ASSESSMENT OF RESAMPLING TECHNIQUES OVER MULTI-MAJORITY DATASETS

Rose Mary Mathew, R. Gunasundari*

Abstract

In the present world, all the events are forecasted based on data extracted from the past experiences, so that machine learning is of having an importance and the dataset used for learning a model is considered as prime. Multiclass imbalanced data classification is a challenging field in machine learning. This kind of skewness can be found in all sectors of data in real world. Multimajority datasets are those datasets having data with multiple majority classes and a few data with minority class label. The spotting of minority class is pivotal in this skewed dataset. This brings about a major issue for machine learning algorithms as they take on the data as balanced distribution of classes. Exploration is going on this field and a great deal of strategies are accessible to manage imbalanced information. Resampling the dataset during pre-processing stage is one of the methods. Oversampling and undersampling are the techniques available with resampling. This work concentrates on an investigational study on the effect of resampling techniques like RUS, Near-Miss, ROS, SMOTE on multimajority class imbalanced data in different sectors.

Keywords: imbalanced, multiclass, multimajority, oversampling, undersampling

I INTRODUCTION

The identification of a class label for a data in datamining is termed as classification. A few occasions are occasionally happened with the goal that the expectation of these kind of occasions are troublesome considering their less data. This issue is looked in some genuine situations. This skewed data will not give correct predictions. Skewed data classification is a kind of supervised learning[1]. Multiclass imbalance can appear in two ways that is either in multimajority case or in multiminority case[2].

Multimajority is the scenario where majority of the classes are having high frequency and only a few classes are having low frequency. Table.1 shows an example of multimajority data distribution. By looking into Table.1 it is evitable that the data distribution is not in a balanced manner. Data is spread across five different classes and frequency of four class are in the the range of 26-22%, however the class5 is of 5%. This is termed as multi majority case. Multiminority is the scenario where majority of the classes are having low frequency and only a few classes are having high frequency[3]. Table.2 shows an example of multiminority data distribution. By looking into Table.2 it is evitable that the data distribution is not in a balanced manner. Data is spread across five different classes and frequency of four class are in the the range of 6-9%, however the class1 is of 70%. This is termed as multi minority case. Most of the machine learning algorithms cannot produce an accurate prediction from these skewed classes. This was an issue from the past years and many techniques are developed by the researchers in machine learning to handle the imbalanced data.

Internal and external are the two methodologies for dealing with imbalanced information. Internal level is also known as algorithmic level is managing the improvement of new algorithm to handle skewed data[4]. External level is also known as data level which utilizes some data preprocessing techniques to make the data balanced. Cost

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

sensitive learning is another approach which utilizes a mix of above levels, and it imposes a high cost for the minority class and try to minimise the high-cost errors[5]. Another way to deal with imbalanced data is ensemble approach, which uses ensemble learning algorithms at the data level approach or by enhancing cost sensitive framework in learning process[5].

Class	class1	class2	class3	class4	class5
Frequency	23	26	24	22	5
Total %	23 %	26%	24%	22%	5%

Table.1 Multimajority Classes

Class	class1	class2	class3	class4	class5
Frequency	70	6	8	7	9
Total %	70%	6%	8%	7%	9%

Table.2 Multiminority Classes

In this paper the first section addresses the importance of multiclass skewed data and its variants. The second section specifies the materials and methods used for the experiment. Third section describes the experiment with its results. Fourth section portrayed the conclusion and future works.

II MATERIALS AND METHODS

In this section of the paper describes the details of the dataset used in this study. The types of resampling techniques, algorithms and the performance measures used are discussed here.

2.1 DATASET DESCRIPTION

In this study we are using two datasets one obtained from the UCI Machine Learning Repository and the second one is from KEEL data repository. The first dataset is related to learning sector, and it consists of 5 attributes and 403 instances. These 403 instances are categorised in 4 classes. The classes are Very Low, Low, Medium, High. Data is distributed unequally to all the classes. There are 130 instances in High Class, 122 instances in middle class, 129 instances in Low class and 50 instances in Very Low class. The dataset can be treated as a multimajority set. Fig.1 shows the data distribution of samples.

