Open Peer Review on Qeios

# Chronic disease treatment default prediction with random sampling optimization.

Michael Owusu-Adjei[1], James Ben Hayfron-Acquah[1], Twum Frimpong[1], Gaddafi Abdul-Salaam[1]

1 Kwame Nkrumah University of Science and Technology

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

## Abstract

A characteristic feature of real-world applications is the occurrence of dataset class imbalance in the output class distribution. Predictive modeling contributions from the minority or underrepresented class are overlooked by most learning algorithms. Addressing this challenge includes applying re-sampling techniques that eliminate class distribution imbalance for a more balanced output class distribution in the training examples. Random sampling techniques such as random over-sampling of the minority class duplicates the minority class examples to achieve a more balanced distribution or random under-sampling to delete training examples in the majority class for a balanced distribution to eliminate class imbalance in the dataset. The usefulness of these random sampling techniques has received attention in several research studies, particularly for binary classifications in two-class or multi-classification problems. This application, to many, is aimed at achieving equal class distribution meant to determine optimal model performance. This comparative assessment of random sampling optimization uses five classification-based algorithms, namely: extreme gradient boosting, gradient boosting, random forest, support vector machines and logistic regression, to evaluate predictive performance in random sampling on a real-world healthcare dataset of patients suffering from hypertension with comorbidities. The average prediction accuracy score (balanced accuracy) obtained shows statistically significant differences between scores obtained at the pre-sampling and post-sampling stages. The lowest score obtained with post-sampling was 85.55%, as against 54% in pre-sampling. However, high auc_roc score recorded in pre-sampling, over-sampling and under-sampling indicate a statistically insignificant impact of over-sampling and under-sampling use in this context. This confirms that the impact and effect of random sampling use in predictive modeling can better be explained in context.

**Michael Owusu-Adjei**[1*], **James Ben Hayfron-Acquah**[1], **Twum Frimpong**[1], **Gaddafi Abdul-Salaam**[1]

[1]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

*Corresponding author's email: mowusuadjei@st.knust.edu.gh

**Keywords:** Random sampling, techniques, prediction, accuracy, score, disease, treatment default

## 1.0 Introduction

Common among real-world applications is the occurrence of unequal class distribution in target labels. In supervised learning, variables are labelled for identification. Output class may consist of two or more classes for predictive classification modeling. Predictive modeling techniques that are capable of correctly identifying predominantly higher number of target classes define its performance potency. However, the challenge of predictive modeling techniques making biased predictions with skewed class distribution persists and continues to attract research interest. This phenomenon leads to a situation in which minority class contribution, effect and impact are ignored by learning algorithms. Meanwhile, in many real-world applications such as healthcare systems, fraud detections etc., minority class contributions are particularly of paramount interest. For example, the detection of fraudulent transactions in a banking system, spam email detection system etc., may reveal an exceedingly higher number of normal transactions than fraudulent ones, but for which the detection of fraudulent transactions is of utmost priority. Similarly, the prediction of infectious diseases[1][2][3] among patients could also show that the majority of patients who may not have the disease but knowledge of patient minority with the disease is of paramount importance to contain any potential spread or large scale outbreak and any devastating impact on public health. Additionally, risk assessment of default by patients to disease treatment[4][5] and other risk assessments for default may also encounter class distribution imbalance. Addressing dataset class imbalance to improve prediction accuracy scores continues to gain attention in many related research works. Solutions to dataset imbalance range from using sampling techniques that address class imbalance through over-sampling the minority class or under-sampling the majority class to other optimization techniques, such as applying class weight optimization to penalize the minority class for each prediction error made. A key concern for addressing class imbalance is why and for what purpose? This legitimate concern arises because a characteristic of real-world applications is the phenomenon of class imbalance; therefore, applied predictive modeling techniques should rather be evaluated to determine effective and efficient performance on such datasets with minority class involvement. This evaluation task should include impact assessment on prediction accuracy, model generalization and model behavior on training examples over a period. This key challenge has received little attention in related research works as a solution to the use of datasets with class imbalance.

### *1.1 Related Works*

In this section, a preview of related research works showcasing random sampling technique use is examined for emphasis. Using transaction details of customers, the application of random sampling techniques showed high over-sampling technique performance over under-sampling [6]. Applied over-sampling techniques on large datasets to address challenges associated with class imbalance drew conclusions about its efficiency against other sampling techniques [7]. Similarly, a comparative study of several re-sampling techniques to determine efficiency also concluded that over-sampling is an efficient sampling approach [8]. Improving prediction accuracy in instance method selection [9] with over-

sampling and under-sampling techniques showed accuracy improvements in minority class prediction. Understanding superior differences in performance of over and under-sampling techniques with an exploration of dataset inner structure [10] to explain the superior performance of over-sampling technique. A new approach [11] for determining the required over-sample size, which is less than the required sample for achieving class balance, has been evaluated to conclude that this approach improves classification performance when used with over-sampling. Comparative analysis of over-sampling and under-sampling influence on physical activities with ensemble machine learning techniques determines that under-sampling with refined features addresses class imbalance more effectively[12]. Another approach to effectively address class imbalance includes the use of synthetic minority over-sampling [13]. A combination of three sampling techniques – over-sampling, under-sampling and combi-sampling [14] – on healthcare datasets with artificial neural networks showed varying performance for different re-sampling techniques, including varying performance on different datasets by ANN.

