International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An effective approach for determining sample size that optimizes the performance of the classifier

Tsehay Admassu Assegie*,1

Submitted: 30/09/2021 **Accepted : 03/04/2022**

Abstract: The goal of machine learning is to create a model that performs well and gives accurate prediction outcomes in a particular set of classification tasks. To achieve higher performance, the machine-learning model has to be optimized. Literature shows that parameter and hyper-parameter tuning is most widely used for model optimization. In most classification tasks, the dataset is divided into training and testing sets with 70% and 30% for training and test respectively. However, the 70% training and 30% testing set division does not guarantee better predictive outcomes for all classifications. Thus, this study proposes a learning curve for analysis of the effect of data size on the performance of the classification model using a real-world heart disease dataset employing a random forest model. The experimental result shows that data sample size has a significant effect on the performance of the random forest model. A learning curve is the best approach for determining the sample size for classification tasks using a machine-learning model.

Keywords: *Learning curve, model optimization, random forest, effective sample size*

1. Introduction

In recent years, machine learning has become one of the most widely researched fields due to its wider application in object recognition, disease diagnosis, and automated recommendation systems [1, 2]. However, the use of machine learning models for medical diagnosis and object recognition is not without negative impacts. Especially, in medical diagnosis the effect of false-negative results in adverse effects for patients and prolonged complications in cases where the machine learning model miss predicts the class of a given input.

The performance of machine learning algorithms largely depends on many parameters such as sample size, number of features, and the complexity of a particular problem under study [3]. One of the causes of false-positive or wrong predictive outcomes is the quality and quantity of the data sample size used for training the predictive model [4]. Moreover, the training and test set is divided by reserving 70% of the dataset for training and 30% for testing generally. The 70% training and 30% training set paradigm for model development is not the perfect choice for all datasets and problems under study.

Many machine-learning techniques have been applied to heart disease dataset classification. For instance, multi-class support vector machine [5], decision tree [6] and logistic regression, k-nearest neighbor, Naïve Bayes and random forest [7] with acceptable classification outcome, random forest producing highest prediction outcome of 92.85%. However, the studies are focused on classification accuracy and the number of features rather than the sample size. In [8], the researchers examined the performance of three classification algorithms namely, random forest, support vector machine, and Naïve Bayes. They found that the Naïve Bayes model outperforms with the highest classification accuracy of 80.42%. In [9?] the researchers investigated the performance of x algorithms for varying sample sizes trained on 44 has a classification accuracy of 78% outperforming other classifiers. In this study, we proposed a learning curve for determining the sample size for classification problems for improved predictive outcomes. Moreover, this study investigates the answers to the following research questions:

1) What are the existing methods used for determining sample size for training machine-learning mode to enchase classification outcome?
2) What is the effect of varying sample sizes on the training score and cross-validated score of the classification model?
3) How do determine the data sample size that maximizes the predictive outcome of the classification model?
4) What is the relationship between sample size and classifier performance?

2. Literature review

In this section, we will focus on literature related to the application of machine learning techniques to heart disease prediction and diagnosis. There have been many types of life-threating diseases among which heart disease is the most common and life-threating disease [9]. Early diagnosis of heart disease is vital to saving human life from heart disease. However, the identification of heart disease at the early stage of its occurrence is more complicated and challenging. This is because heart disease requires experienced experts and more accurate automated methods for better precision during the identification of heart disease in the early stages of its occurrence [10].

[1] *Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara-40, Ethiopia*
ORCID ID: 0000-0003-1566-0901
** Corresponding Author Email: tsehayadmassu2006@gmail.com*

A literature survey [11], shows that there have been automated methods and machine learning models for heart disease prediction in the early stage. Moreover, with the rapid growth of artificial intelligence and digital technology for data storage, a large volume of heart disease-related data is stored using digital systems all over the world. Machine learning has been found to have wide application in the medical field for the diagnosis and prognosis of diseases. In [12], the researchers used a decision tree for classifying the heart disease dataset. The researchers evaluated the proposed model and the result shows that classification accuracy of 83% is achieved with the decision tree model.

From the above reviews, we observe that the performance of the machine learning algorithm has scope for improvement. Thus, we aimed to optimize the performance of the existing classification model by determining the optimal sample size using a learning curve as a method for determining a better number of training instances.

3. Research method and materials

This section discusses the dataset employed for testing the proposed technique and the algorithm used for developing the model that predicts kidney and heart disease patients. A learning curve is employed to determine the optimal sample size by relating the cross-validation test score and training score to varying sizes of the data sample. Random forest, Naïve Bayes, and support vector machines are used to develop the model. The proposed model is tested on two medical datasets namely, the kidney and heart disease datasets collected from the Kaggle data repository. The performance of the model is evaluated using accuracy as a measure of performance. Each algorithm is trained on different sample sizes and the performance is measured on different training instances.

