Performance Evaluation of Statistical and Machine Learning Models

Mervat M. Ramadan
Professor of Statistics, Statistics, Mathematics, and Insurance Faculty of Business, Benha University

Dina S. Eltelbany
Lecturer of Statistics, Statistics, Mathematics, and Insurance Faculty of Business, Benha University

Ayman S. Hegazy
Teaching Assistant, Statistics, Mathematics, and Insurance, Faculty of Business, Benha University

Performance Evaluation of Statistical and Machine Learning Models

Mervat M. Ramadan [a] , Ayman S. Hegazy [a] , Dina S. Eltelbany [a]

[a] *Department of Statistics, Mathematics, and Insurance,*
Faculty of Business, Benha University

Abstract:

Machine learning algorithms have gained popularity in recent years in many fields due to their promising results in predictive performance of classification problems. The application of machine-learning algorithms has also been highly simplified in the last years due to their well-documented integration in commonly used statistical programming languages (such as R or Python). Machine learning is a subsection of Artificial Intelligence (AI), it is one of the most promising tools in classification and it a model that aims to discover the unknown function, dependence, or structure between input and output variables. This study proposes statistical and machine learning models to diagnose anemia disease. Some machine learning techniques have been used in this work to avoid overfitting, pre-process the data and adjust the outliers to give better results. Three classifiers, including Logistic Regression, k-Nearest Neighbor and Decision Tree are implemented in this work. The performance of the models is evaluated based confusion matrix, recall, precision, f1-score, accuracy, Matthews correlation coefficient and ROC curve to compute area under the curve (AUC). The results show the logistic regression has the highest accuracy of 99.57%, with recall values of 99.41%, precision values of 99.61, f1-score of 99.51% and Matthews correlation coefficient values of 99.13%. Decision tree has the second highest accuracy of 98.64%, with recall values of 99.02%, precision values of 97.87, f1-score of 98.44%, and Matthews correlation coefficient values of 97.23%.

Keywords: Machine Learning, ROC curve, Confusion Matrix

1 Introduction

The concept of machine learning is concluded in the use of algorithms based on statistical models to train computers to extract knowledge from large sets of data. The data set represents historical data from reality, for example medical observation history for real patients. A machine learning model will have its own unique algorithm that would guide it to use this data set and discover different patterns to either classify the data or predict future data. The algorithm that a model uses to process or analyze the data is based on a set mathematical equation involving linear algebra, logarithmic, arithmetic, statistics, probability, and calculus. There are many machine learning models with multiple variations of algorithms.

Machine learning plays a vital role in predicting classification problems effectively and efficiently with less cost. Different machine learning algorithms such as logistic regression, support vector classification, and random forest have been used for many years in medical fields and have successfully analyzed, characterized, and correctly predicted the results. This study shows the optimal result for classification prediction using statistical and machine learning models.

Our contribution to this proposed work:

- ❒ Some statistical techniques have been used to remove the correlation or dependencies between features in a dataset to improve the performance of the classifiers.
- ❒ Three classifiers and different machine learning techniques are performed in this research work to get the best result.

- ❒ Among the three classifiers, Logistic Regression has the highest accuracy of 99.57%, with recall values of 99.41%, precision values of 99.61, f1-score of 99.51% and matthews correlation coefficient values of 99.13%.

2 Related Works

There are various studies on generating predictions and evaluating the performance of machine learning models in classification in several applications from researchers of diversified areas of economic, statistics, engineering, and science. Here is an overview of some of these studies reviewing statistical and machine learning models related work:

Sabbeh, S. F. (2018) used machine-learning models to analyze customers' personal and behavioral data to give organization a competitive advantage by increasing customer retention rate. Those models predicted customers who are expected to churn and reasons for churn. This paper tried to compare and analyze the performance of different machine-learning techniques that are used for churn prediction problem included Discriminant Analysis, Decision Trees, K-Nearest Neighbors, Support Vector Machines, Logistic Regression, ensemble–based learning techniques, Naïve Bayesian, and Multi-layer perceptron. Results show that both random forest and ADA boost outperform all other techniques with almost the same accuracy 96%. Both Multi-layer perceptron and Support vector machine can be recommended as well with 94% accuracy. Decision tree achieved 90%, naïve Bayesian 88% and finally logistic regression and Linear Discriminant Analysis (LDA) with accuracy 86.7%.

Sanni and Guruprasad (2021) analyzed machine learning algorithms based on the percentage of various performance metrics (such as, Accuracy, Precision and Recall) to predict heart failure. The researchers used different supervised machine learning algorithms which include Decision Tree, Logistic Regression, KNN and Random Forest. The researchers obtained the performance metrics, Accuracy, Precision and Recall, obtained by the four machine learning algorithms to compare between. The results observed that the highest accuracy is obtained by decision tree, the highest precision is obtained by Logistic Regression and the highest recall is obtained by KNN. The random forest algorithm gives promising results across all the performance metrics.

