …, m, given the value of the class as $y \in \{-1, +1\}$ is

$$p((x_1\ , x_2\ , \ldots, x_n|y) = \prod_{i=1}^{n} \frac{p((x_i).p((y|x_i)}{p(y)} \quad (9)$$

The instance E is classified as $y = +1$ if and only if

$$n_b(E_k) = \frac{p(y=+1|x_1\ , x_2\ , \ldots, x_n)}{p(y=-1|x_1\ , x_2\ , \ldots, x_n)} \geq 1 \quad (10)$$

where $n_b(E_k)$ is called Bayesian classifier.

2.3. Random Forest

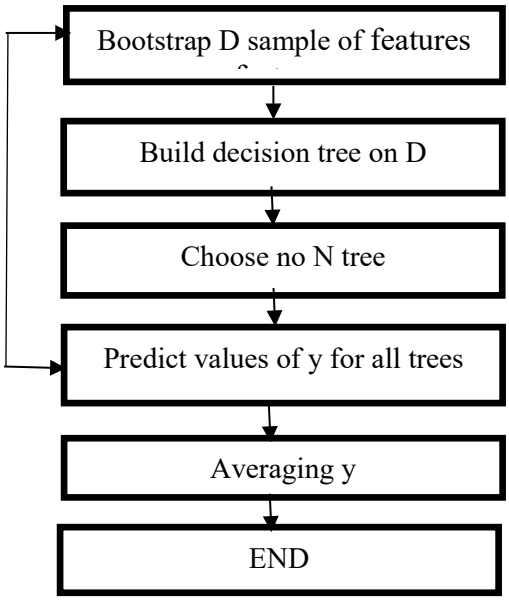
A random Forest [20,21,32] is a machine learning technique that is used to solve regression and classification problems. Random Forest shows an excellent performance with both small and high dimensional data sets with less training time. It is a supervised machine learning method that is constructed from a decision tree. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random Forest algorithm consists of many decision trees. Every decision tree is associated with a set of bootstrap samples of the features that are generated from the original data set. The sample set of the features is the same size as the original data. Each node of the decision tree is divided according to the entropy associated with a subset of the features. Then an approach called the bagging (bootstrap aggregated) technique is used to select the best trees with a voting scheme. Bagging repeatedly selects a random sample of features with replacement from the training set and fits trees to these samples. Hence, the 'Forest' generated by the random Forest algorithm is trained through bagging or bootstrap aggregating. Applying of bagging results in high improvement in performance and gives substantial gains in accuracy than using individual classifier [33]
The steps for construction random Forest are as follows: -

1. Pic a bootstrapped sample of features from the original training data set.
2. Build a decision tree using this bootstrapped sample.
3. Select the number of N tree of tree you want to build.
4. Predict the value of the class label y of each decision tree and assign the new data points to the category that wins the majority votes.
5. Repeat the previous steps.
6. Aggregate all predicted y values.

These steps are shown in figure 1

Figure 1: Random Forest' flow chart



3 The Proposed Method

In this paper, we propose a method based on two facts: first, logistic regression requires the features to be uncorrected, and second, usually, only a few features will be relevant when building the model. In the first step, we will take the number of features that equal the number of instances divided by 10 according to the rule. We will apply logistic regression to these features. We will keep the significant features for the second step.

Following that, we will add the next features, ensuring that the number of the added features plus the significant features does not exceed the number of instances divided by 10. This procedure will be repeated until all features are selected for building the model. The classification accuracy of the model with be measure. This proposed method will be valid and effective when the number of significant features does not exceed the number of instances divided by 10.

The procedure of the method is: -

1- Count the number of the instances, say N.

2- Choose number (N/10) starting from the first column.

3- Built the LR on these features & test the significance using $\alpha = 0.5$.

4- Count the number of the significant feature/s (S) out of these (N/10).

5- Keep the significant features (S) and choose number of [(N/10) -S] features to perform step 3.

6- Repeat 5 until all features have been selected.

7- If the number S in step 4 is greater than (N/10), stop the procedure.

8- Build the final model and compute it accuracy.