## 2.0 Methods

To address the challenge of minority class use in predictive modeling, we adopt two approaches. One is to perform predictive modeling using healthcare context-based datasets with extreme class imbalance without sampling techniques to evaluate model performance on prediction accuracy score, model generalization and model behavior regarding training examples with five (5) classification models. The second approach is to use the same dataset with class imbalance with two sampling techniques (over-sampling the minority class and under-sampling the majority class for comparative evaluation analysis to determine the impact of addressing dataset class imbalance. For this purpose, we adopt a real-world healthcare dataset obtained with expressed permission referenced DCS/S.1/VOL.1 from a district hospital in Ghana (Kwahu Government Hospital) noted for its long-standing involvement in managing chronic diseases. Identifiable features, such as patient names, were blocked from the records for privacy protection. Records obtained showed biological data, clinical notes, patient visits and performance metrics for prescriber evaluation. Records with no relevancy to this research were excluded in the collection process, and some of these were body temperature readings, clinical notes on eye treatments, dental notes, obstetric notes etc. Gold standards used for feature labeling include patient attendance records and clinical note descriptions without identifiable patient features to address patient privacy concerns. We adopt five classification-based machine learning techniques: Gradient boosting, Extreme gradient boosting, Logistic regression, Support vector machines and Random forest classifiers for all predictive modeling.

### *2.1 Exploratory dataset analysis*

Dataset exploration for patterns and features is described in the figures below.

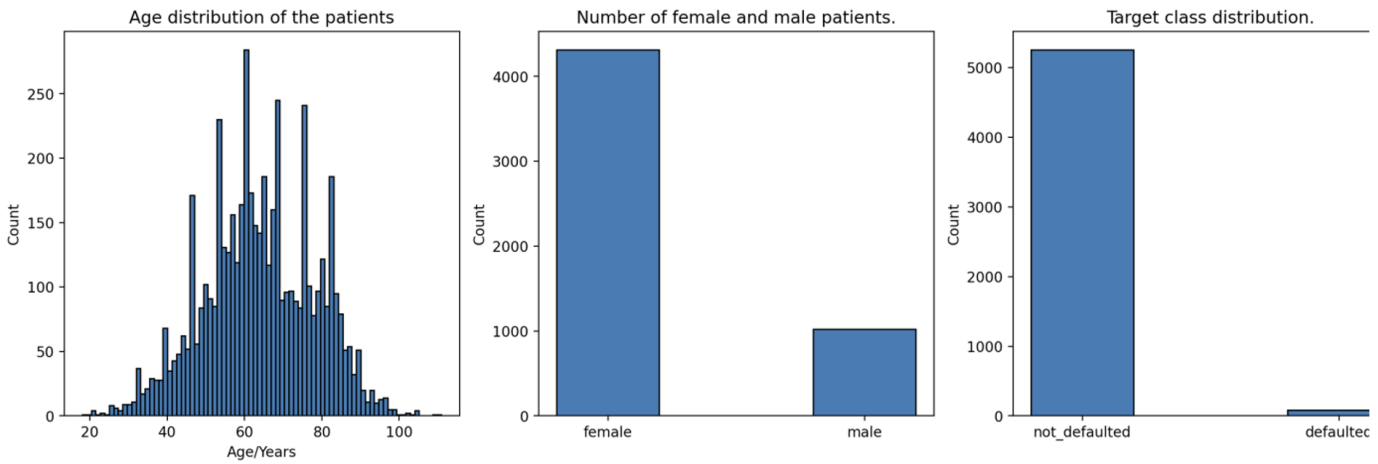



**Figure 2.1 Distribution of age, gender and target class**

The age distribution graph (first graph) of the sampled dataset from Figure 2.1 can be described as normal or with a Gaussian distribution. The gender distribution graph (second graph) shows the number of female and male patients. This graph describes a higher number of female patients than males in the sampled dataset. The last graph (third graph) shows the distribution of the target class (number of patients described as defaulters of chronic disease and those who are described as non-defaulting patients). The distribution of female numbers in the sampled population was 4,312, constituting 80.86%, and 1,021 males made up 19.14% of the entire sampled population, bringing the total sample population to 5,333.