Bekele, W. T. (2022) focused predicting the low birth weight (LBW) and used the Ethiopia Demographic and Health Survey 2016 to develop predictive LBW models. The research used Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (K-NN), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB) to compare and find the best classifier for predictive classification. Before applying the predictive models, data preprocessing was carried out, including data cleaning. The results shown that the RF was the best classifier, predicting LBW with 91.60 percent accuracy, 91.60 percent Recall, 96.80 percent ROC-AUC, 91.60 percent F1 Score, 1.05 percent Hamming loss, and 81.86 percent Jaccard score, according to the research, RF predicted the occurrence of LBW correctly and more effectively than other classifiers.

Sharma and Mishra (2022) focused on prediction of various diseases in early stages (as breast cancer diagnosis) by using machine learning and deep learning evolution. The researchers chose well established ML algorithms such as K-NN, LR, DT, ANN, SVM, RF, Adaboost etc. based on various feature selection techniques such as CFS, SFS and Information Gain. Features selected using CFS achieved higher accuracy as compared to other techniques. Finally, they built a voting classifier by combining results of the best three models SVM, LR and ANN to classify the new test samples. Results shown that voting classifier can predicted results with 99.41% accuracy.

3 Machine Learning Models

After the pre-processing of the data, three different machine learning classifiers have been used namely Logistic Regression, k-Nearest Neighbor, and Decision Tree for anemia prediction. After that, the optimal classifier is evaluated with others models according to performance metrics.

3.1. Logistic Regression [LR]

Logistic regression, even if called regression, is a popular classification classifier that is based on the probability of a sample belonging to a class. It uses categorical and continuous variables to predict a categorical outcome. This algorithm is closely related to linear regression, but it uses the sigmoid function to assign probabilities to discrete outcomes, which transforms numerical values into an expression of probability between 0 and 1. The probability is calculated by the equation (3).

$$P(Y) = \frac{1}{1 + e^{-(\alpha + \beta x)}} \quad (1)$$

Where α is the intercept, β is the beta coefficient, and x is the value of the predictor.

3.2. K-Nearest Neighbor [K-NN]

K-Nearest Neighbors is a simple supervised learning classifier that predicts the label of data points by looking at what is the majority in its closest neighbors. It tries to predict the correct class for the new data by calculating the distance between the new data and all the training points. Then select the K number of points which is closest to the new data. To show how does K-NN classifier works. As seen from Figure 1, if we set K= 3, then we predict that the new point (test data) belongs to class B, and if we continue with K= 6, then we predict that this point belongs to class A.

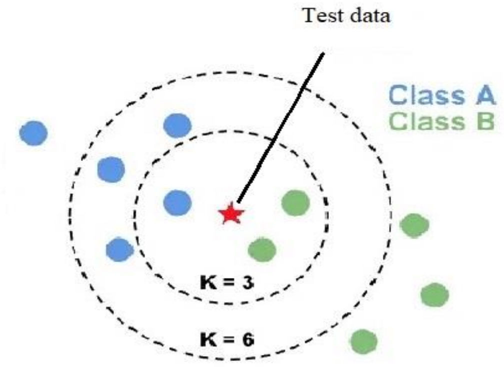

Figure 1. K-nearest neighbor algorithm illustration.

Equation (2), (3), and (4) represent the formula of Euclidean, Manhattan and Minkowski distance functions, respectively which them calculate the distance between the train data and test data points.

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{2}$$

$$Manhattan = \sum_{i=1}^{k}\left|x_i - y_i\right| \tag{3}$$

$$Minkowski = \left(\sum_{i=1}^{k} \left(|x_i - y_i| \right)^p \right)^{1/p} \qquad (4)$$

Manhattan function has been used in this work which achieved the highest accuracy is 97.44% compared to other distance equations.

3.3. Decision Tree [DT]

The Decision Tree is a non-parametric supervised learning classifier, that constructs a tree-like structure consisting of branch nodes and leaf nodes. It is used for both classification and regression problems by splitting the dataset at each branch node using a set of criteria (impurity measures). The structure of a binary Decision Tree starts from the first decision node, known as the root node. And it contains the entire dataset, which is divided into two or more sub-trees/branches (the splitting is calculated according to the impurity measures). The decision nodes represent the dataset’s features, branches denote the decision rules, and each leaf node represents the classification outcome. Figure 2 shows the structure of a binary decision tree.