The second dataset is Balance which consists of 625 instances and having three different classes. Frequency of data distributed over these three classes are 49,288 and 288. The dataset can be treated as a multimajority set. Fig.2 shows the data distribution of samples.

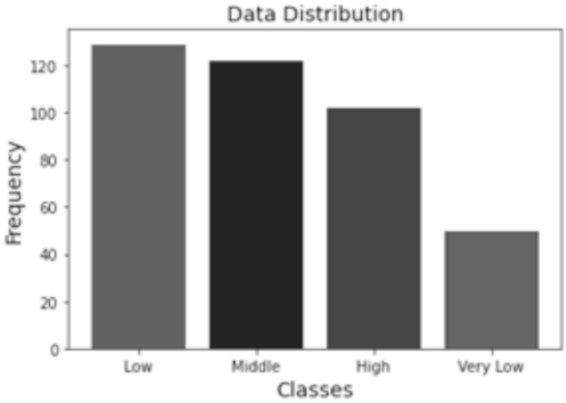

Fig.1 Data Distribution of the Education Dataset

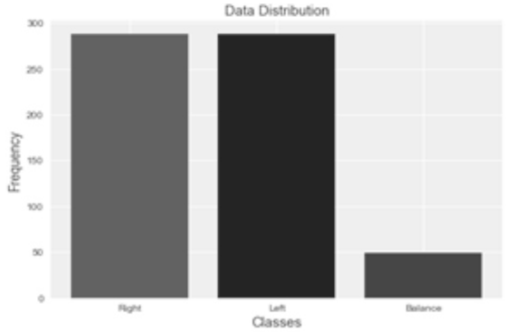

Fig.2 Data Distribution of the Balance Dataset

2.2. RESAMPLING TECHNIQUES

Resampling strategies are implemented within the data pre-processing stage to make the skewed data to a balanced one. These strategies may be implemented in methods both through including new samples to the minority classes or removing samples from majority classes to make the dataset balanced. Adding new samples to the minority classes is termed as oversampling and eradicating samples from the majority classes is termed as undersampling[5]. The Fig.3 shows the representation of oversampling and undersampling in binary data. The same can be applied in multiclass data.

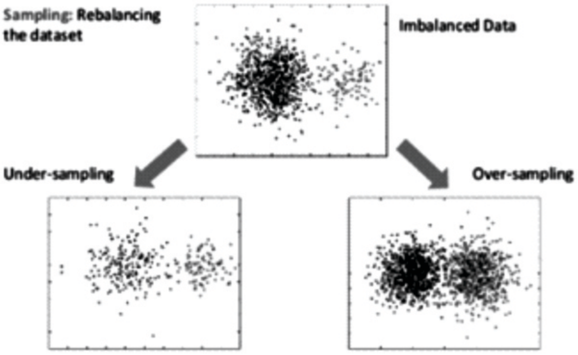

Fig.3 Comparison of undersampling and oversampling[24]

For this study four resampling techniques are used to balance the skewed data. They are ROS (Random Oversampling), RUS (Random Undersampling), NearMiss, SMOTE (Synthetic Minority Oversampling Technique) .

2.2.1 Random Oversampling (ROS)

ROS is the simplest form of oversampling approach that can be applied in the data preprocessing stage. In this technique the data in the minority classes are selected randomly and replicated in this type of manner that they ought to make a balance with the count of the majority class data[6]. This is carried out for all minority classes in the multiclass problem. Since the existing datas are replicated there will be an increase in the probability for overfitting of data [7].

2.2.2 Random Undersampling (RUS)

RUS is the simplest form of undersampling approach that can be applied in the data preprocessing stage. In this technique the data in the majority classes are selected randomly and the selected data are eleminated in such a way that they ought to make a balance with the count of the minority class data[8]. This is carried out for all majority classes in the multiclass problem. Since the original data points are eliminated that will make a loss in learning process of the data [1].

2.2.3Synthetic Minority Oversampling Technique (SMOTE)

SMOTE generates new synthetic datapoints rather than duplicating the existing datapoints and hence it is an oversampling technique. To generate new datapoints k-nearest neighbours method is used. The value of k is based on the number of datapoints generated. Using the distance formula the distance between feature vector and neighbouring points are assessed. The difference in the distance is noted for various points and this difference is multiplied with a random value in the set (0,1). The value of the product is added to the feature vector as the new data value[9].