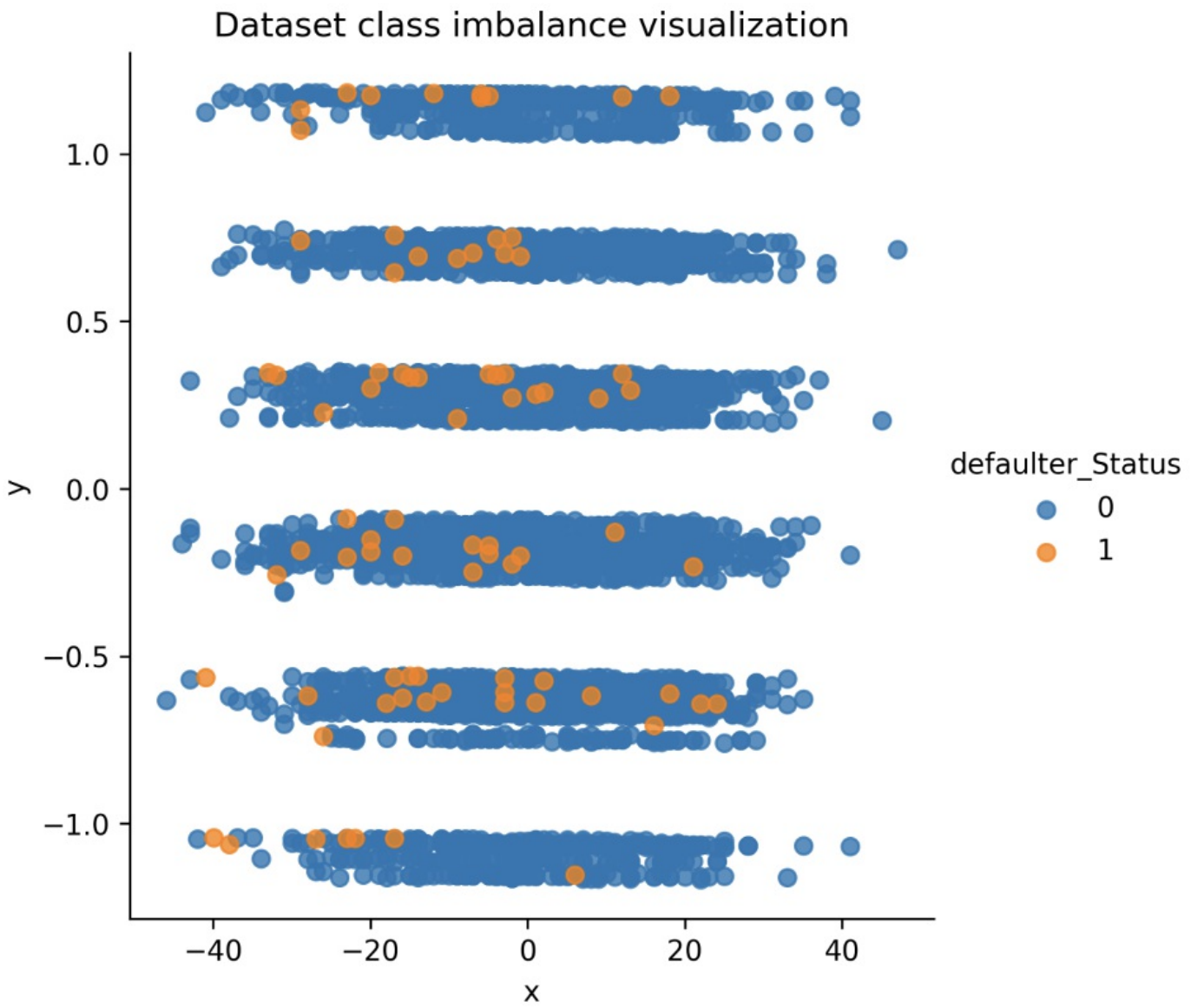



Figure 2.2 Class imbalance visualization

The dataset class output class imbalance visualization graph to demonstrate class distribution is shown in Figure 2.2 above.

### *2.2 Data preprocessing stages*

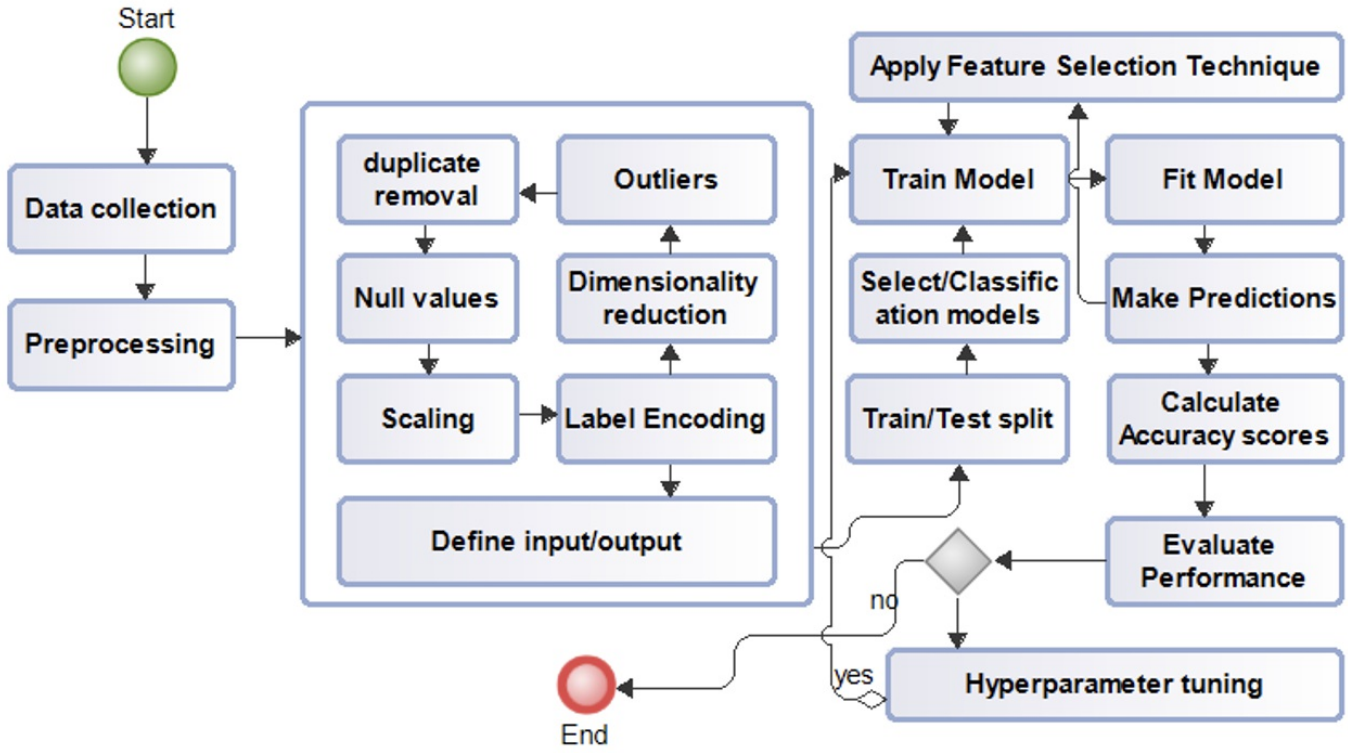



Figure 2.3 Flowchart of data preprocessing.

Figure 2.3 is a display of data preprocessing processes to improve dataset quality for model construction. Stages involving sub-processes are boxed.

