Journal of Systemics, Cybernetics and Informatics (2025) 23(2), 15-22
https://doi.org/10.54808/JSCI.23.02.15

Effect of Data Imbalance in Predicting Student Performance in a Structural Analysis Graduate Attribute-based Module Using Random Forest Machine Learning

Masikini LUGOMA
Department of Mining, Minerals and Geomatics Engineering,
University of South Africa
Pretoria, South Africa.

Abel Omphemetse ZIMBILI
Department of Civil and Environmental Engineering and Building Science,
University of South Africa
Pretoria, South Africa.

Masengo ILUNGA
Department of Civil and Environmental Engineering and Building Science,
University of South Africa,
Pretoria, South Africa

Ngaka MOSIA
Department of Industrial Engineering and Engineering Management,
University of South Africa,
Pretoria, South Africa

Agarwal ABHISHEK
Department of Mechanical Engineering, College of Science and Technology,
Royal University of Bhutan,
Phuentsholing, Bhutan

ABSTRACT

This study uses Random Forest algorithm to model students' final year mark in an engineering technology module taught by the University of South Africa. The algorithm uses a supervised learning classification technique to map the different assessment marks and the final mark. Hence, the latter are labelled instances whereas the former constitute the features. Random Forest (RF) has been applied to Structural Analysis 3, which takes into consideration the graduate attribute concept or level of competence as far as assessments are concerned. Firstly, the RF is subjected to imbalanced binary classes, then balanced classes are achieved by Synthetic Minority Oversampling Technique (SMOTE) and class weights adjustment techniques. The results showed that SMOTE brought an improvement in accuracy of 3%. It was also revealed that an increase of 4, 15 and 9% in precision, recall and F1-Score were observed in predicting non-competent students. An increase of 4 and 3% was noticed in the case of the precision and F1-Score respectively in predicting competent students, whereas the recall did not display any change. Despite the RF with SMOTE overperformed standard RF and RF class weights adjustment, all three algorithms were good candidates in the prediction of student performance. RF-SMOTE could be suggested as a guiding instrument when dealing with imbalanced data.

Keywords: Random forest, algorithm, prediction, assessment, teaching, learning, class imbalance

1. INTRODUCTION

Universities and tertiary colleges usually decide the students' fate based on the outcome of assessments, which determine ultimately the competency of the students. In this regard, teaching staff have the task to prepare their students before any assessment is written. This is done for all the modules in the curriculum of a given programme. Teaching and learning should be linked to pedagogy to ensure learning goal is reached. For instance, peer learning will require staff to explore teaching strategy that stimulates collaboration [1]. Students' individual learning style could be affected by the interactive pedagogical approaches, which could enhance a reflective learning style in adult students [2]. Through interaction with the teacher, students had the capability of augmenting their lexical repertoire by selecting words and expressions from texts, and audio-visuals [3]. Therefore, the module should be reviewed to accommodate incorporated diverse teaching strategies, which should align with students' experiences.

Weighting the different assessments for the calculation of the final mark/score is explicitly used in education.

For engineering technology qualifications offered in South Africa, the accent is on graduate attributes (GA), the student should demonstrate for him or her to be declared competent, throughout the academic programme. This is mandatory as stipulated by the Engineering Council of South Africa, for further registration of the student, as professional. Therefore, through well-structured course and module design, GAs should be embedded in teaching and learning and proved to be assessed. For a module that includes GA assessments, the student must achieve 50% overall and demonstrate the required competence level in specific assessments.

Currently, technology and intelligent tools have increasingly enhanced the efficiency of teaching and learning activities. Therefore, educators, students and institutions of higher education at large have embraced it, where possible.

The inherent necessity of machine learning (ML) techniques requiring datasets, has painted data processing, simulation and optimization applications. Students' performance in education has been noticed in the current literature, e.g. [19], [20] and [21]. There is no doubt there are a myriad of ML algorithms, e.g. just to name few artificial neural network (ANN), support vector machine (SVM), logistic regression, K-nearest neighbor (K-NN), classification tree and Naïve Bayes [10]. Thus, there is no size that fits all in this sub-field of Artificial Intelligence (AI). The current study employs Random Forest (RF) algorithm for the purpose of predicting student performance based solely on continuous assessments. In some instances, RF has been appreciated to offer some advantages, over other algorithms, e.g. RF offers a variable importance metric, and can deal with datasets of high dimensions [11]. Moreover, RF may outperform classification algorithms that use ensemble learning models like bagging and boosting [12]. Hence, their popularity is due to the fact the RF displays higher accuracy and resistance to noise than a single classification technique [13].