4 Materials and methods

4.1 The data sets

The data sets used in this study are composed of 16 data sets with binary class attributes, from the UCI repository (UCI Machine Learning Repository: Data Sets) and Kaggle at Kaggle Datasets. These data sets are of different types and sizes, five of them are almost balanced and the remaining eleven are unbalanced. six of them are considered as overfitted data sets as the number of the features is relatively larger than the number of the instances. Table 1 gives a numerical summary of the data sets.

Table 1: summary of the data sets

Data set	Data size	Number of variables	Type of features	Data type
Breast cancer	683	9	Numerical	unbalanced
Diabetes	768	8	Numerical	unbalanced
Liver disorder	345	6	Categorical+ Real	balanced
Spam	4601	57	Numerical	unbalanced
Ionosphere	351	34	Numerical	unbalanced
Heart	270	13	Categorical+	balanced
German	1000	20	Numerical	unbalanced
Surgical-deepnet	9567	25	Numerical	Unbalanced
QSAR biodegradation	1054	41	Numerical	Unbalanced
Sports articles	1000	59	Numerical	Unbalanced
The overfitting data sets				
Sonar	208	60	Numerical	balanced

Digital Colposcopies	287	69	Numerical	unbalanced
Cervical Cancer	72	19	Numerical	Unbalanced
Students' performance	145	31	Numerical	Unbalanced
Lymphography	142	18	Categorical	Balanced
Musk(version1)	476	168	Numerical	Balanced

4.2 The performance measures

The various performance metrics assess different tradeoffs in the predictions made by the algorithm, and it is possible for the learning algorithm to outshine in one metric while being suboptimal in others [33]. As a result, a range of machine learning metrics have been employed to evaluate the method's performance. The commonly used machine learning performance metrics include accuracy, specificity, sensitivity, precision, F-score, and Kappa.

Furthermore, the ROC (receiver operating characteristic) visually compares the algorithm's performance across all possible probability thresholds. The ROC curve plots the observed sensitivity against (1-specificity) for every potential classification threshold. It also measures the algorithms' ability to distinguish instances of different classes. The strength of the ROC curve lies in its depiction of the classification model's performance as a curve rather than a single point. Therefore, we utilized the area under the ROC curve to compare the class probability estimators of the algorithms.

4.3 Experimental Setup

The experiment performed a binary classification of the data sets using logistic regression, naïve Bayes, and random Forest. The methods are applied in accordance with standard approaches. As the standard method for logistic regression, when the number of features is not much larger than the number of features, when the thumb rule is met. Most statisticians and researchers use the method in its standard form due to its simplicity and lack of expertise. For the overfitting data sets, we apply the proposed method for logistic regression because the standard form of logistic regression cannot be applied to such data sets.

The original data set is divided into two sets for training and testing: 70 % of the data is used for training the model while 30% is allocated for testing the model. The statistical metrics are computed from the testing data. The ROC curve is also calculated from test data. For the overfitting data, the whole data is used for building the model and the statistical metrics are computed based on the training data. The proposed method has been applied using the above-mentioned procedure in SPSS 22. The results of the method on the overfitting data set are compared with the results of naïve Bayes and random Forest which are applied via R version 4.1.2 as well as the results of the logistic regression, naïve Bayes and random Forest on the data sets that are not suffering from overfitting. The ROC curve analysis and the nonparametric Wilcoxon signed ranked test is used to compare the performance of the algorithms.

5 Results, Analysis and Discussions

The results of logistic regression, naïve Bayes and random Forest have been carried out using R version 4.1.2 available at Download R-4.2.1 for Windows. The R-project for statistical computing. while the results of the logistic regression on the overfitting data sets have been carried out using SPSS 22.

5. 1 Performance by measures.

The results of naïve Bayes, logistic regression and random Forest on the data sets for each machine learning metric are shown in Table 2, Table 3, and Table 4 respectively.

