Grenze International Journal of Engineering and Technology, June Issue

Machine Learning Approach for Handling Imbalanced Students' Performance Data

E. Sujatha[1], S.Divya[2], R.G.Sakthivelan[3] and Gopirajan PV[4]
[1]Department of Computer Science and Engineering, Saveetha Engineering College (Autonomous), Thandalam, Chennai.
Email: sanjaymohankumar@gmail.com, https://orcid.org/0000-0003-2500-2937
[2]Department of Artificial Intelligence and Machine Learning, Rajalakshmi Engineering College (Autonomous), Chennai.
Email: divya.s@rajalakshmi.edu.in
[3]Department of Computer Science and Engineering, Gopalan College of Engineering and Management, Bengaluru,
Email: sakthi21245@gmail.com, https://orcid.org/0000-0002-3444-9821
[4]Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Kattankulathur campus, Chennai, India
Email: gopivrajan@gmail.com

***Abstract*— Imbalanced student performance data in educational institutions is crucial for any machine learning prediction model. It affects the efficiency of classifiers and challenges the sampling methods for having a more significant number of features. The proposed model was designed to predict the student's performance by comparing the results with popular similar models. The Kaggle dataset having 386 rows and 33 columns of student data, was used in this study. In addition, the Portugal database was considered for experimental analysis, containing 394 students and 19 attributes. It proves that the Random Forest classifier yields the highest accuracy at 84.15% in handling imbalanced datasets. The results show that the SVM-SMOTE is higher accuracy as, 94.76%, than the other sampling methods in predicting the student's performance with various features.**

***Index Terms*— Machine learning, Imbalanced data, Prediction, Student Performance**

I. INTRODUCTION

Many higher education institutions have Management Information Systems to maintain and monitor students' academic performance. But the challenge lies in developing a prediction model in imbalanced data. The crucial thing is that dataset has many features which are not significant for prediction. Many researchers have suggested many prediction models for different classifiers and sampling methods.

Solomon et al. discussed the challenges faced in higher educational institutions [1]. Bujang et al. achieved 99% accuracy using SMOTE to predict student grades[2]. Migues et al. and P. M. Moreno-Marcos et al. have illustrated external factors such as socioeconomic demographics which are influencing the learning behavior of learner's performance [3], [4]. A. E. Tatar et al. demonstrated the prediction model for academic performance at the UG level with grades [5].

The significance of the predictive model was discussed by K. L. M. Ang et al. for academic institutions that are in high demand to produce high accuracy [6]. A. Hellas et al. discussed the possible Machine learning techniques used for predicting the learner's performance but uncovered few issues which are hindering learners' performance[7]. L. M. Abu Zohair et al. have improved the quality of the student's performance in imbalanced

Grenze ID: 01.GIJET.10.2.163

data [8]. X. Zhang et al. discussed multi-classification models for predicting student performance [9].
SMOTE method was used by D. Berrar et al. to improve the accuracy in the prediction model [10]. A. Hellas et al. and N. V. Chawla et al., have implemented a resampling data method using Cross-validation to predict errors in the performance models and enhances the quality of the parameters [11], [12]. Ashfaq, U. et al. and L. E. O. Breiman et al. implemented Random Forest classifier which focuses on the strength of the individual tree and the correlation between them [13], [14].
C.Jalota et al., and Buenaño-Fernández et al. have found that academic performance results with numerous numbers of features from various Feature selection algorithms and classifiers will help the researcher to discover the most excellent mixture of filter feature selection algorithms and its associated classifiers [15], [16].
S.Chinna Gopi et al. proved that SVM exhibits better generalization in class imbalanced problems. It is less sensitive when compared to other classification algorithms. SVM was analyzed with feature selection methods [17].
P.Nair et al. and S. Zhang et al. applied KNN for regression, classification, and missing data imputation. Different datasets have been tested and compared on standard KNN, SMOTE with KNN, and SRAS with KNN. Preprocessing model SRAS produces superior and enhanced outcomes for the KNN classifier in terms of accuracy, recall, F Measure, and Area Under the Curve [18], [19].
Hanyuan Hang et al. experimented with the expected sub-sample size can be much smaller than the number of training data at each bagging round, and the number of nearest neighbors can be reduced simultaneously, especially when the data are highly imbalanced, which leads to substantially lower time complexity and roughly the same space complexity [20].
Benjamin Lu et al. suggested that well-adapted for prediction interval estimation and simulations proves prediction intervals are competitive with, and in some settings outperform existing methods [21].
Ashfaq, U. et al. proved that the ensembling method is adopted to obtain better accuracy compared to other methods with various combinations of classifiers and sampling techniques [22].