### *2.3 Supervised learning types*

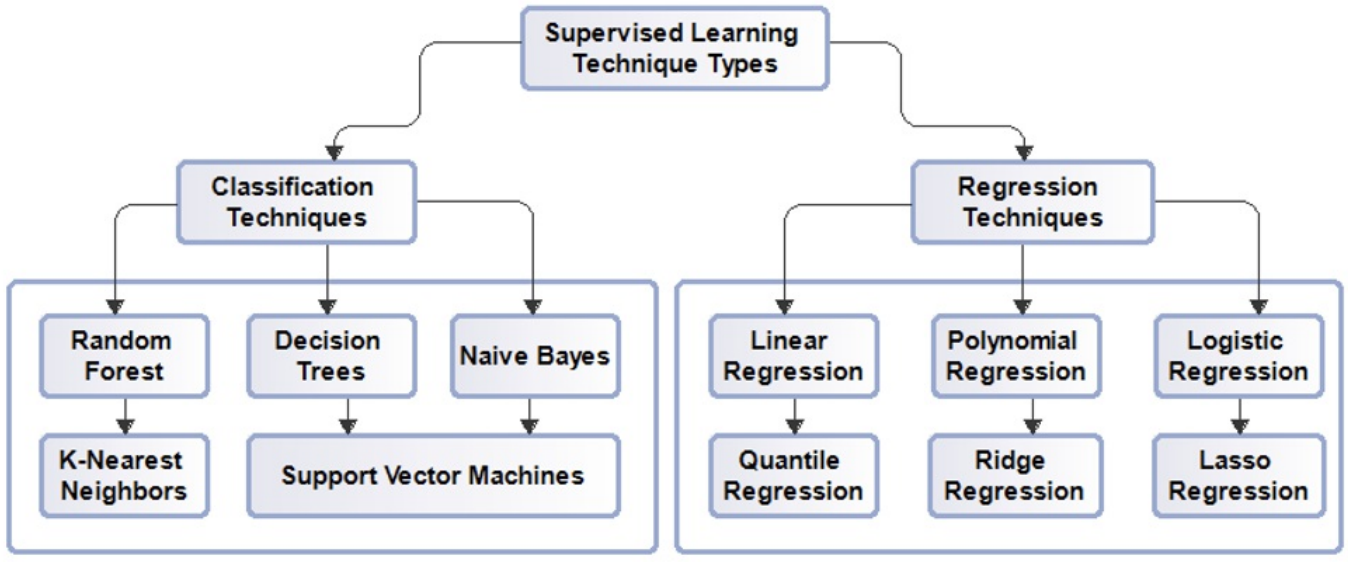



Figure 2.4 Model types

Figure 2.4 shows supervised model selection types for each problem category and lists learning application choices for each problem domain. Supervised learning has two main branches of use, classification and regression, and each choice has predefined model selections.

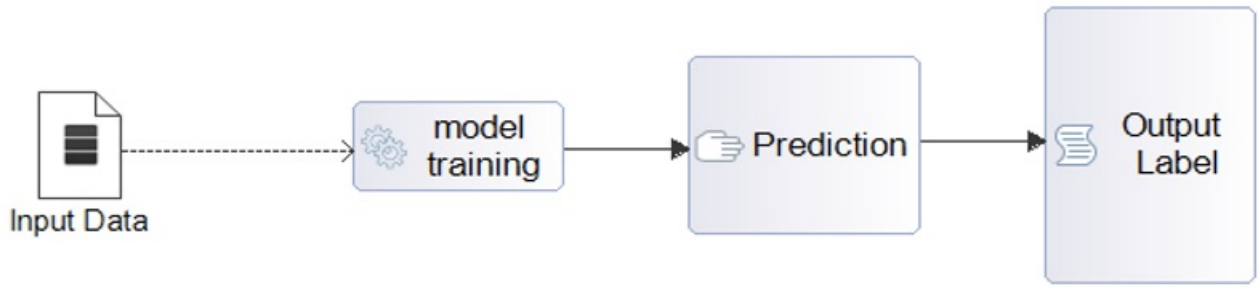



Figure 2.5 Simplified model building process

Labeled data, as shown in Figure 2.5 (processed input), is given to a machine learning (supervised) algorithm with computational functions to train for predictions as an output label. One of the key advantages of Supervised modeling use is the prediction of an output based on prior experience for which knowledge of dataset class use is required.

## 3.0 Results and discussions

The presentation of results is in two parts. The first part deals with a display of results without random sampling techniques and covers performance metrics such as receiver operating characteristics and learning curves with cross-validation, true positive rates, true negative rates, false positives, false negatives, positive predicted values, negative predicted values and the second part with random sampling technique.

Table 2.0 Sensitivity and Specificity evaluation results (pre-sampling)

| Model | FNR(%) | TNR(%) | FPR(%) | PPV(%) | NPV(%) | TPR(%) | AUC SCORE(%) | PREDICTION ACCURACY(%) | BALANCED ACCURACY(%) |
|---|---|---|---|---|---|---|---|---|---|
| **Gradient boosting** | **0.13** | **9.09** | **90.91** | 98.75 | 50 | **99.87** | **97.00** | **99.00** | **55.00** |
| **XGBC** | **0.51** | **13.64** | **86.36** | 98.80 | 27.27 | **99.49** | **99.00** | **98.00** | **57.00** |
| **Logistic regression** | **0.06** | **13.64** | **86.36** | 98.81 | 75 | **99.94** | **94.00** | **99.00** | **57.00** |
| **Random forest** | **0.57** | **9.09** | **90.91** | 98.74 | 20.16 | **99.25** | **97.00** | **98.00** | **54.00** |
| **SupportVector machine** | **0.00** | **0.00** | **100** | 98.62 | 66.67 | **100** | **89.00** | **99.00** | **57.00** |