3.1. Performance metrics

To evaluate the performance of the proposed model on varying sizes of heart disease dataset samples, accuracy is employed as a classification metric. The accuracy of the classification model is detrained by the formula defined in equation (1).

$$Accuracy = \left(\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} * 100\right) \qquad (1)$$

4. Experimental result and discussion

In this study, we have employed Python 3.4 with Jupyter Notebook as the experimental platform for training the classification model with five-fold cross-validation. The performance of the support vector machine, random forest, and Naïve Bayes is recorded and the result is compared. The experimental result for the three algorithms on different sample sizes is discussed in figure 1.

As shown in figure 1, the training and cross-validation score converges. The model performs well when more data samples are used for training the model. In addition, the training score is much greater than the cross-validation score revealing that the model requires more training samples to generalize more effectively. The learning curve shown in figure x demonstrates the relationship between training score and cross-validation score for varying data samples. Hence, the grid search reveals training and cross-validated test scores, which is important to tune a model to find a better balance between error due to bias and variance. The learning curve shown in figure 1 shows that there is more variability around the cross-validation score revealing that the model suffers from error due to variance. The score values for various training instances are demonstrated in Table 1.

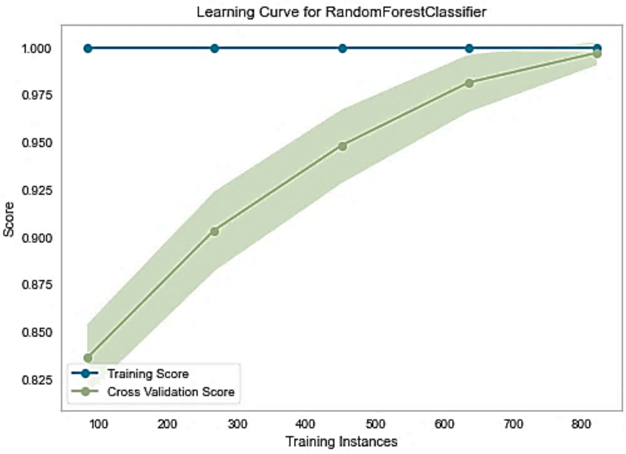

Fig. 1. Performance of random forest on heart disease dataset.

Table 1. Performance of random forest model on different training instances

Number of raining instances	Accuracy in %
100	84%
250	89%
450	96%
650	98%
800	99.02%

As demonstrated in table 1, the performance of the random forest model varies between 84% and 99.02% for varying sizes of training instances between 100 and 800. Classification accuracy for the model has shown a significant variation with lower accuracy of 84% using training instances of size 100 and an increase of 15.02% with training instances of 800. Thus, the accuracy of the random forest classifier largely depends on the training instance used for model development. Moreover, the learning curve is an important technique for determining the optimal number of training instances for developing a more accurate classification model.

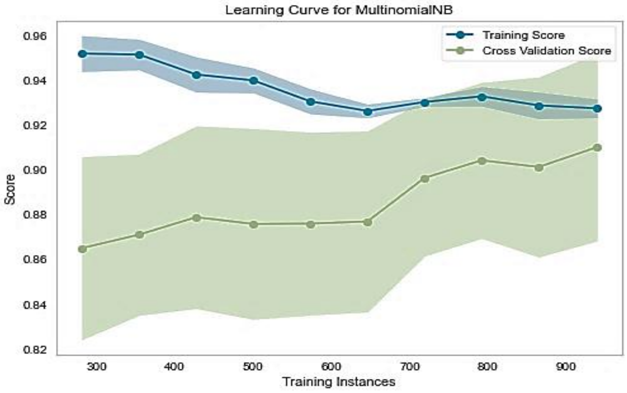

Fig. 2. Performance of Naïve Bayes on heart disease dataset.

Table 2. Performance of Naïve Bayes model on heart disease dataset

Number of raining instances	Accuracy in %
100	83%
250	85%
450	87%
650	88%
800	91%

As demonstrated in table 2, the Naïve Bayes model performs well when more training instances are used to train the model. However, the performance of the Naïve Bayes model is 91% when trained on 800 training instances. Thus, the random forest model performs well for heart disease diagnosis as compared to the Naïve Bayes model.