II. Proposed System Methodology

This proposed methodology includes the essential metrics for evaluating imbalanced classification models, where one class significantly outnumbers the other.
Following Key Metrics were considered for performance calculation:

- *Confusion Matrix*: Visualizes true positives, false positives, true negatives, and false negatives, offering insights into specific prediction errors.
- *Precision*: Measures the fraction of predicted positive cases that are truly positive.
- *Recall*: Measures the proportion of actual positive cases that are correctly identified.
- *F1-Score*: Balances precision and recall into a single score, useful for imbalanced data.
- *ROC AUC*: Measures the model's ability to distinguish between classes using the ROC curve. Can be optimistic for highly imbalanced data.
- *PR AUC*: Measures the model's performance on the minority class using the precision-recall curve. A robust metric for imbalanced classification.

III. Results And Discussion

The proposed system attempts to examine different sampling methods with various classifiers. There are two datasets used for analysis. The Portugal dataset has information about two secondary education Portuguese schools. It contains 394 students with 19 attributes, and the Kaggle dataset contains 396 students with 33 attributes. Four categories are available in the dataset: Poor, Medium, Good, and Excellent. Hence, the multi-classification problem occurs. Many datasets suffer from the imbalanced problem. In an imbalanced situation, the majority class dominates the minority class. In the Portugal dataset, the Poor class samples are 15%, the medium course is 44%, the good course is 35%, and the Excellent class is 6%.
This project uses shuffle 5-fold cross-validation for testing and training the dataset. The dataset is divided into five segments one segment is used for testing, and the other four segments are used for training the dataset each time. It is termed the hold-out method and is repeated five times. It is a more robust and trustworthy method for validation. Hence it improves the accuracy of the proposed predicting model. Cross-validation model applied for all sampling methods.
The results shown are based on the 2 different datasets. The distribution of average grades of students is shown in the figure B1 and the average lies between 7.5 and 10.00.

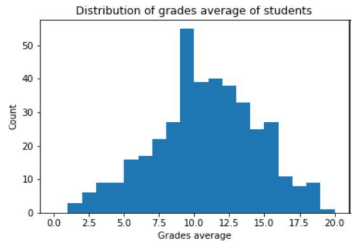

Figure 1. Histogram distribution of average grade of students

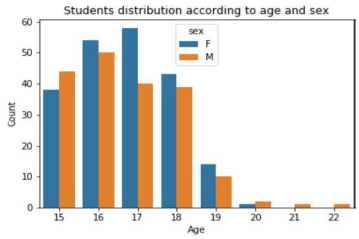

Figure 2. Distribution of students according to age and sex.

Figure 1 shows the frequency distribution of students' average grades on the horizontal axis, which ranges from 0.0 to 20.0, and the count of students on the vertical axis.

Figure 2 illustrates the distribution of students by age and further categorizes them by sex, with 'F' representing female students and 'M' representing male students.

The horizontal axis of the chart represents the age of the students, ranging from 15 to 22 years. The vertical axis represents the count of students within each age group. Each age group has two bars adjacent to each other, one for female students (in blue) and one for male students (in orange), allowing for a direct comparison between the number of male and female students within each age category.