Table 2.1 Sensitivity and Specificity results (Over-sampling)

| Model | FNR(%) | TNR(%) | FPR(%) | PPV(%) | NPV(%) | TPR(%) | AUC SCORE(%) | PREDICTION ACCURACY(%) | BALANCED ACCURACY(%) |
|---|---|---|---|---|---|---|---|---|---|
| **Gradient boosting** | **20.84** | **98.91** | **1.09** | 98.67 | 82.19 | **79.16** | **95.00** | **88.90** | **89.03** |
| **XGBC** | **2.38** | **100** | **0.00** | 100 | 97.61 | **97.62** | **100.00** | **98.79** | **98.81** |
| **Logistic regression** | **17.46** | **94.59** | **5.41** | 94.01 | 84.04 | **82.54** | **93.00** | **87.84** | **88.57** |
| **Random forest** | **1.81** | **100** | **0.00** | 100 | 98.17 | **98.19** | **100.00** | **99.08** | **99.09** |
| **SupportVector machine** | **20.53** | **91.63** | **8.37** | 90.71 | 81.28 | **79.47** | **91.00** | **85.47** | **85.55** |

Table 2.2 Sensitivity and Specificity evaluation results (Under-sampling)

| Model | FNR(%) | TNR(%) | FPR(%) | PPV(%) | NPV(%) | TPR(%) | AUC SCORE(%) | PREDICTION ACCURACY(%) | BALANCED ACCURACY(%) |
|---|---|---|---|---|---|---|---|---|---|
| **Gradient boosting** | **11.25** | **98.26** | **1.74** | 98.00 | 90.06 | **88.75** | **98.00** | **93.59** | **93.5** |
| **XGBC** | **2.13** | **100.00** | **0.00** | 100.00 | 97.99 | **97.87** | **100.00** | **98.97** | **98.93** |
| **Logistic regression** | **16.61** | **94.58** | **5.42** | 93.68 | 85.52 | **83.39** | **93.00** | **89.09** | **88.98** |
| **Random forest** | **1.29** | **100** | **0.00** | 100.00 | 98.77 | **98.71** | **100.00** | **99.37** | **99.35** |
| **SupportVector machine** | **18.55** | **91.15** | **8.85** | 89.87 | 83.6 | **81.45** | **92.00** | **86.39** | **86.3** |

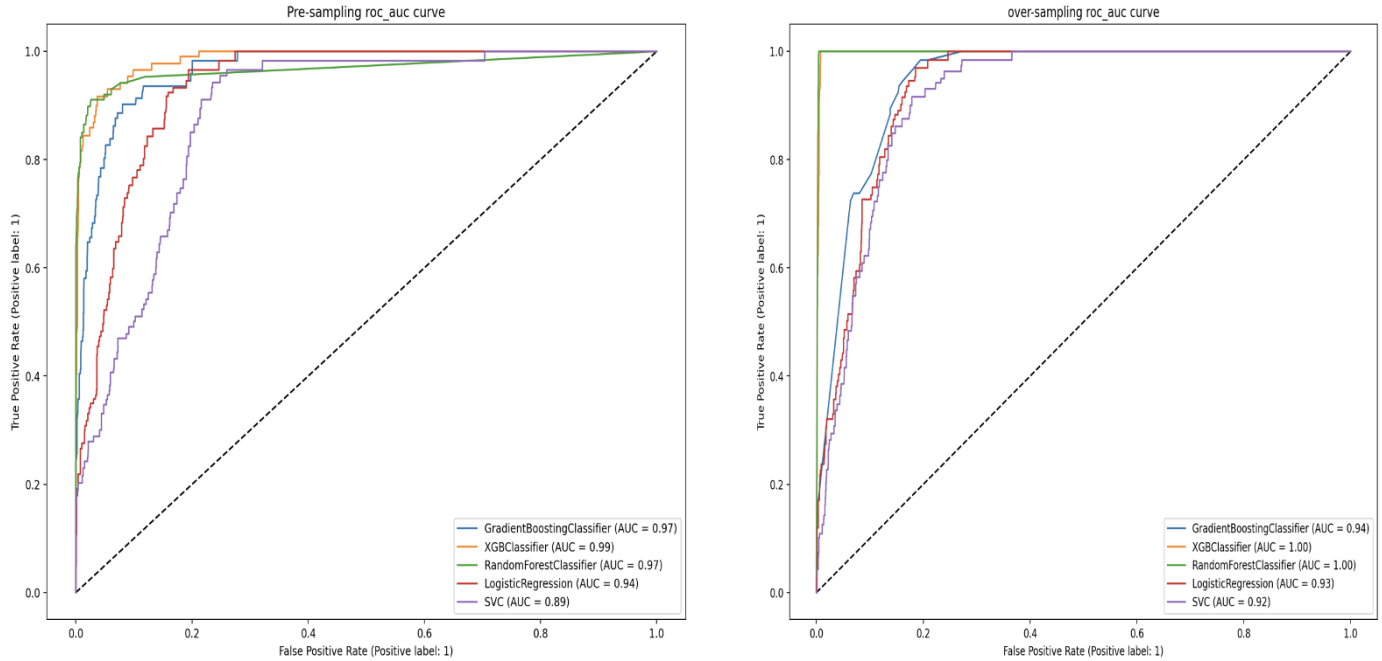



Figure 2.6 roc_auc characteristic curves. a=pre-sampling roc_auc curve ; b=over-sampling roc_auc curve