The popularity of RF among machine learning models is undeniable. It is one of the ensemble learning techniques, which draws its foundation from several decision trees such that their predictions are combined for increasing accuracy, by reducing overfitting.

Besides, the issue of imbalanced data is common for ML applications, that could affect the accuracy of the ML during training and more during prediction because of the minority instances being under-represented. This may lead to a situation where ML models have higher performance on the majority class as opposed to the minority class [7]. Techniques to balance the datasets have been documented, e.g. [16] to remove the model bias towards the majority class.

Binary class, even multi-classes are common for classification problems in supervised ML.

The rest of the paper is organised as follows: random forest foundation in a supervised machine learning context, including hyperparameter tuning. The data and methods are presented, in the pursuit of the purpose of the study, for the predication of student performance. Then, the findings are presented, accompanied by a discussion. At the end, the findings yield to a conclusion and suggestions.

In what follows, model, technique and algorithm will be used interchangeably in the context of machine learning. Similarly, module and course will have the same meaning.

2. RANDOM FOREST CLASSIFICATION SUPERVISED MACHINE LEARNING

Random Forest ML basis

In a supervised classification learning, RF maps the features to the labels, which are represented by classes. As an ensemble learning technique, RF works on a set of forest, which represents decision trees. Different data sub-sets are used to train these trees; however, training happens independently from each tree to the other such that the model prediction is performed by deriving performance from all trees. Figure 1 displays a RF structural flow. In a classification problem, the prediction is made from the majority vote whereas in regression problems, an average of predictions from the training of all trees is used. It is understood that several trees (models) are combined during training of RF, in a way that the accuracy of the final model increases. This may avoid overfitting if individual trees could be used without being aggregated into the final model.

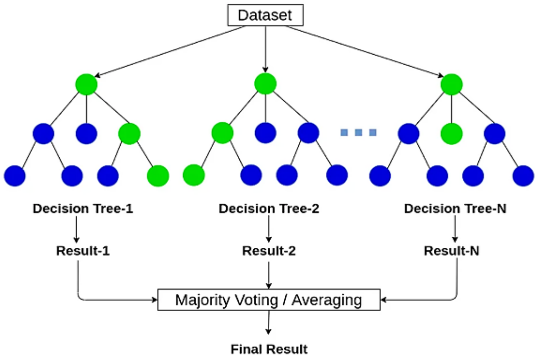


Figure 1. Random Forest structural flow, source: [4]

Dataset splitting

For supervised classification applications, there is no universal rule to split the dataset into training and testing. In this respect, there is no robust methodology that has been adopted unequivocally for machine learning. The 80/20 split for training and testing has been quite dominant [15]. However, other splits have been used, e.g. 75/25; 65/35, and 70/30 respectively. When dealing with sufficient large datasets, testing can be further used to accommodate validation, which is used to divide testing into equal proportions.

Hyperparameter tuning of RF

Hyperparameters tunning enables to find optimum parameters, which are combined to increase the performance of a ML model, e.g. RF model. Important hyperparameters for this model include the following [4]:

-Number of Trees (n_estimators), which increase accuracy; however, yields to computational burden.

-Maximum number of features, which should be considered when splits are made at each node.

-Maximum tree depth, which is the limit of the individual trees, and can lead to overfitting when shallow trees are used. Underfitting may occur when trees are too large.

-Minimum samples per split (min_samples_split), which is needed for splitting an internal node and enables to reduce overfitting when higher values are used.

-Minimum Samples per Leaf (min_samples_leaf), which is at the leaf node as required, by avoiding overfitting when small leaves are used.

-Bootstrap Sampling (bootstrap), for controlling bootstrap samples during construction of trees. A False setting of this parameter yields to using the whole dataset for training each tree.

Grid Search, Random Search, and Bayes Search are some of the optimization techniques for hyperparameter tuning that can be used for RF models.