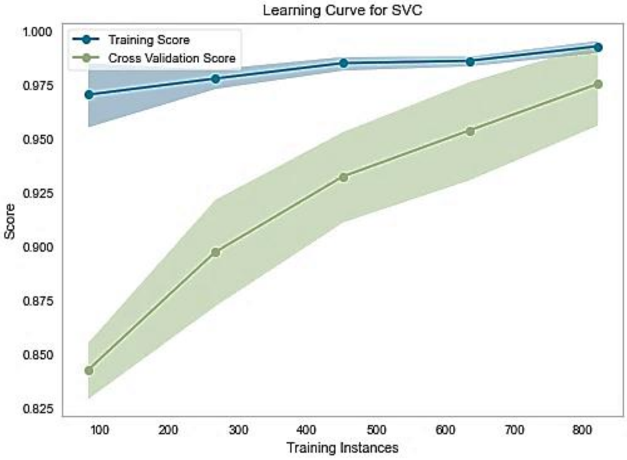

Fig. 3. Performance of support vector machine on heart disease dataset.

Table 3. Performance of support vector machine on heart disease dataset

Number of raining instances	Accuracy in %
100	82%
250	88%
450	94%
650	96%
800	97%

As demonstrated in table 3, the support vector machine performs better for higher training instances in a similar way to the random forest and Naïve Bayes model. The more the data points are in the training set, the better the model performance. In addition, the support vector machines performed better as compared to Naïve Bayes. However, the performance of the support vector machine is lower than the random forest model for heart disease prediction. Thus, the random forest model outperforms as compared to Naive Bayes and support vector machine although the Naive Bayes and support vector machine performed with promising or acceptable accuracy on heart disease prediction.

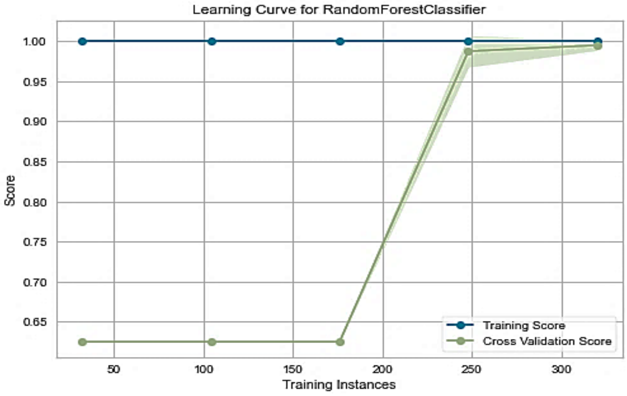

Fig. 4. Performance of random forest on kidney dataset.

Table 4. Performance of random forest on kidney dataset

Number of raining instances	Accuracy in %
25	55%
125	55%
175	55%
250	96%
350	97%

As demonstrated in table 4, the performance of random forest tends to increase for an increase in training instances. The highest accuracy archived with the random forest is 97% on the kidney dataset using the optimal training instances or sample size.

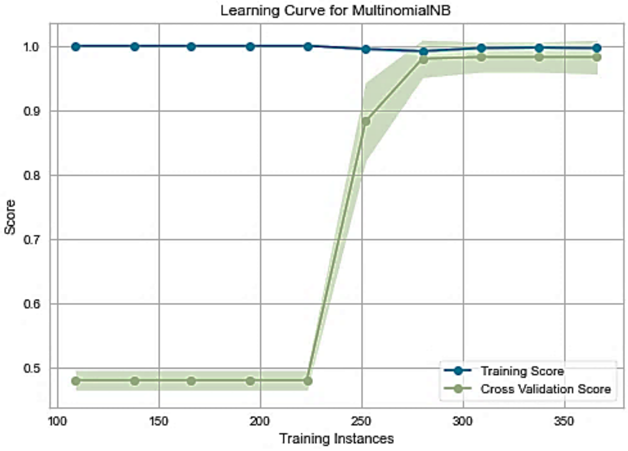

Fig. 5. Performance of Naïve Bayes on kidney dataset.

Table 5. Performance of Naïve Bayes on kidney dataset

Number of raining instances	Accuracy in %
25	40%
125	40%
175	40%
250	89%
350	99%

As demonstrated in table 5, the Naïve Bayes model tends to increase for an increase in training instances. The highest accuracy archived with the random forest is 99% on the kidney dataset using the optimal training instances or sample size. Initially, the performance of Naïve Bayes remained constant for training instances below 250.

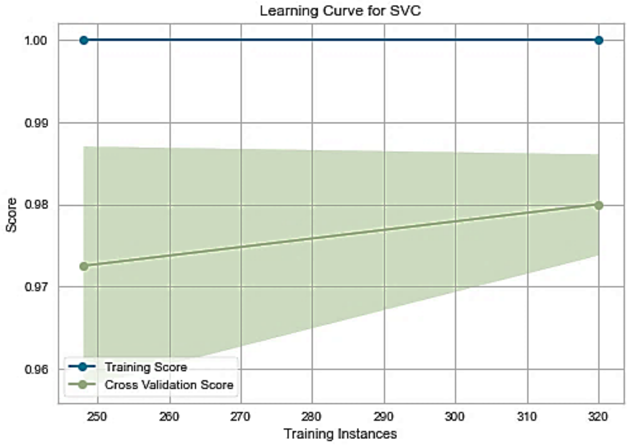

Fig. 6. Performance of support vector machine on kidney dataset.