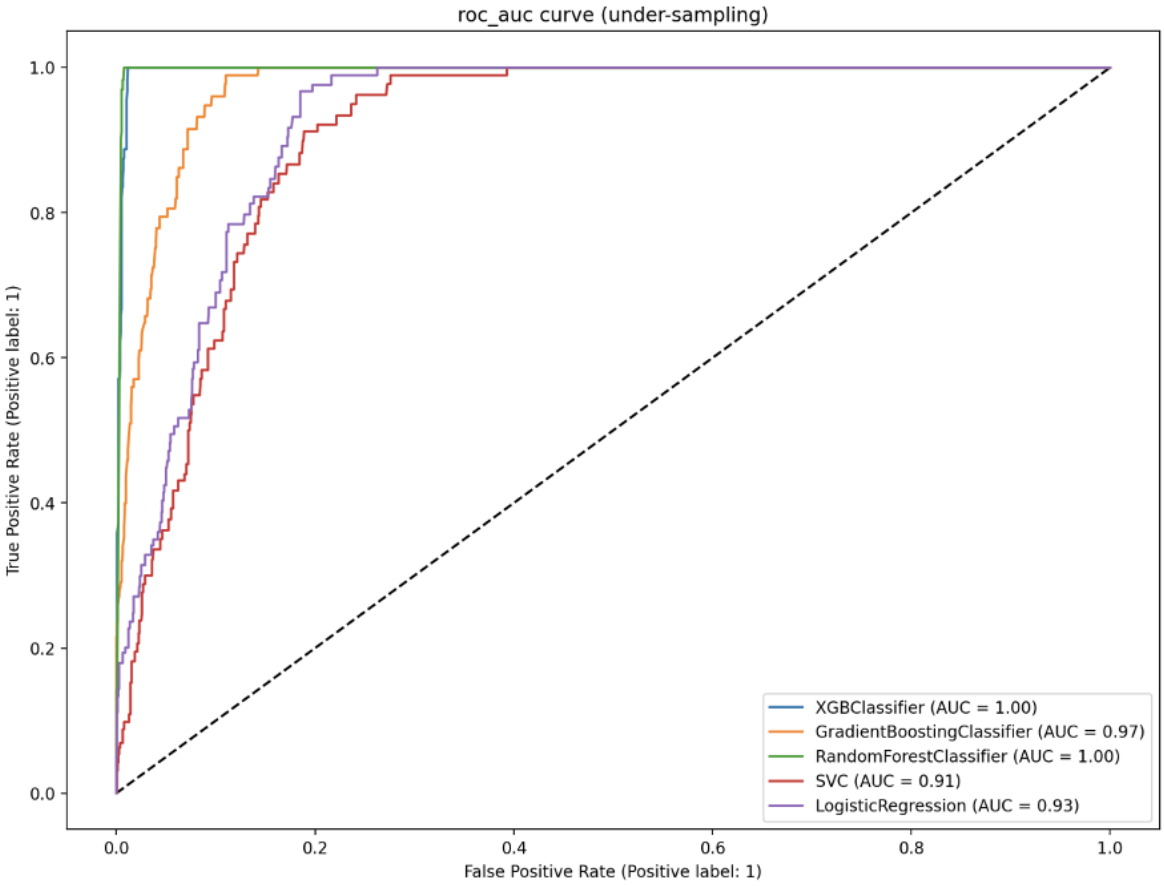
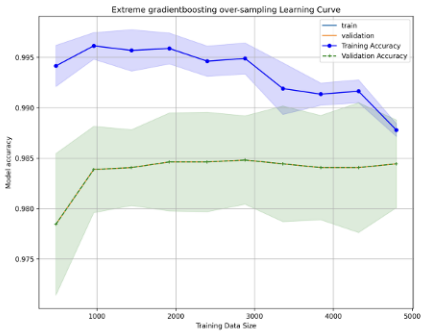
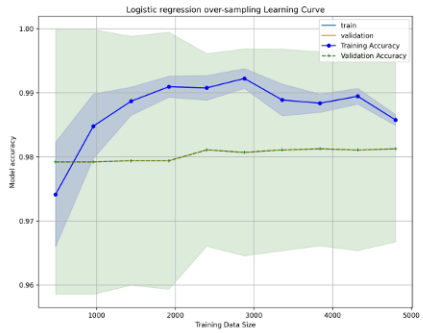


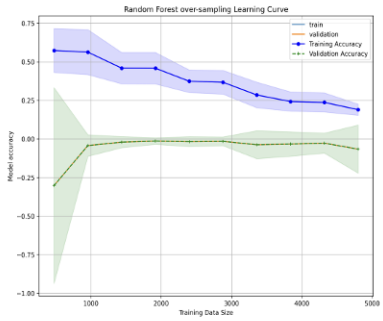

Figure 2.7 roc_auc under-sampling



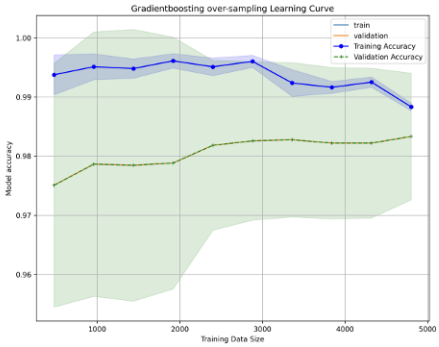

(a= Extreme gradient boosting)



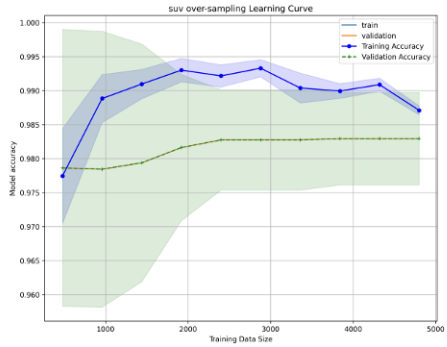

(b=Logistic regression)




(c=Random forest)

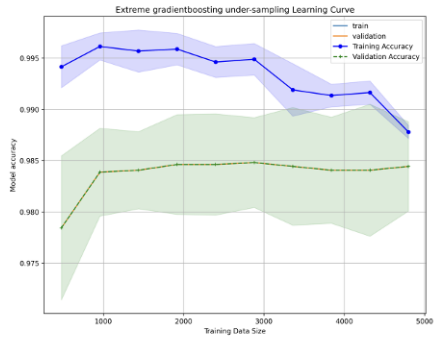



(d= Gradient boosting classifier)

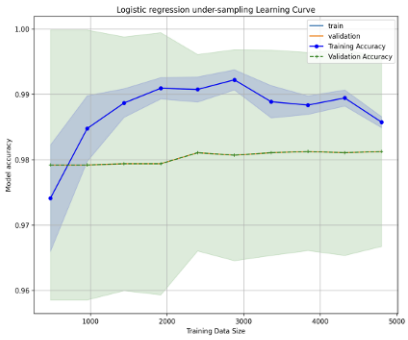



(e=Support vector machines)