2.2.4 Near Miss

Near-miss is an undersampling technique. This technique randomly eliminates samples from the larger class. If two data points are belonging to different classes and are very close to each other in the data distribution, this strategy eliminates the datapoint of the majority class. This technique finds the distance between all the points in the majority class with the points in the minority class. Then it selects points of the majority class that is having the shortest distance with the points of the minority class. These n points need to be taken for elimination.

2.3 CLASSIFIERS SELECTED

In this study, three different classifiers are used to create the models. The different classifiers are K-NN, Support Vector Machines (SVM) and Random Forest(RF) .

2.3.1K-NN

K-Nearest Neighbour classifier is an algorithm which follows lazy learning technique[10]. In this method k number of nearest neighbours are identified for the test data and from these k neighbours there classes are identified and the class which is having the highest frequency is fixed as the class of test data[11][12]. Fig.4 shows the representation of k-nn classification.

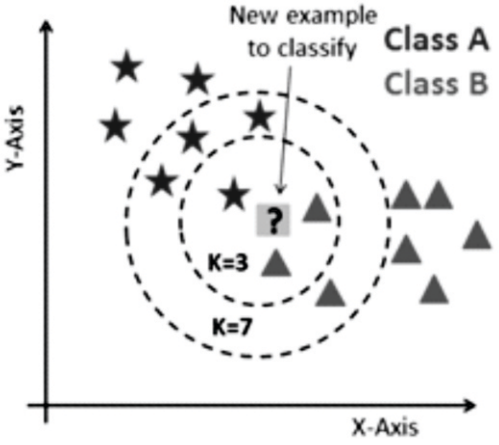

Fig.3 Comparison of undersampling and oversampling[24]

2.3.2 Support Vector Machines

Support vector machine is a popular algorithm for doing the classification tasks. SVM creates a decision boundary for the data points and this decision boundary is called hyperplane. SVM makes this hyperplane with maximum distance away from the data points. The hyperplane used by SVM is entitled as maximum margin hyperplane[13][14]. Fig.6 represents the support vector machine[15].

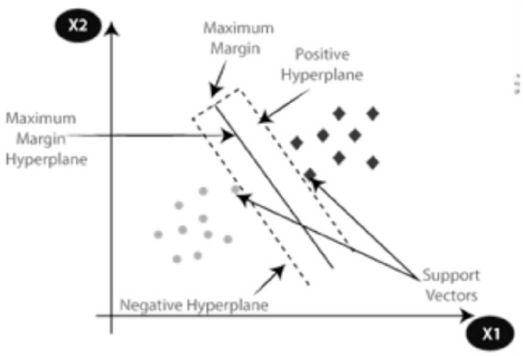

Fig.6 Support Vector Machine [26]

2.3.3 Random Forest

Random forest is an ensemble algorithm which uses bagging method[16]. Random forest is a combination of different decision trees. This algorithm takes the results from all the trees, and it takes an average or considers a majority poll to produce the outcome. This strategy improves the predictive accuracy of the model[17]. Fig.7 shows the representation of random forest in general.

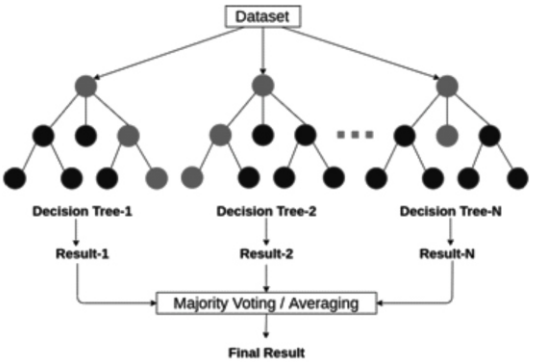

Fig.7 Random Forest [27]

III PERFORMANCE MEASURES

Classification models are constructed with several resampling techniques. The performance of these models is evaluated by using confusion matrices. Confusion matrices provide a view of actual values and predicted values[18]. If the predicted value and actual value are same, then we can call it as correct classification otherwise it is termed as incorrect classification. The values in confusion matrix are True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN)[18]. Using these values accuracy of the model can be predicted. In the case of imbalanced data, actual performance on skewed data cannot be correctly evaluated by this accuracy metric, so in addition to this the metrics like F1-Score, Precision, Recall, cross validation score, roc_auc curve values are considered.