Table 2 : The results of the performance measures for Naïve Bayes (NB)

Data set	Accuracy	Specificit	Sensitivity	Precision	F_score	Kappa	AUC
Breast cancer	0.971	0.962	0.976	0.976	0.976	0.938	0.995
Diabetes	0.770	0.577	0.868	0.852	0.860	0.464	0.822
Liver disorder	0.631	0.687	0.528	0.475	0.500	0.210	0.629
Spam	0.713	0.948	0.555	0.464	0.506	0.457	0.886
Ionosphere	0.914	0.940	0.865	0.889	0.877	0.810	0.923
Heart	0.815	0.790	0.837	0.818	0.828	0.628	0.847
German	0.723	0.408	0.830	0.805	0.818	0.246	0.740
Surgical-deepnet	0.795	0.756	0.820	0.8395	0.8298	0.537	0.880
QSAR biodegradation	0.718	0.954	0.594	0.961	0.734	0.468	0.876
Sports articles	0.813	0.646	0.914	0.816	0.861	0.580	0.849
The overfitting data sets							
Sonar	0.731	0.595	0.887	0.857	0.702	0.471	0.842
Digital Colposcopies	0.784	0.875	0.507	0.571	0.537	0.340	0.762
Cervical Cancer	0.931	0.857	0.961	0.942	0.952	0.830	0.993
Students' performance	0.772	0.772	0.772	0.840	0.805	0.533	0.827
Lymphography	**0.845**	**0.901**	**0.771**	**0.839**	**0.869**	**0.570**	**0.893**
Musk(version1)	**0.805**	**0.845**	**0.773**	**0.867**	**0.817**	**0.609**	**0.901**

Table 3: The results of the performance measures for logistic regression (LR)

Data set	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**F_score**	**Kappa**	**AUC**
Breast cancer	0.966	0.937	0.984	0.961	0.972	0.927	0.995
Diabetes	0.783	0.577	0.888	0.804	0.844	0.480	0.819
Liver disorder	0.660	0.731	0.528	0.516	0.521	0.258	0.740
Spam	0.918	0.883	0.842	0.923	0.932	0.829	0.963
Ionosphere	0.837	0.896	0.730	0.794	0.761	0.637	0.830
Heart	0.790	0.737	0.837	0.783	0.810	0.577	0.834
Surgical-deepnet	0.871	0.906	0.817	0.8849	0.8954	0.728	0.936
German	0.740	0.421	0.848	0.812	0.828	0.282	0.728
QSAR biodegradation	0.877	0.789	0.923	0.893	0.907	0.723	0.936
Sports articles	0.828	0.709	0.895	0.814	0.867	0.618	0.867
The overfitting data sets							
Sonar	0.813	0.790	0.833	0.812	0.822	0.624	0.876
Digital Colposcopies	0.812	0.829	0.707	0.409	0.518	0.413	0.763
Cervical Cancer	0.944	0.905	0.975	0.961	0.961	0.866	0.975
Students' performance	0.772	0.761	0.778	0.875	0.824	0.506	0.777
Lymphography	0.838	0.828	0.845	0.877	0.861	0.667	0.937
Musk(version_1)	0.859	0.840	0.874	0.877	0.876	0.713	0.938

Table 4: The results of the performance measures for Random Forest (RF)

Data set	Accuracy	Specificit	Sensitivity	Precisio	F_score	Kappa	AUC
Breast cancer	0.980	0.975	0.984	0.984	0.984	0.959	0.994
Diabetes	0.765	0.603	0.849	0.806	0.827	0.463	0.833
Liver disorder	0.728	0.761	0.667	0.600	0.632	0.417	0.772
Spam	0.942	0.890	0.977	0.930	0.953	0.878	0.982
Ionosphere	0.923	0.955	0.865	0.914	0.889	0.830	0.967
Heart	0.778	0.684	0.861	0.755	0.804	0.550	0.860
Surgical-deepnet	0.943	0.888	0.979	0.9314	0.9218	0.879	0.982
German	0.750	0.421	0.861	0.814	0.837	0.300	0.734
QSAR biodegradation	0.880	0.807	0.918	0.901	0.909	0.732	0.933
Sports articles	0.817	0.718	0.874	0.843	0.858	0.600	0.889
The overfitting data sets							
Sonar	0.947	0.897	0.991	0.917	0.952	0.893	0.997
Digital Colposcopies	0.892	0.977	0.634	0.900	0.744	0.678	0.974
Cervical Cancer	0.944	0.857	0.980	0.943	0.961	0.862	0.994
Students' performance	0.786	0.597	0.909	0.781	0.841	0.530	0.841
Lymphography	0.873	0.803	0.926	0.862	0.893	0.738	0.930
Musk(version_1)	0.996	0.995	0.996	0.996	0.996	0.994	0.999