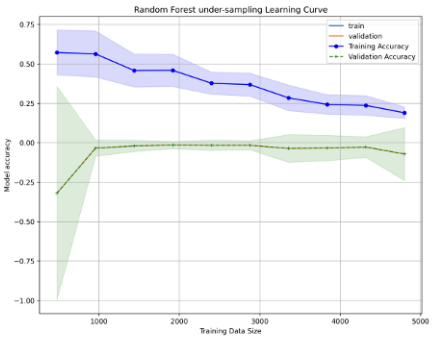
Figure 2.8 Over-sampling roc_auc learning curve



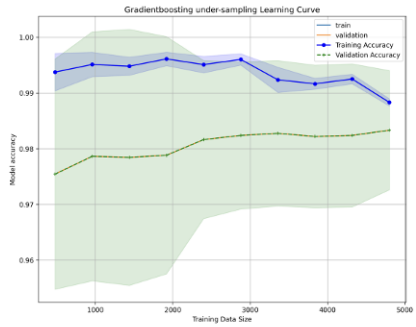

(a=Extreme gradient boosting)



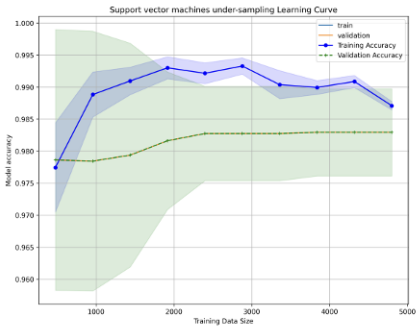

(b=Logistic regression)




(c=Random forest)

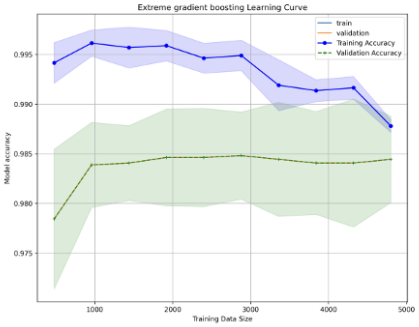



(d=Gradient boosting)

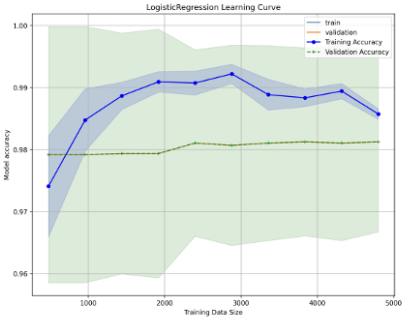



(e=Support vector machines)

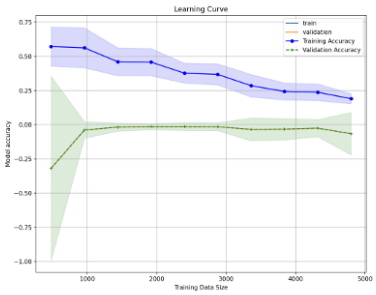
Figure 2.9 Under- sampling roc_auc learning curves



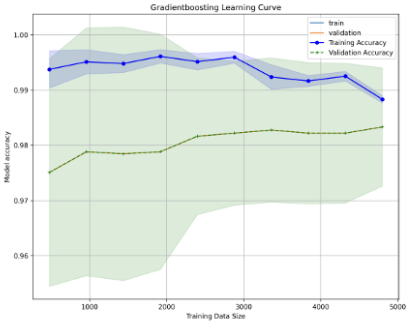

(a=extreme gradient boosting)



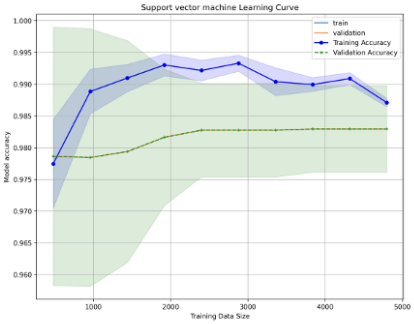

(b=logistic regression)




(c=random forest)




(d=gradient boosting)




(e=support vector machines)

Figure 2.10 Pre-sampling learning curves

## Discussions

Evaluation metrics and predictive performance results are presented as follows: Tables 2.0, 2.1 and 2.2 are results of specificity and sensitivity from pre-sampling, over-sampling and under-sampling activities. Table 2.0 describes the results obtained before applying random sampling techniques (pre-sampling), Table 2.1 represents results from applied over-sampling, and Table 2.2 represents the under-sampling technique.

Receiver operating characteristic curve scores (roc_auc) recorded in Tables 2.0, 2.1 and 2.2 were obtained from the display of graphs in Figure 2.6 (a=pre-sampling), 2.6 (b=oversampling) and Figure 2.7 (under-sampling).

Additional evaluation metrics presented in Figure 2.8 (a, b, c, d, e) represent over-sampling learning curves, Figure 2.9 (a, b, c, d, e) shows results from under-sampling and Figure 2.10 (a, b, c, d, e).

In Figure 2.1, three graphs represent the dataset distribution of patient age, gender and output class. These graphs are represented as the first graph- age distribution, the second graph- distribution by gender and the third graph is represented by outcome or output distribution. The mean age is approximated at 64 years, the minimum age is 18 years, and the maximum age is 111. The distribution of age, as shown in Figure 2.1 (first graph), follows a normal Gaussian distribution. The distribution of gender, as shown in the second graph, indicates an unequal distribution between the female majority and the male minority. Output class distribution, as indicated (third graph), shows an imbalance distribution, and this is confirmed in Figure 2.2, visualization of output class distribution.