Accuracy of the model is termed as the success rate of the model. The accuracy can be evaluated using the confusion matrix data[19].

Accuracy=(TP+TN)/(TP+ TN+FP+FN)

Precision of a model is defined as the rate of correctly classified samples. It is considered as the rate of true positives[19].
Precision=TP/(TP+FP)

Recall is treated as the measure of completeness of the results. It is also termed as the sensitivity of the model. It represents the rate of positive samples that were correctly classified.

Recall=TP/(TP+FN)

F1-Score is a measure obtained by combining precision and recall. This can be calculated as the harmonic mean between precision and recall[19].

F1 Score=2 * (Precision*Recall)/(Precision+Recall)
roc_auc score is a measure which discuss the capability of the model to distinguish between the classes. ROC is receiver operating characteristics is a probability curve and AUC area under curve represents the degree of separability[19].

Cross validation is an evaluation technique for models in machine learning by training the different models with subsets of convenient data and complementary subset is used for testing the model. K-fold cross validation is the popular method. Cross validation score is the average of the produced output, that is laboured by the amount of folds[20].

IV EXPERIMENTAL STUDY AND RESULTS

Pythons Jupyter Notebook is used to perform the study. The Fig.8 shows the diagrammatic representation of this experimental study. The different stages that are adopted in this study are,

Selecting the dataset.

Data pre-processing is performed on the selected dataset to remove the noise data.

Splitting the dataset into training dataset and testing dataset in the ratio 80:20.

Evaluate the performance of classifiers like k-NN, SVM and Random Forest.

Apply resampling techniques like ROS, RUS, SMOTE and NearMiss to the to the training dataset one by one.

After applying each resampling technique, it evaluates the performance of the various classifiers like k-NN, SVM and Random Forest.

Analyse the performance of these classifiers before applying resampling technique and after applying the resampling technique.

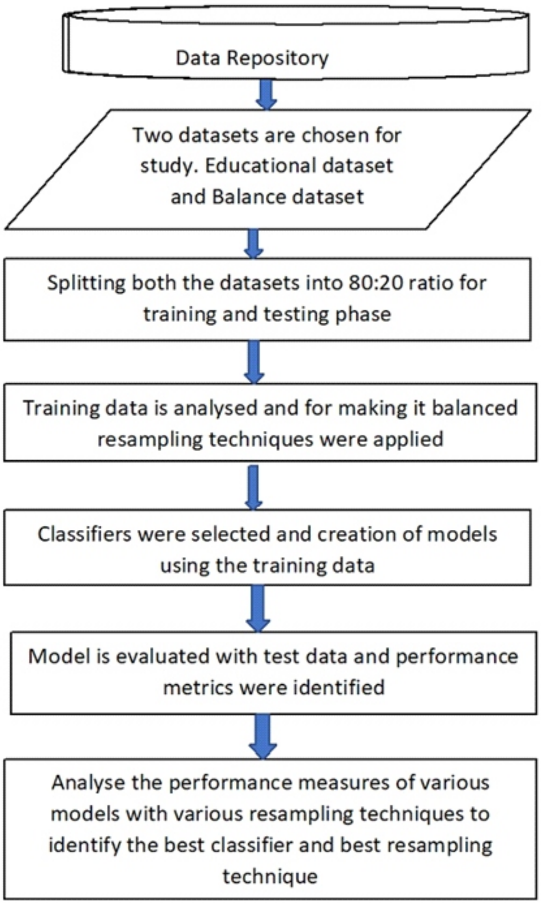
Data Repository
Two datasets are chosen for study. Educational dataset and Balance dataset
Splitting both the datasets into 80:20 ratio for training and testing phase
Training data is analysed and for making it balanced resampling techniques were applied
Classifiers were selected and creation of models using the training data
Model is evaluated with test data and performance metrics were identified
Analyse the performance measures of various models with various resampling techniques to identify the best classifier and best resampling technique

Fig.8 Steps in Experimental Study

For this experimental study the dataset used is a multiclass dataset. In this multiclass scenario the results will appear based on these multiple classes. For instance, the accuracy of prediction for all the classes will be displayed, likewise all the other measures. For the easiness of evaluation, we can take the average of results produced by all the classes. The average values can be obtained as macro average using the libraries of python[21].