Overall, the figure 2 shows that the distribution of students by age and sex varies, with a general trend of more male students in the middle age groups (16-17 years) and a decline in student counts as age increases, particularly after age 18. This could suggest that most students are concentrated in the typical high school age range, with fewer students continuing at older ages, which might be indicative of graduation or other factors leading to a decrease in the student population.

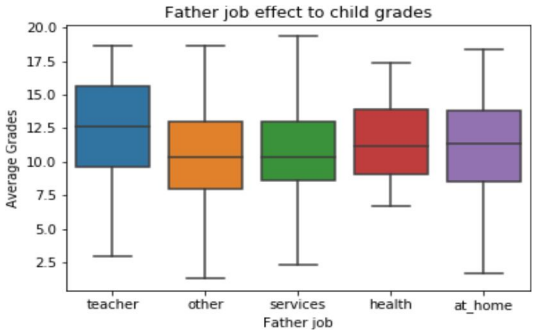

Figure 3. Effects of father's job on student's grade. Error bars indicate the standard deviation.

Figure 3 compares the average grades of children based on their fathers' occupations. The occupations are categorized as "teacher," "other," "services," "health," and "at home." Each box plot represents the distribution of average grades for each category of father's job. The bottom and top of each box indicate the first (Q1) and third (Q3) quartiles, respectively, and the band inside the box is the median (Q2). The "whiskers" extending from the boxes represent the range of the data, excluding outliers.

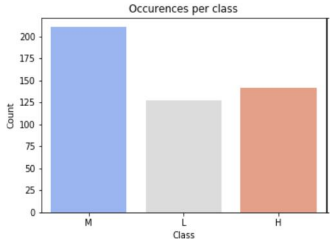

Figure 4: Students per grade interval each semester

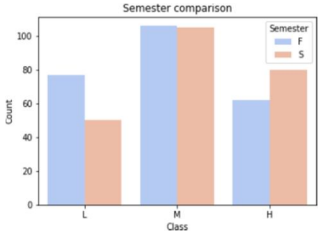

Figure 5: Comparison of student grades each semester

Students' performance in occurrences of class per grade interval such Medium, Low and High is shown in figure 4.

Comparison of the student grades in first and second semester shows that medium category has higher count than other categories.

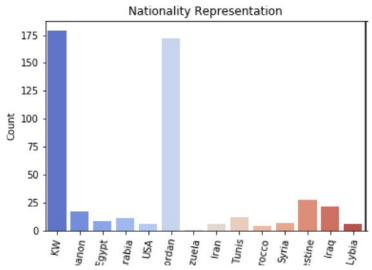

Figure 6: Impact the nationality of comparison based on gender

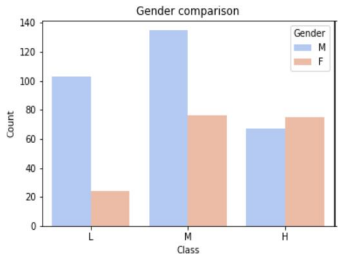

Figure 7: Gender distribution and grade the students.

Nationality of the students also have impact on the performance of the students. Both male and female have obtained different grades and their distributions are compared.

The significant of the features are ranked in Random Forest classifier with respect to different attributes available in the dataset is visualized in Figure 10.

TABLE I. PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS

Classifiers	**5-fold**	**PR-AUC**	**ROC-AUC**	**Precision**	**Recall**	**F1**	**Accuracy**
Linear	55.48	0.76	0.81	0.56	0.51	0.43	59.50
Random Forest	70.95	0.85	0.83	0.79	0.68	0.69	84.15
KNN	61.90	0.61	0.79	0.65	0.53	0.54	74.24
Neural Network	54.87	0.72	0.72	0.69	0.57	0.49	65.72
XG Boost	63.47	0.64	0.75	0.63	0.59	0.60	82.54
SVM	58.98	0.74	0.69	0.78	0.58	0.55	80.05
Decision Tree	56.76	0.78	0.68	0.68	0.61	0.51	67.89