5.2. The statistical analysis

The nonparametric Wilcoxon signed _ranked test is performed for comparing the performance of proposed methods (SFM) logistic regression, the naïve bayes, and random Forest under all the metrics using 0.05 as a level of significance. The test is carried out using SPSS 22. The results of the p_values of comparing NB and LR, NB and RF, LR and RF are shown in Table 5, Table 6, and Table 7 respectively.

Table 5 : The results of the Wilcoxon test of naïve Bayes (NB) and logistic regression (LR)

	Accuracy	**Specificity**	**Sensitivity**	**Precision**	**F_score**	**Kappa**	**AUC**
P_value	**0.088**	**0.514**	**1.09**	**0.623**	**0.171**	**0.103**	**0.379**

Table 6. The results of Wilcoxon test for naïve Bayes (NB) and Random Forest (RF)

	Accuracy	**Specificity**	**Sensitivity**	**Precision**	**F_score**	**Kappa**	**AUC**
P_value	**0.011**	**0.82 ***	**0.003**	**0.156***	**0.007**	**0.013**	**0.006**

Table 7. The results of the Wilcoxon test for logistic regression (LR) and Random Forest (RF)

	Accuracy	**Specificity**	**Sensitivity**	**Precision**	**F_score**	**Kappa**	**AUC**
P_value	**0.009**	**0.32***	**0.018**	**0.056***	**0.006**	**0.007**	**0.002**

5.3. The ROC curve analysis

The ROC curves for the naïve Bayes, logistic regression, and random Forest models are presented in the figures below. Figure 2 and Figure 3 display the ROC curves for the liver and spam datasets, respectively, serving as examples of balanced and unbalanced non-overfitting datasets. On the other hand, Figure 4 and Figure 5 exhibit the ROC curves for

the solar and Lymphography datasets, respectively, illustrating balanced and unbalanced overfitting datasets. The relationship between the RUC values of the naïve Bayes, logistic regression, and random Forest models for these datasets is depicted in Figure 6, with the relationship for the overfitting data shown on the right side of the figure. Figure 6 represents the AUC values for each dataset, following the same order as presented in Table 1 for naïve Bayes, logistic regression, and random Forest. The figure demonstrates that random Forest consistently outperforms naïve Bayes and logistic regression for almost all the datasets. Furthermore, for the overfitting data, the proposed methods (SFM) logistic regression and naïve Bayes are very similar. To assess the statistically significant differences of AUC, the Wilcoxon signed ranked test is conducted. The p-value for comparing naïve Bayes and logistic regression is 0.379, indicating that there is no significant difference between these two algorithms. However, the p-values for comparing random Forest with naïve Bayes and logistic regression are 0.006 and 0.002, respectively, suggesting a high significant difference between random Forest and both naïve Bayes and logistic regression. Therefore, random Forest outperforms both naïve Bayes and logistic regression in terms of the area under the ROC curves (AUC).

Figure 2: ROC curve for liver.