Class imbalance handling

There are different techniques to deal with minority instances in the datasets. For example, [6] used Synthetic Minority Oversampling Technique (SMOTE) technique for oversampling the minority class and adjusting class weights techniques. These 2 techniques will be used in this study. The process of augmenting the number of minority class instances, through duplication or synthesis, leads to creating new instances or oversampling [7]. This process makes interpolation among the existing minority instances to have realistic generated samples. For data imbalance, SOMTE is among the most dominant techniques for extra sampling generation [14]. Although SMOTE can significantly improve the learning, its main drawback is that when generating the synthetic examples, it does not take into consideration the neighboring examples from other classes, which can increase the overlapping of classes and introduce additional noise [16]. Other variants of SMOTE have been developed, i.e. Borderline SMOTE, Support Vector Machine (SVM-SMOTE), and Adaptive Synthetic Sampling (ADASYN) [8]. It is also possible to proceed with undersampling, which reduces the majority class to bring it to the size of the minority class [8]; [7]. [8] stated that, empirically the data ratio of at least 25%, between the minority class and the size of the majority class does not affect the model performance by large margins. Advanced ensemble techniques may be used for handling imbalanced data, e.g. through Balanced Random Forest (Adjusts class weights in random forests), Cost-Sensitive Learning (incorrect predictions of the minority class are penalised more than the majority class) and boosting with weighted Loss [7].

Statistical metrics for algorithm evaluation

During training, parameters of ML are obtained and during testing, the parameters are used for prediction on the dataset ML was not exposed to before. It is during this last phase that the robustness of ML is assessed. There are several indicators for classification problems [5], however this study was limited mainly to the classification report, besides accuracy. Hence, the following metrics were used: Accuracy, Precision, Recall, and F1-Score. These metrics are essentially derived from the elements of the confusion matrix. Therefore, the performance of the ML classification algorithms, including RF can carried out by the following equations.

$$Confusion\ matrix = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precison\ (positive) = \frac{TP}{TP+FP} \quad (3)$$

$$Precision\ (negative) = \frac{TN}{TN+FN} \quad (4)$$

$$Recall\ (positive) = \frac{TP}{TP+FN} \quad (5)$$

$$Recall(negative) = \frac{TN}{TN+FP} \quad (6)$$

$$F_1 Score = 2\frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

Where:

True Positives (TP) and True Negatives (TN) correspond to the correctly predicted positive classes and negative classes respectively,

False positives (FP) and False negatives (FN) correspond to the incorrectly predicted positive classes and negative classes.

The precision positive(negative) displays the fraction TP (TN) predicted, with respect to all the positive (negative).

The above equations can be applied to binary classification problems.

The recall (negative) or True Negative Rate called (Specificity) shows the fraction of correctly negative predicted (TN) instances with respect to the actual total number of negatives. The values 1.0 and 0.0 show the best and worse specificity respectively. Hence, the recall (positive) or sensitivity shows the Rate of True Positives, which is the proportion of correctly predicted (TP) instances with respect to the actual total number of positives.

F1-Score ranges from 0 to 1. It is a metric which determines the test accuracy. It is computed from precision and reminders.

It should be noted that accuracy, precision and recall may suffer from imbalanced data, by favoring the majority class, whereas F1-Score seems to cope with imbalanced data [9]; [13]. Due to their shortcomings, [9] introduced the probabilistic approach (dealing with confidence interval) of the precision, recall and F1-Score. For the same reason, other metrics like the Kappa coefficient and Matthews' correlation coefficient (MCC) were proposed. The MCC generated scores more informatively and truthfully for classification problems than accuracy and F1 score [17].

3. DATA AVAILABILITY AND METHODS

Data used

Structural Analysis 3 is a 12-credit course taught at advanced diploma course, with 1 credit equating to 10 hours. It is part of the curriculum of the Department of Civil and Environmental Engineering, and Building Science, at the University of South Africa. Graduate Attribute 1 (GA1) is the focus of this course and deals with identifying, formulating, analysing and solving broadly defined problems in Civil Engineering. This GA is assessed in the project component of the course. Hence, the dataset had 146 instances during semester 2 of the year 2023.

Methods.

From the dataset, the categorical class (fail, pass) was transformed into the binary class (0, 1). This enables a good mapping between assessments (features) and labelled binary class instances for the Structural Analysis 3 course. Missing data were inexistent; however, the python code was run in this respect. The output confirmed, indeed there was no missing value. The dataset was split into 80/20 for training and testing respectively.