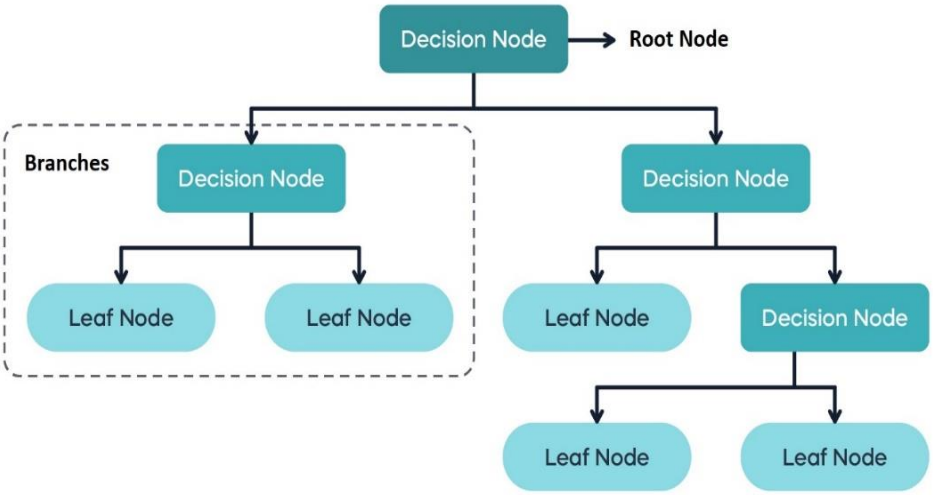

Figure 2. The structure of a binary decision tree.

For a binary classification for a node *j*, representing a region D_j with observations N_j and the proportion of class *c* observations in this node is calculated by the equation (5).

$$P_{jc} = \frac{1}{N_j} \sum_{y \in D_j} I(y = c) \qquad (5)$$

Where *I* represent the impurity measure which is used. The entropy impurity measure is used in this work.

4 Results and Discussion

The proposed work has been tested and trained in 25% and 75% of data respectively. Different machine learning classifiers have been evaluated such as LR, K-NN and DT. Among these three classifiers, the best machine learning model is discovered using performance metrics including confusion matrix, recall, precision, f1-score, accuracy, matthews correlation coefficient (MCC) and ROC curve to compute the area under the curve (AUC). This section covers the performance evaluation metrics for each classifier. Next, each model classifier has been evaluated by using the performance metrics which are shown in table 1.

Table 1. Performance metrics are used in this work.

Performance Metrics	Mathematical Formula
Accuracy [*A*]	$A = \frac{TP + TN}{TP + FP + FN + TN}$
Precision [*P*]	$P = \frac{TP}{TP + FP}$
Recall [*R*]	$R = \frac{TP}{TP + FN}$
F1-score [*F*]	$F = \frac{2 \times R \times P}{R + P}$
Matthews Correlation Coefficient [*MCC*]	$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

The performance of each classifier has been evaluated by using the above performance evaluation metrics, and results are shown in Table 2. We note that LR has the highest accuracy of 99.57%, with 99.41% recall, precision values of 99.61%, f1-score values of 99.51%, MCC values of 99.13% and AUC values of 1%. DT achieves the second highest accuracy of 98.64%, with 99.02% recall, precision values of 97.87%, f1-score values of 98.44%, MCC values of 97.23% and AUC values of 0.99%. Finally, K-NN the third highest accuracy of 97.69%, with 97.84% recall, precision values of 96.89%, f1-score values of 97.37%, MCC values of 95.33% and AUC values of 0.98%.

Table 2. Performance evaluation metrics for each classifier.

Model	Accuracy	Recall	Precision	F1-score	MCC	AUC
LR	0.99574	0.99412	0.99607	0.99509	0.99133	1.00
K-NN	0.97698	0.97843	0.96893	0.97366	0.95326	0.98
DT	0.98636	0.99019	0.97868	0.98441	0.97234	0.99

5 Conclusion

Machine learning classifiers play a vital role in predicting the dataset class in the first stage. This paper includes predicting classification problems based on different machine learning classifiers. The proposed work mainly contains many stages including loading the dataset, data pre-processing, and performance evaluation of classifiers. The results show the best accuracy for logistic regression, decision tree**,** and k-nearest neighbor are 99.57% with 99.41% recall, 98.64% with 99.02% recall, and 97.69% with 97.84% recall, respectively. We are hopeful that this experiment can help treat this disease more effectively and another research study will be conducted using neural networks and advanced deep learning models to achieve accurate results in future.

References

[1] Bekele, W. T. (2022). Machine learning algorithms for predicting low birth weight in Ethiopia. *BMC Medical Informatics and Decision Making*, *22*(1), 1-16.

[2] Bisong, E., & Bisong, E. (2019). Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, 243-250.

[3] Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.

[4] Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*, 171-186.

[5] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310.

[6] Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, *13*(8), 603-605.

[7] Rokach, L., & Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, 165-192.

[8] Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of advanced computer Science and applications*, *9*(2).

[9] Sanni, R. R., & Guruprasad, H. S. (2021). Analysis of performance metrics of heart failured patients using python and machine learning algorithms. *Global transitions proceedings*, *2*(2), 233-237.

[10] Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, *14*(4), 1949-1960.

[11] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1255-1260). IEEE.