For the original dataset the data distribution of the training set was in the range of sixty. In the case of resampling techniques, it tries to distribute approximately the same amount of data among the different classes of the dataset. Random under sampling technique and NearMiss under samples the different class data to the frequency of data in the minority class. Random oversampling and SMOTE techniques oversample all the class data to the frequency of the majority class. The Table.3 shows the data distribution of the training data among the various classes present in the dataset[22].

Dataset	Class Labels	Unbalanced dataset	After Resampling Techniques			
			RUS	ROS	SMOTE	NearMiss
Education Dataset[23,24,25]	High	83	38	104	104	38
	Middle	97	38	104	104	38
	Low	104	38	104	104	38
	Very Low	38	38	104	104	38
Balance Dataset[26,27,28]	Left	288	36	233	233	36
	Right	288	36	233	233	36
	Balance	49	36	233	233	36

Table.3 distribution of training data among various classes before and after the application of resampling techniques

Name of the resampling technique	Name of the classifier	Accuracy	Precision	Recall	F-Score
Unsampled	K-NN	0.83	0.88	0.78	0.8
	SVM	0.93	0.94	0.9	0.92
	Random Forest	0.91	0.9	0.88	0.9
RUS	K-NN	0.68	0.69	0.68	0.628
	SVM	0.696	0.727	0.73	0.655
	Random Forest	0.68	0.647	0.6	0.59
ROS	K-NN	0.656	0.689	0.66	0.61
	SVM	0.72	0.728	0.75	0.67
	Random Forest	0.66	0.64	0.63	0.6
SMOTE	K-NN	0.728	0.626	0.6	0.6
	SVM	**0.86**	**0.84**	**0.8479**	**0.84**
	Random Forest	0.82	0.59	0.61	0.6
NearMiss	K-NN	0.77	0.81	0.809	0.8
	SVM	0.85	0.84	0.847	0.84
	Random Forest	0.85	0.84	0.847	0.84

Table.4 shows the performance metrics of Education dataset with different resampling techniques and classifiers

The performance of the model can be evaluated using the confusion matrix. The evaluation metrics like precision, recall, F-score will be evaluated for these multiple classes. The average value for all these measures can also be identified by python. The detailed evaluation is shown in the following tables. The table Table.4 is showing the precision, recall and F-score of the Education dataset while applying the various classifiers on unsampled and resampled datasets.

From Table.4 we can analyse the various performance measures for the education dataset. For the unsampled dataset the dataset gives satisfactory measures for the dataset however we must consider the count of very low class. So, we need to balance the dataset. We applied several balancing techniques like RUS, Near Miss, ROS and SMOTE to balance the dataset. SMOTE resampling technique produces a better performance than other resampling techniques.

Name of the resampling technique	Name of the classifier	Accuracy	Precision	Recall	F-Score
Unsampled	K-NN	0.8	0.57	0.59	0.58
	SVM	0.89	0.59	0.66	0.63
	Random Forest	0.83	0.59	0.61	0.6
RUS	K-NN	0.696	0.715	0.735	0.696
	SVM	0.744	0.76	0.81	0.7
	Random Forest	0.728	0.65	0.62	0.621
ROS	K-NN	0.71	0.57	0.52	0.549
	SVM	**0.87**	**0.808**	**0.8853**	**0.815**
	Random Forest	0.816	0.591	0.607	0.599
SMOTE	K-NN	0.736	0.61	0.587	0.596
	SVM	0.848	0.79	0.867	0.79
	Random Forest	0.808	0.59	0.6	0.596
NearMiss	K-NN	0.67	0.698	0.677	0.625
	SVM	0.776	0.736	0.77	0.71
	Random Forest	0.672	0.62	0.57	0.58

Table.5 shows the performance metrics of Balance dataset with various resampling techniques and classifiers.