The correlations among features, and between features and labels were adopted from the companion paper.

The dataset imbalance check was conducted by counting each class. SMOTE and class weights adjustments techniques were used for the application of RF.

The random forest algorithm was chosen with the following hyperparameters, n_estimators=100, max_features= 100, and random_state = 42.

The model evaluation was summarized through the confusion matrix, subsequently the classification report.

Three cases were compared, in terms of performance: RF with data imbalance, RF with data imbalanced handled through SMOTE and through class weights adjustment.

No cross validation was performed at this stage.

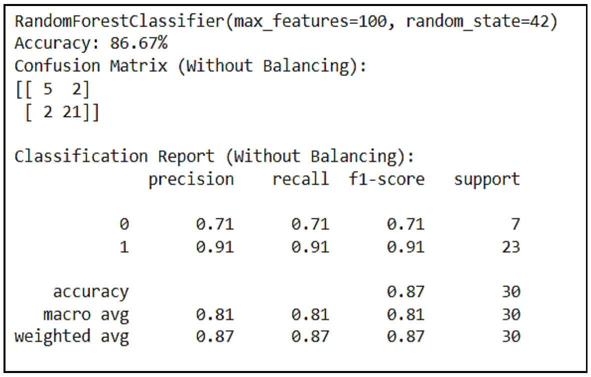
4. RESULTS AND DISCUSSION

Assessments marks and the final score were represented by Ass1, 2, 3, 4 and Result, respectively. These are respectively the features and labels respectively. The statistical characteristics of all features as well as the correlations among features, and between features and the labels were already computed and analysed in the companion paper.

The status of dataset as displayed in Figure 2 showed that the classes were imbalanced with 0 forming the minority, hence 1 forming the majority. This could justify the introduction of the 2 techniques for handling data imbalance. The ratio between the instance minority size to instance majority size was 30 to 116, close to 26%.

Results displayed in Table 1 below showed that RF without balancing data, had a relatively high accuracy of 87%, which tends to 100%. This means that the higher percentage of TP+TN out of the total number testing size sample, has been correctly predicted. Based on the confusion matrix entries, the sum TP and TN (26), is far bigger than the sum FP and FN (4). The values of precision, recall and F1-Score metrics were relatively higher (91%) for the "pass" class than those of the "fail" class (71%). Hence, the rate of the performing students predicted was 91% as opposed to that of the underperforming students (71%). In both cases, RF was shown to be a good candidate for prediction of the students' performance. Figure 2 is a visual representation of Table 1.

Table 1. Classification for the Random Forest, without balancing minority class.



```
RandomForestClassifier(max_features=100, random_state=42)
Accuracy: 86.67%
Confusion Matrix (Without Balancing):
[[ 5  2]
 [ 2 21]]

Classification Report (Without Balancing):
              precision    recall  f1-score   support

           0       0.71      0.71      0.71         7
           1       0.91      0.91      0.91        23

    accuracy                           0.87        30
   macro avg       0.81      0.81      0.81        30
weighted avg       0.87      0.87      0.87        30
```

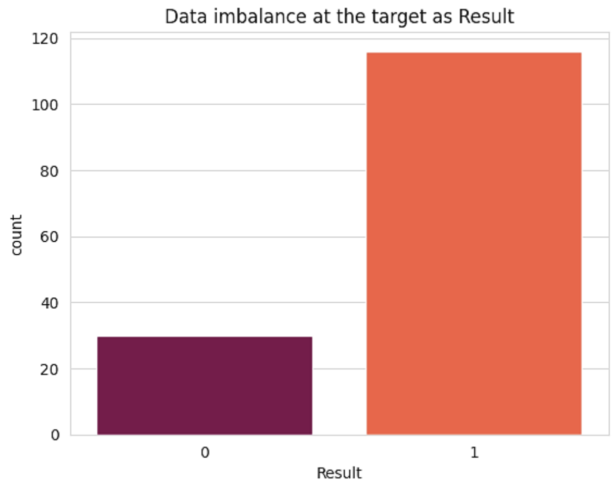


Figure 2. Class imbalance status at the target variable.