Evaluation of performance with area under the curve (auc_roc) for pre-sampling (Table 2.0), over-sampling (Table 2.1) and under-sampling (Table 2.2) show minimum prediction auc_roc score of 89% in pre-sampling by support vector machines. The high auc_roc scores in pre-sampling, over-sampling and under-sampling indicate a statistically insignificant impact of over-sampling and under-sampling application of random sampling techniques used in this context. However, as recorded in Tables 2.0, 2.1 and 2.2, average prediction accuracy scores (balanced accuracy) obtained from minority and majority classes on sensitivity and specificity performance show a statistically significant difference between scores obtained at the pre-sampling stage and those using applied random sampling techniques. The lowest score obtained with applied sampling is 85.55%, as against 54% in pre-sampling.

Other performance evaluation metrics indicated in Tables 2.0, 2.1, and 2.2 show differences between pre-sampling scores against over and under-sampling use and these are FNR, TNR, FPR, NPV and balanced accuracy. However, in sharp contrast to observed differences, scores as shown for FPR and TPR determination show close score associations with insignificant differences for pre-sampling, over-sampling and under-sampling.

Evaluating model performance behavior with Learning curves indicates the following: No significant difference in performance behavior as observed in pre-sampling, over-sampling and under-sampling. The performance behavior of models shows similarities in both instances. Overall performance of the gradient boosting classifier shows continuous improvements over time with more training examples, increasing accuracy score over more training examples. It generalizes well on unseen datasets on cross-validation in both pre-sampling and post-sampling (over-sampling and

under-sampling).

## Conclusions

Similar performance metrics showing close associations and diverse performance in other measures, as shown in this research study, confirm that extensive model evaluation is needed to determine the best model performance in both balanced and imbalanced dataset classifications. We have demonstrated with learning curves and other important metrics how future predictive modeling performance determination in healthcare systems can be influenced by these results. It also brings into focus an understanding of model performance evaluation metric use, especially in class imbalanced datasets for the determination of the best performing predictive model.

## Declarations

Authors Contributions: Concept, Michael Owusu-Adjei,
Methodology**:** Michael Owusu-Adjei, Twum Frimpong, Gaddafi Abdul-Salaam
Supervision: James Ben Hayfron-Acquah.

### Funding

This publication is an extract from an academic research that is self-funded by the student. No external funding is involved.

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Approved

### Availability of data and Materials

The dataset used for analysis is available through upload upon request.

### Conflicts of Interest

The author(s) declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

We acknowledge the support and cooperation of the management and staff of Kwahu Government Hospital, Ghana.

## References

[1] W. Yang, J. Zhang, and R. Ma, “The prediction of infectious diseases: A bibliometric analysis,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 17, pp. 1–19, 2020, doi: 10.3390/ijerph17176218.

[2] O. E. Santangelo, V. Gentile, S. Pizzo, D. Giordano, and F. Cedrone, “Machine Learning and Prediction of Infectious Diseases: A Systematic Review,” *Mach. Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 175–198, 2023, doi: 10.3390/make5010013.

[3] E. L. Ray and N. G. Reich, “Prediction of infectious disease epidemics via weighted density ensembles,” *PLoS Comput. Biol.*, vol. 14, no. 2, pp. 1–23, 2018, doi: 10.1371/journal.pcbi.1005910.

[4] H. W. Gichuhi, M. Magumba, M. Kumar, and R. W. Mayega, “A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in Mukono district.,” *PLOS Glob. Public Heal.*, vol. 3, no. 7, p. e0001466, 2023, doi: 10.1371/journal.pgph.0001466.

[5] N. V. Korneev, J. V. Korneeva, S. P. Yurkevichyus, and G. I. Bakhturin, “An Approach to Risk Assessment and Threat Prediction for Complex Object Security Based on a Predicative Self-Configuring Neural System,” *Symmetry (Basel).*, vol. 14, no. 1, 2022, doi: 10.3390/sym14010102.

[6] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, pp. 243–248, Apr. 2020, doi: 10.1109/ICICS49469.2020.239556.

[7] F. Rodríguez-Torres, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, “An Oversampling Method for Class Imbalance Problems on Large Datasets,” *Appl. Sci.*, vol. 12, no. 7, 2022, doi: 10.3390/app12073424.

[8] I. C. S. Overview, “on the Performance of,” *IEEE Veh. Technol. Conf.*, vol. 24, no. 1, pp. 645–660, 2014, doi: 10.1007/978-3-030-47436-2.

[9] J. Hernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8258 LNCS, no. PART 1, pp. 262–269, 2013, doi: 10.1007/978-3-642-41822-8_33/COVER.

[10] V. García, J. S. Sánchez, A. I. Marqués, R. Florencia, and G. Rivera, “Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data,” *Expert Syst. Appl.*, vol. 158, p. 113026, Nov. 2020, doi: 10.1016/J.ESWA.2019.113026.

[11] D. Lee and K. Kim, “An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data,” *Expert Syst. Appl.*, vol. 184, p. 115442, Dec. 2021, doi: 10.1016/J.ESWA.2021.115442.

[12] D. H. Jeong, S. E. Kim, W. H. Choi, and S. H. Ahn, “A Comparative Study on the Influence of Undersampling and Oversampling Techniques for the Classification of Physical Activities Using an Imbalanced Accelerometer Dataset,” *Healthc.*, vol. 10, no. 7, 2022, doi: 10.3390/healthcare10071255.

[13] A. M. Sowjanya and O. Mrudula, “Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms,” *Appl. Nanosci.*, vol. 13, no. 3, pp. 1829–1840, 2023, doi: 10.1007/s13204-021-02063-4.

[14] M. Saul and S. Rostami, “Assessing performance of artificial neural networks and re-sampling techniques for healthcare datasets,” *Health Informatics J.*, vol. 28, no. 1, 2022, doi: 10.1177/14604582221087109.