The table Table.5 is showing the precision, recall and F-score of the balance dataset while applying the various classifiers with different resampling techniques. Balance dataset is also a multimajority class dataset. The resampling techniques like ROS, SMOTE works well with the Balance dataset. As the dataset is having multiple majority classes it is better to oversample the minority class to balance the dataset. The classifier SVM shows good performance over the ROS resampled data.

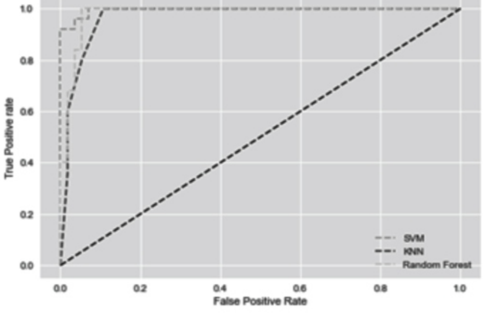

Fig.9 ROC curve for Smote sampled data of Educational Dataset

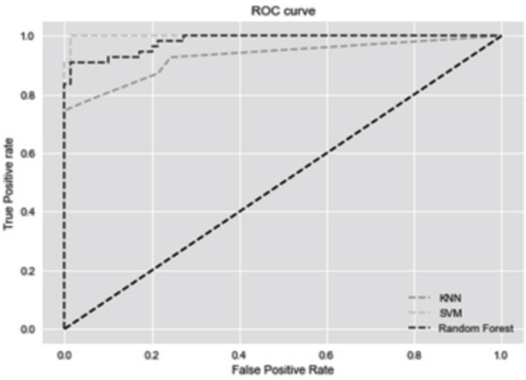

Fig.10 ROC curve for ROS resampled data of Balance Dataset

The receiver operating characteristics curve for Smote resampled dataset and ROS resampled dataset of educational dataset and balance dataset are plotted in Fig.9 and Fig.10. Expect KNN classifier all others shown comparatively good positive rate in the case of SMOTE resampled data of educational dataset. ROS resampled Balance dataset SVM classifier shown the highest positivity rate than others. From these figures it is evident that oversampled data will give more performance than the under sampled data.

In this work, primary importance is given for resampling techniques and secondary importance is given for the classifiers. For imbalanced data in learning scenario, it is better to apply oversampling technique to balance the data than applying under sampling techniques. Under sampling techniques will sometimes leads to vital data loss and this will affect the performance of the model.

V CONCLUSION

Data of most of the real-world are imbalanced for some classes. For making the data balanced it is better to use resampling techniques on the data during the pre-processing stage. In this work, it is identified that oversampling techniques like SMOTE and ROS produces better result than the under sampled data. It is desirable to use oversampling techniques on skewed data for making it balanced. Under sampling always leads to loss of data which may be somewhat important. This study also shown that the classifiers performance improved after applying the resampling techniques. Oversampling techniques can be used for data in all sectors of life to make it balanced. As we are dealing with multimajority dataset which means multiple majority classes are already there, so that we must oversample the minority ones to balance the data. In the case of multimajority datasets oversampling techniques produces promising results. As a future work we can extend this study to datasets with different imbalance ratio and the the effect of different resampling techniques over them. The performance can also be analysed for these datasets with different classifiers.

REFERENCES

[1] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” Pattern Recognit., vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[2] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” Prog. Artif. Intell., vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[3] S. Wang and X. Yao, “Multiclass imbalance problems: Analysis and potential solutions,” IEEE Trans. Syst. Man, Cybern. Part B Cybern., vol. 42, no. 4, pp. 1119–1130, 2012, doi: 10.1109/TSMCB.2012.2187280.

[4] R. Singh and R. Raut, “Review on Class Imbalance Learning: Binary and Multiclass,” Int. J. Comput. Appl., vol. 131, no. 16, pp. 4–8, 2015, doi: 10.5120/ijca 2015907573.

[5] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, “Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches,” Knowledge-Based Syst., vol. 42, pp. 97–110, 2013, doi: 10.1016/j.knosys. 2013.01.018.

[6] V. S. Spelmen and R. Porkodi, “A Review on Handling Imbalanced Data,” Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018, no. December, pp. 1–11, 2018, doi: 10.1109/ICCTCT. 2018.8551020.