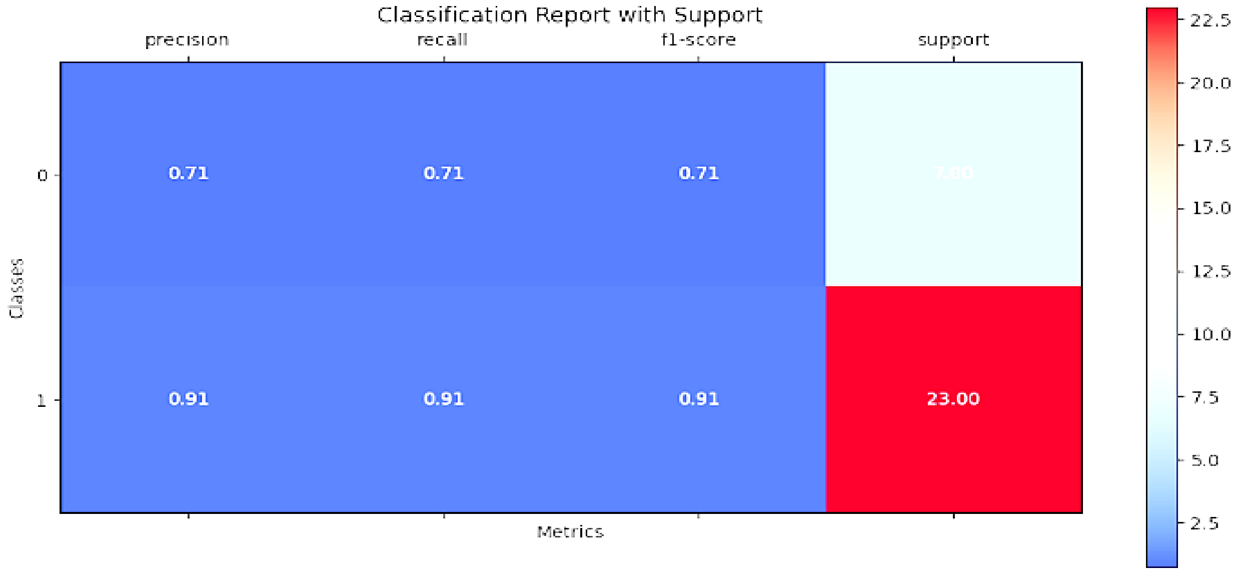
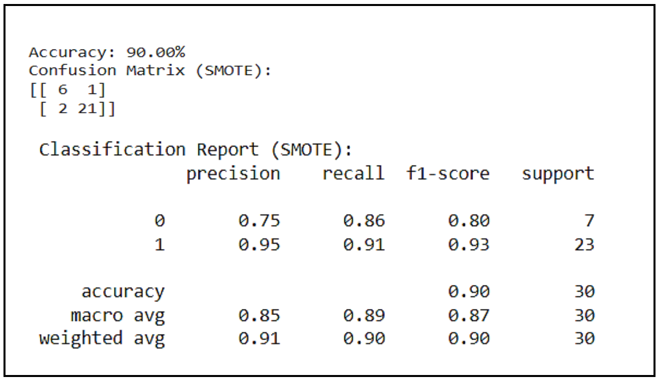


Figure 3. Classification report with support for the Random Forest, without balancing of minority class

From Table 2 and Figure 4 below, the results revealed that RF with balanced data (e.g. with oversampling SMOTE), had a relatively high accuracy of 90%, which is close to 100%. This means that the higher percentage of TP+TN out of the total number of testing size sample, has been correctly predicted. The confusion matrix entries show that the sum TP and TN (27), is way bigger than the sum FP and FN (3). The values of precision, recall and F1-Score metrics were 75%, 86% and 80% respectively for the "fail" class. In the case of "pass" class, the values of metrics were 95, 91 and 93% respectively. The predicted true positive instances represented 95% of all the positive, whereas negative instances truly predicted, with respect to all the negative, represented 75%. The predicted true positive instances represented 91% of the actual positives, whereas negative instances truly predicted, with respect to actual negatives, represented 86%. Finally, the balanced score between precision and recall was 80% and 93% for the non-performing students and performing students.