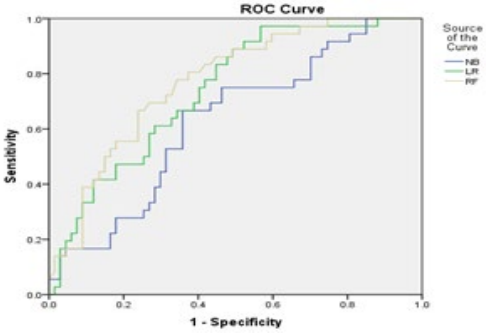


Figure 3: ROC curve for spam

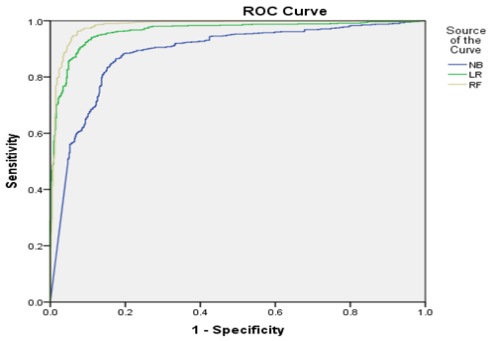
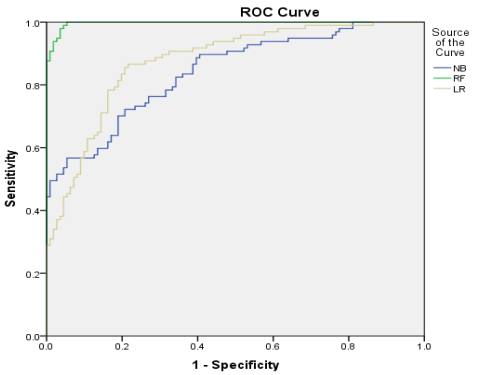


Figure 4: ROC curve for solar



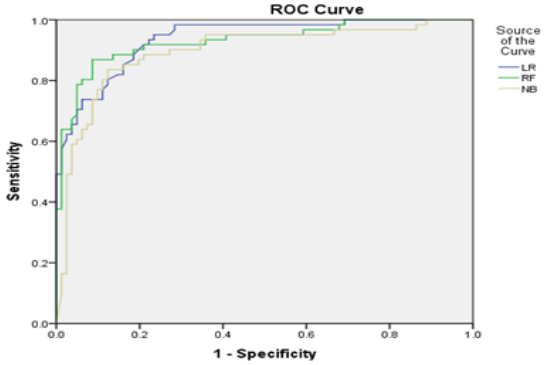
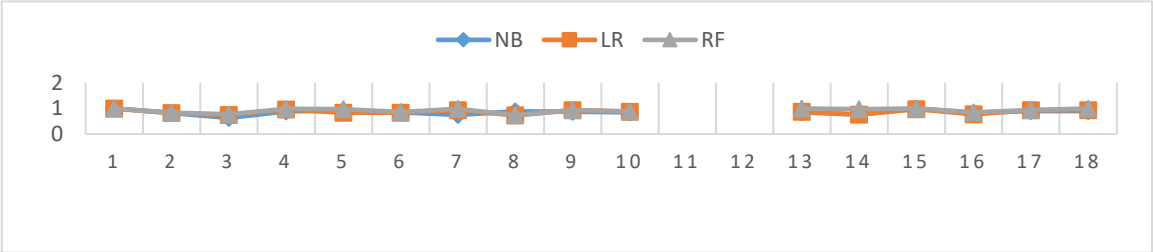


Figure 5 : ROC curve for Lymphography

Figure 6: The relationship of the RUC of the NB, LR and RF for the data sets



.

5.4 Discussion

In this study, various datasets with distinct characteristics were utilized, employing diverse machine learning measures to facilitate comprehension and the development of a comprehensive outcome. The overall findings of the experiments have demonstrated that the three algorithms exhibit satisfactory performance. According to Table_5, the p-value obtained from the Wilcoxon signed-ranked test, which compares the performance of naïve Bayes and logistic regression across all metrics, exceeds the level of significance. The minimum p-value observed for the accuracy metric was 0.088, indicating that there is no statistically significant difference in performance between naïve Bayes and logistic regression. On the other hand, Table_6 reveals that the p-values obtained from the Wilcoxon signed-ranked test, comparing the performance of naïve Bayes and random Forest across all metrics, are below the significance level of 0.05, except for the Specificity and Precision metrics, where the p-values are 0.82 and 0.156, respectively. These results suggest that there is a statistically significant difference in performance between naïve Bayes and random Forest; random Forest is performing better than naïve Bayes. The p-values obtained for the Specificity and Precision metrics can be recognized to the bias of naïve Bayes towards the bigger class. According to the results obtained from Table_7, the p-values of the Wilcoxon signed-rank test for all metrics, except for Specificity and Precision, are below the significance level of 0.05. The p-values for Specificity and Precision are 0.32 and 0.056, respectively. These findings indicate a significant difference in performance between logistic regression and random Forest, with random Forest outperforming logistic regression. The p-values for Specificity and Precision suggest that logistic regression is biased towards the bigger class, like what was observed for the naïve Bayes method. Since the p-values for these metrics are lower than the p-values obtained when comparing naïve Bayes and random Forest, it can be concluded that logistic regression is less biased towards the bigger class compared to naïve Bayes. Random Forest performs better than both logistic regression and naïve Bayes, with a higher performance observed in 13 out of the 16 data sets, accounting for approximately 80% of the data across all measures. Although the statistical test does not show a significant difference between logistic regression and naïve Bayes, logistic regression performs better in 12 out of the 16 data sets, representing approximately 75% of the data in terms of accuracy.