Table 6. Performance of support vector machine on kidney dataset

Number of raining instances	Accuracy in %
25	90%
125	90%
175	90%
250	97.5%
350	98%

As demonstrated in table 6, the support vector machine model tends to increase for an increase in training instances. The highest accuracy archived with the random forest is 98% on the kidney dataset using the optimal training instances or sample size. Initially, the performance of Naïve Bayes remained constant for training instances below 250. We observe from table 4, table 5, and table 6 that, the Naïve Bayes model performs better on kidney disease diagnosis as compared to the random forest and support vector machine. The highest accuracy score by Naïve Bayes, random forest, and support vector machine is 99%, 98%, and 97% respectively. Hence, the Naïve Bayes model performs well on imbalanced classification tasks as compared to the support vector machine and random forest.

Table 7. Performance of classifiers on heart and kidney disease dataset

Classifier	Accuracy of heart disease dataset	Accuracy of kidney disease dataset
Random forest	99.02%	97%
Naïve Bayes	97%	99%
SVM	98%	98%

As shown in Table 7, the support vector machine performs with similar accuracy on the heart and kidney disease dataset. But the performance of the random forest model is less as compared to the heart disease dataset and the Naïve Bayes model performs better on the kidney disease dataset as compared to the heart disease dataset.

5. Conclusion

In this study, we have evaluated three machine learning models on the heart disease dataset. The performance of the classification model is compared on different sample sizes. The experimental result appears to prove that the sample size significantly affects the performance of the classification model. Based on the experiment conducted on varying sample sizes, we have concluded that the performance of the classification model largely depends on the sample size used for training a particular model. A learning curve is a good method for determining the training size that maximizes the performance of the classification model. The learning curve provides the relationship between the cross-validation test score and training score on varying sample sizes. Overall, with an optimal sample size of 78% for training and 22% of the dataset for testing the highest accuracy score was 98.02% using the random forest model on heart disease prediction and 99% using Naïve Bayes on kidney disease prediction.

For future work, we recommend researchers conduct an empirical study on other supervised learning methods such as decision trees, logistic regression, and K-nearest neighbor using different real-world medical datasets such as diabetes and liver disease datasets.

References

[1] T.A. Assegie, and P.S Nair, “Handwritten digits recognition with decision tree classification: a machine learning approach,” International Journal of Electrical and Computer Engineering., Vol. 9, No. 5, October 2019, Art. no. pp. 4446~4451.

[2] T.A Assegie, R.L Tulasi, N.K Kumar, “Breast cancer prediction model with decision tree and adaptive boosting,” IAES International Journal of Artificial Intelligence, Vol. 10, No. 1, March 2021, Art. no. pp. 184~190.

[3] X. Chen, “Coronary Artery Disease Detection by Machine Learning with Coronary Bifurcation Features, “Appl. Sci. **2020**, 10, 7656; doi: 10.3390/app10217656.

[4] P. Sujatha and K. Mahalakshmi, “Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart disease,” 2020 IEEE International Conference for Innovation in Technology, Nov 6-8, 2020.

[5] T.N Nguyen, “Multi-class Support Vector Machine Algorithm for Heart Disease Classification”, *International Conference on Green Technology and Sustainable Development, 2020.*

[6] M. Benllarch, “Improve Extremely Fast Decision Tree Performance through Training Dataset Size for Early Prediction of Heart Diseases”, IEEE, 2019.

[7] F. Tasnim, S.U Habiba, “A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection”, International Conference on Robotics, Electrical and Signal Processing Techniques, IEEE, 2021.

[8] Ni. Gupta, “Intelligent heart disease prediction in cloud environment through ensembling”, Expert Systems, https://doi.org/10.1111/exsy.12207, 2017.

[9] R.G. Franklin, “***Survey of Heart Disease Prediction and Identification using Machine Learning Approaches***”, Proceedings of the Third International Conference on Intelligent Sustainable Systems, 2020.

[10] R.L Fueroa, Q.Z Treitler, S. Kndula and L.H Ngo, “Predicting sample size required for classification performance”, BMC Medical Informatics and Decision Making 2012, 12:8.

[11] T.A Assegie, R. L Tulasi, N.K Kumar, “Breast cancer prediction model with decision tree and adaptive boosting”, IAES International Journal of Artificial Intelligence, Vol. 10, No. 1, March 2021, pp. 184-190.

[12] K.M Almustafa, “Prediction of heart disease and classifiers’ sensitivity analysis”, BMC, Bioinformatics, 2020, 21:278 https://doi.org/10.1186/s12859-020-03626-y.