Table 2. Classification for the Random Forest, with balancing minority class, using SMOTE



```
Accuracy: 90.00%
Confusion Matrix (SMOTE):
[[ 6  1]
 [ 2 21]]

 Classification Report (SMOTE):
              precision    recall  f1-score   support

           0       0.75      0.86      0.80         7
           1       0.95      0.91      0.93        23

    accuracy                           0.90        30
   macro avg       0.85      0.89      0.87        30
weighted avg       0.91      0.90      0.90        30
```

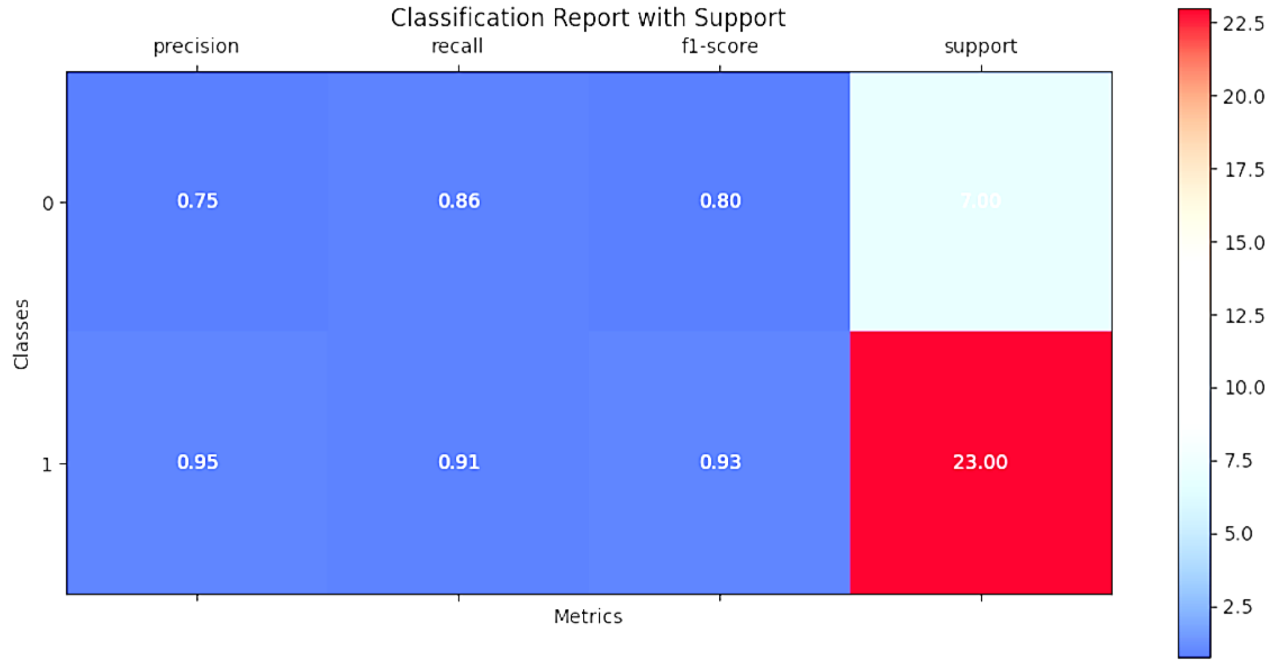
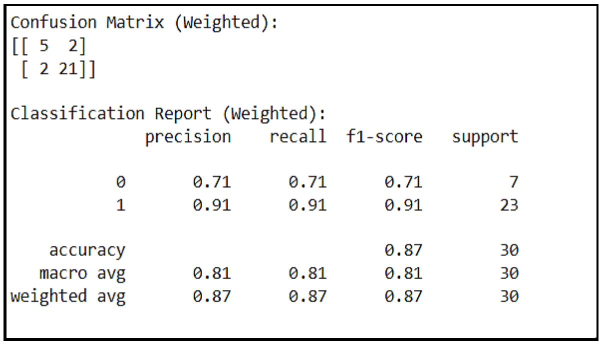


Figure 4. Classification report for the Random Forest, with balancing of minority class (SMOTE)

Comparing Table 1 (Figure 3) and Table 2 (Figure 4), there is a significant improvement in metrics for class 0 after class balance, as well as the accuracy and precision. A slight improvement was noticed, only F1-Score and no improvement was noticed for the recall.