[7] Y. Pristyanto, I. Pratama, and A. F. Nugraha, “Data level approach for imbalanced class handling on educational data mining multiclass classification,” in 2018 International Conference on Information and Communications Technology, ICOIACT 2018, 2018, vol. 2018-Janua, doi: 10.1109/ICOIACT.2018.8350792.

[8] B. S. Raghuwanshi and S. Shukla, “Class imbalance learning using UnderBagging based kernelized extreme learning machine,” Neurocomputing, vol. 329, pp. 172–187, Feb. 2019, doi: 10.1016/j.neucom. 2018.10.056.

[9] G. Kovács, “Smote-variants: A python implementation of 85 minority oversampling techniques,” Neurocomputing, vol. 366, no. June, pp. 352–354, 2019, doi: 10.1016/j.neucom.2019.06.100.

[10] M. Ashraf, M. Zaman, and M. Ahmed, “An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 1471–1483, 2020, doi: 10.1016/j.procs.2020.03.358.

[11] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, “Hybrid resampling to handle imbalanced class on classification

of student performance in classroom," Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017, vol. 2018-Janua, pp. 207–212, 2017, doi: 10.1109/ICICOS. 2017.8276363.

[12] E. A. Yekun and A. T. Haile, "Student Performance Prediction with Optimum Multilabel Ensemble Model," J. Intell. Syst., vol. 30, no. 1, pp. 511–523, 2021, doi: 10.1515/jisys-2021-0016.

[13] M. Imran, M. Afroze, S. Kumar Sanampudi, and A. Abdul Moiz Qyser, "Data Mining of Imbalanced Dataset in Educational Data Using Weka Tool," Int. J. Eng. Sci. Comput. IJESC, vol. 6, no. 6, pp. 7666–7669, 2016, doi: 10.4010/2016.1809.

[14] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access, vol. 4, pp. 2379–2387, 2016, doi: 10.1109/ACCESS. 2016.2568756.

[15] X. Li, S. Wu, X. Li, H. Yuan, and D. Zhao, "Particle Swarm Optimization-Support Vector Machine Model for Machinery Fault Diagnoses in High-Voltage Circuit Breakers," J. Mech. Eng, vol. 33, p. 6, 2020, doi: 10.1186/s10033-019-0428-5.

[16] V. T. N. Chau and N. H. Phung, "Imbalanced educational data classification: An effective approach with resampling and random forest," Proc. - 2013 RIVF Int. Conf. Comput. Commun. Technol. Res. Innov. Vis. Futur. RIVF 2013, no. January, pp. 135–140, 2013, doi: 10.1109/RIVF.2013.6719882.

[17] Y. Pristyanto, A. F. Nugraha, I. Pratama, and A. Dahlan, "Ensemble Model Approach for Imbalanced Class Handling on Dataset," 2020 3rd Int. Conf. Inf. Commun. Technol. ICOIACT 2020, pp. 17–21, 2020, doi: 10.1109/ICOIACT50329.2020.9331984.

[18] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[19] E. Mortaz, "Imbalance accuracy metric for model selection in multi-class imbalance classification problems," Knowledge-Based Syst., vol. 210, Dec. 2020, doi: 10.1016/j.knosys.2020.106490.

[20] A. Maxwell et al., "Deep learning architectures for multi-label classification of intelligent health risk prediction," BMC Bioinformatics, vol. 18, no. Suppl 14, 2017, doi: 10.1186/s12859-017-1898-z.

[21] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," J. Mach. Learn. Res., vol. 18, no. September, pp. 1–5, 2017.

[22] Nitesh V. Chawla et al. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321–357.

[23] Haibo He et al. "Adaptive Synthetic Sam-pling Approach for Imbalanced Learning Neural Networks - ADASYN", IJCNN 2008.1-8

[24] https://www.kaggle.com/rafjaa/ resampling- strategies-for-imbalanced-datasets.

[25]https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8

[26]https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm [27] https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

[27]https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

[28]R. Mary Mathew and R. Gunasundari, "A Review on Handling Multiclass Imbalanced Data Classification In Education Domain," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 752-755, doi: 10.1109/ICACITE51222.2021.9404626.