The study also, demonstrates that the performance of naïve Bayes is influenced by the number of training data. Increasing the size of the data leads to a decrease in the performance of naïve Bayes, as observed in the spam and surgical data sets, which are the largest data sets. In contrast, naïve Bayes performs better for the heart data set, which is the smallest data set. Furthermore, the results indicate that both logistic regression and naïve Bayes exhibit bias in classification with respect to the bigger class. Random Forest, on the other hand, performs better for unbalanced data sets. Additionally, when the data

set is unbalanced and relatively small compared to the number of features, logistic regression does not perform well. This is evident in the ionosphere data set, which has the largest number of features. For overfitting data sets, logistic regression is not applicable. In such cases, the proposed method (SFM) is applied to these data sets, and the results are compared to those obtained by applying naïve Bayes and random Forest. The majority of the values of the area under ROC curves (AUC) for the three algorithms are closely aligned, as depicted in the right part of Figure 6. The only exception is the Lymphography data set, which is categorical in nature. This is because logistic regression becomes slightly confusing when applied to categorical data. Consequently, naïve Bayes and random Forest do not outperform the proposed method (SFM).

6 Conclusion

The paper presents a comprehensive comparative study that examines logistic regression, naïve Bayes, and random Forest as statistical algorithms for classifying and modeling binary data. The comparison is conducted using datasets of varying sizes and types. Various machine learning measures are employed to ensure a fair and clear evaluation, as different datasets may perform better for certain metrics while not performing well for others. The study applies the three methods using the standard approach commonly used in research. Additionally, the study proposes a new method to address the overfitting problem that renders logistic regression unsuitable for such data types. The results indicate that there is no statistical difference between naïve Bayes and logistic regression. However, there is a significant difference between logistic regression and random Forest, as well as between naïve Bayes and random Forest. Random Forest outperforms logistic regression and naïve Bayes, particularly for unbalanced datasets. The results also reveal that both logistic regression and naïve Bayes exhibit bias towards the larger class. Furthermore, naïve Bayes does not perform well when the training data size is large, while logistic regression struggles when the training data is small relative to the number of features.
Moreover, the study concludes that the proposed method (SFM) successfully applies logistic regression to overfitting datasets and can compete with naïve Bayes and random Forest.

References

1. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley series in probability and statistics, Wiley, Inc, New York
2. Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression (Vol. 398). John Wiley & Sons.
3. Menard, S., 2002. Applied logistic regression analysis (Vol. 106). Sage.
4. Thabtah, F., Abdelhamid, N. and Peebles, D., 2019. A machine learning autism classification based on logistic regression analysis. Health information science and systems, 7(1), pp.1-11.