Table 3. Classification for the Random Forest, with balancing minority class, using class weights adjustment



```
Confusion Matrix (Weighted):
[[ 5  2]
 [ 2 21]]

Classification Report (Weighted):
              precision    recall  f1-score   support

           0       0.71      0.71      0.71         7
           1       0.91      0.91      0.91        23

    accuracy                           0.87        30
   macro avg       0.81      0.81      0.81        30
weighted avg       0.87      0.87      0.87        30
```

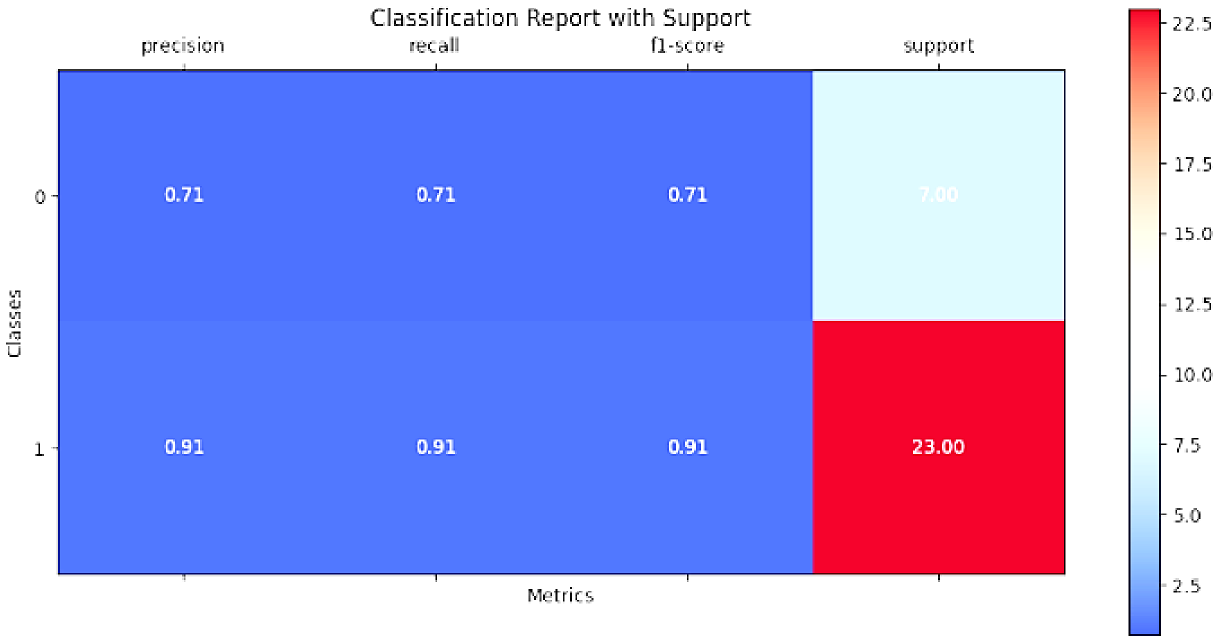


Figure 5. Classification report for the Random Forest, with balancing of minority class (with class weights adjustment)

From Table 3 above and Figure 5, the results revealed that RF with balanced data (e.g. with adjustment of class weights), were the same as those obtained in the case without data balance. Hence, one should refer to the discussion conducted on Table 1 and Figure 2. However, the comparison of the 2 algorithms after data balance revealed that RF with SMOTE outperformed the RF with class weight adjustments.

Therefore, using the classification report, the performance comparison between RF algorithm, RF with balanced data with SMOTE, RF with class weights adjustments, revealed that RF-SOMOTE outperformed the first 2 algorithms.

The RF with imbalance data was still found to be a very good candidate for the purpose of prediction of the binary class (fail or pass) undertaken in this paper. This is contrary to the ratio of least 25%, between the minority class and the majority class, that could impact on the performance the ML [8]. Unlike [18], there was no substantial reduction of false negatives after handling imbalance, as observed from the confusion matrix. This finding was only restricted to the Structural Analysis 3 module as far as precision, recall, and f1-score were concerned.

The RF, and its two balanced variants, were revealed to have the capability of predicting both competent and non-competent students. Of particular attention, the academic department together the module lecturer could put in place mechanisms to prepare underperforming students to for assessments. The RF and its variants should be reviewed as more data becomes available.

5. CONCLUSION

This study has assessed RF algorithm, RF with balanced data with SMOTE, RF with class weights adjustments, to model the prediction of students' performance in a specific engineering technology course. Assessments were instrumental as input variables. It was revealed that the non-linear mapping between input variables could be achieved with these models. Nonetheless, the RF with SMOTE achieved a relatively higher performance than the rest of models, by using accuracy, precision, recall and F1-score as statistical metrics. These findings could be explored by instructors for guidance as far as assessments for graduate attribute based-courses are concerned. Machine learning could be used as an efficient tool for quick identification of both competent and non-competent students and for modeling of their respective scores. The increase in data availability regarding assessments will mean a re-evaluation of the performance of the different algorithms. More techniques for handling imbalance data should be investigated for the RF case and could be extended to other algorithms. Further hyperparameter tuning should be carried out.

6. REFERENCES

[1] A. Raymond, E. Jacob, D. Jacob, J. Lyons, "Peer learning a pedagogical approach to enhance online learning: A qualitative exploration", **Nurse education today,** 2016 Sep 1;44:165-9.
[2] C.C. Li, M.A. Aldosari, S.E. Park, Understanding pedagogical approaches on student learning styles. **Ann Dent Oral Health**. 2021;4(1):1039-45.

[3] A. Pellicer-Sánchez, F. Boers, "Pedagogical approaches to the teaching and learning of formulaic language". Understanding formulaic language. In A. Siyanova-Chanturia & A. Pellicer- Sánchez (Eds.), Understanding formulaic language: A second language acquisition perspective, Routledge, 2018 Sep 3, pp. 153-173.

[4] A. Jain, Everything about Random Forest. 2024. https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989

[5] Ž. Vujović, Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 2021, *12*(6), pp. 599-606.

[6] N. Novriansyah, Handling Class Imbalance in Loan Approval Prediction Using XGBoost, 2024. https://medium.com/artificial-intelligence-101/handling-class-imbalance-in-loan-approval-prediction-using-xgboost-805a54ca4523.

[7] A.P. Pandelu, Handling Imbalanced Data-Oversampling, Undersampling, SMOTE, 2024. https://medium.com/@bhatadithya54764118/day-31-handling-imbalanced-data-oversampling-undersampling-smote-a9dba8c363b7

[8] H. Aguiar, What Is Imbalanced Data and How to Handle It?, 2024 https://www.turintech.ai/what-is-imbalanced-data-and-how-to-handle-it/

[9] R. Yacouby, and D. Axman, Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 2020, pp. 79–91, Online. Association for Computational Linguistics

[10] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease". ***Neural Computing and Applications***, 2018, *29*, 685-693.

[11] L. Brieman, J. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees. *Chapman and Hall/CRC,* 1984. https://doi.org.10.1201/9781315139470

[12] P.O. Gislason, J.A. Benediktsson, and J.R. Sveinsson, Random Forest for land cover classification. *Pattern Recognition Letters,* 2006, *27(4),* pp. 294-300. https://doi.org.10.1016/j.patrec.2005.08.011

[13] S. M. Malakouti, Heart disease classification based on ECG using machine learning models. *Biomedical Signal Processing and Control*, 2023, *84*, 104796.

[14] D. Elreedy, & A. F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences*, 2019, *505*, 32-64.

[15] Y. Pristyanto, A. F. Nugraha, R. F. A. Aziza, I. H. Purwanto, M. Sulistiyono, & A. Dahlan, Comparison of ensemble models as solutions for imbalanced class classification of datasets. In *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2023, January, (pp. 1-4). IEEE.

[16] J. A. Sáez, B. Krawczyk, & M. Woźniak, Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 2016, *57*, 164-178.

[17] D. Chicco, M. J. Warrens, & G. Jurman, The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *Ieee Access*, 2021, *9*, 78368-78381.

[18] N. Novriansyah, Handling Class Imbalance in Loan Approval Prediction Using XGBoost, 2024 https://medium.com/artificial-intelligence-101/handling-class-imbalance-in-loan-approval-prediction-using-xgboost-805a54ca4523

[19] R. Singh, and S. Pal, Machine learning algorithms and ensemble technique to improve prediction of students performance. *International Journal of Advanced Trends in Computer Science and Engineering*, 2020, *9*(3).

[20] R. Sharma, S. S. Shrivastava, & A. Sharma, Predicting Student Performance Using Educational Data Mining and Learning Analytics Technique. Journal of Intelligent Systems and Internet of Things, 2023, 10(2), 24-37.

[21] A. A. Saa, Educational data mining & students' performance prediction. *International journal of advanced computer science and applications*, 2016, *7*(5).