5. Srivastava, N., 2005, November. A logistic regression model for predicting the occurrence of intense geomagnetic storms. In Annales geophysicae (Vol. 23, No. 9, pp. 2969-2974). Copernicus GmbH.
6. Jiang, X., El-Kareh, R. and Ohno-Machado, L., 2011. Improving predictions in imbalanced data using pairwise expanded logistic regression. In AMIA annual symposium proceedings (Vol. 2011, p. 625). American Medical Informatics Association.
7. Reed, P. and Wu, Y., 2013. Logistic regression for risk factor modelling in stuttering research. Journal of fluency disorders, 38(2), pp.88-101.
8. Musa, A.B., 2014. A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression. International Journal of Machine Learning and Cybernetics, 5(6), pp.861-873.
9. Leung, K.M., 2007. Naïve bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, pp.123-156.
10. Murphy, K.P., 2006. Naïve Bayes classifiers. University of British Columbia, 18(60), pp.1-8.
11. Kazmierska, J. and Malicki, J., 2008. Application of the Naïve Bayesian Classifier to optimize treatment decisions. Radiotherapy and Oncology, 86(2), pp.211-216.
12. Langarizadeh, M. and Moghbeli, F., 2016. Applying naïve bayesian networks to disease prediction: a systematic review. Acta Informatica Medica, 24(5), p.364.
13. Miranda, E., Irwansyah, E., Amelga, A.Y., Maribondang, M.M. and Salim, M., 2016. Detection of cardiovascular disease risk's level for adults using naïve Bayes classifier. Healthcare informatics research, 22(3), pp.196-205.
14. Berrar, D., 2018. Bayes' theorem and naïve Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.
15. Celin, S. and Vasanth, K., 2018. ECG signal classification using various machine learning techniques. Journal of medical systems, 42(12), pp.1-11.
16. Shimodaira, H., 2014. Text classification using naïve bayes. Learning and Data Note, 7, pp.1-9.
17. Shimodaira, H., 2014. Text classification using naïve bayes. Learning and Data Note, 7, pp.1-9.
18. Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.
19. Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D. and Sharma, M., 2018. Credit card fraud detection using Naïve Bayes model based and KNN classifier. International Journal of Advance Research, Ideas and Innovations in Technoloy, 4(3).
20. Breiman, L., 2001. Random Forests. Machine learning, 45(1), pp.5-32.
21. Biau, G. and Scornet, E., 2016. A random Forest guided tour. Test, 25(2), pp.197-227.

22. Konukoglu, E. and Glocker, B., 2020. Random Forests in medical image computing. In Handbook of Medical Image Computing and Computer Assisted Intervention (pp. 457-480). Academic Press.
23. Alickovic, E. and Subasi, A., 2016. Medical decision support system for diagnosis of heart arrhythmia using DWT and random Forests classifier. Journal of medical systems, 40(4), p.108.
24. Zimmerman, N., Presto, A.A., Kumar, S.P., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L. and Subramanian, R., 2018. A machine learning calibration model using random Forests to improve sensor performance for lower-cost air quality monitoring. Atmospheric Measurement Techniques, 11(1), pp.291-313.
25. Hartshorn, S., 2016. Machine learning with random Forests and decision trees. Kindle edition
26. Bansal, D., Chhikara, R., Khanna, K. and Gupta, P., 2018. Comparative analysis of various machine learning algorithms for detecting dementia. Procedia computer science, 132, pp.1497-1502.
27. Devika, R., Avilala, S.V. and Subramaniyaswamy, V., 2019, March. Comparative study of classifier for chronic kidney disease prediction using naïve bayes, KNN and random Forest. In 2019 3rd International conference on computing methodologies and communication (ICCMC) (pp. 679-684). IEEE.
28. Itoo, F. and Singh, S., 2021. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), pp.1503-1511.
29. Aborisade, O. and Anwar, M., 2018, July. Classification for authorship of tweets by comparing logistic regression and naïve Bayes classifiers. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 269-276). IEEE.
30. Trigila, A., Iadanza, C., Esposito, C. and Scarascia-Mugnozza, G., 2015. Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology, 249, pp.119-136.
31. Yoo, W., Ference, B.A., Cote, M.L. and Schwartz, A., 2012. A comparison of logistic regression, logic regression, classification tree, and random Forests to identify effective gene-gene and gene-environmental interactions. International journal of applied science and technology, 2(7), p.268
32. Shah, K., Patel, H., Sanghvi, D. and Shah, M., 2020. A comparative analysis of logistic regression, random Forest and KNN models for the text classification. Augmented Human Research, 5(1), pp.1-16.
33. Musa, A.B., 2013. Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics, 4(1), pp.13-24.

Notes on contributor.

Abdallah B. Musa: Received the B.Sc. degree in Computer/Statistics from University of Gezira, Sudan, in 2000, The M.Sc. degree with distinction in Applied Statistics from University of Gezira in 2006. Doctor of Engineering in Management sciences and Engineering from Hebei University, Baoding, China in 2014. Area of research applied Statistics and Machine Learning. He has published several research articles in journals of mathematics statistics and machine learning.