Sampling methods Borderline SMOTE, SMOTE, SVM-SMOTE, SMOTE-ENN and SMOTE-Tomek are examined with performance metrics and proves that SVM-SMOTE provides higher accuracy is shown in Table II. This table II showcases the impact of various sampling methods on imbalanced student performance prediction using an SVM classifier. 5-fold cross-validation ensures generalizability. SVM-SMOTE stands out significantly with the highest accuracy (94.76%) and strong performance across other metrics. This suggests it effectively addresses the imbalanced data problem. SMOTE and SMOTE-ENN shows moderate improvement over base performance (without sampling) but fall short of SVM-SMOTE. Borderline SMOTE and SMOTE-Tomek method offers slight improvements compared to the base model but remain less effective than SVM-SMOTE.

SVM-SMOTE's high accuracy and precision-recall balance indicate its ability to accurately identify true high performers/passes even in the imbalanced dataset. SMOTE and SMOTE-ENN provide some improvement, suggesting their potential for imbalanced classification. Borderline SMOTE and SMOTE-Tomek seem less effective in this specific context.

TABLE II. PERFORMANCE ANALYSIS OF SAMPLING METHODS

Sampling Methods	5-fold	PR-AUC	ROC-AUC	Precision	Recall	F1	Accuracy
Borderline SMOTE	59.79	0.65	0.71	0.70	0.62	0.54	79.63
SMOTE	57.84	0.71	0.80	0.71	0.64	0.58	80.10
SVM-SMOTE	74.93	0.89	0.87	0.86	0.81	0.71	94.76
SMOTE-ENN	62.90	0.80	0.77	0.76	0.66	0.56	81.43
SMOTE-Tomek	63.15	0.84	0.74	0.67	0.64	0.59	78.67

The Table III discusses about the different authors have proved in their research that different classification algorithms have produced various results.

This table III summarizes the performance of different machine learning models used in various studies for predicting student performance. Several interesting trends emerge:

TABLE III. COMPARISON FOR RESULTS PERFORMANCE OF VARIOUS AUTHORS AND THEIR CLASSIFIERS

Sl. No.	Authors	Classifiers	Results %
1	E.Sujatha et. al., (2024) Proposed Method	Linear	59.50
		Decision Tree	67.89
		KNN	74.24
		Neural Network	65.72
		XG Boost	82.54
		SVM	80.05
		Random Forest	84.15
2	Awujoola J. et al., (2021)	Logistic Regression	91
		Decision Tree	88
		Random Forest	91 and 93
		Naïve Bayes	88
		Support Vector Machine	93
3	Ramin Ghorbani et al., (2020)	Naive Bayes	47.72
		Artificial Neural Network	67.97
		XG-Boost	69.24
		Support Vector Machine	69.11
		k-Nearest Neighbor	75.56
		Logistic Regression	69.24
		Decision Tree	56.58
		Random Forest	76.83
4	Kian Lam et al., (2019)	Logistic Regression	58.50
		Decision Tree	58.84
		Naïve Bayes	57.66
		k-Nearest Neighbor	57.44
		Support Vector Machine	59.01
		Neural Network	59.04
5	Casuat C.D. et al., (2019)	Random Forest	84
		Decision Tree	85
		Support Vector Machine	91.22
6	Pothuganti M. et al., (2019)	Decision Tree	84
		Random Forest	86
7	Othman et al., (2018)	Support Vector Machines	66.09
		Artificial Neural Networks	65.29
		Decision Tree	66.06
8	Hugo L. (2018)	Logistic Regression	72.17
		Discriminant Analysis	69.81
		Decision Tree	71.70
		Artificial Neural Network	73.11
		Support Vector Machine	87.26

IV. CONCLUSION AND FUTURE WORK

The challenge for today's scenario is to analyse and predict the student's performance in educational institutions and handling imbalanced dataset. The proposed model uses machine learning techniques and experiments with the metrics such as accuracy, precision, recall, F1, ROC-AUC, and PR-AUC.

The proposed method establishes the prediction model for student's performance in handling imbalanced dataset with different classifiers and sampling methods. It attempts to prove best classifier among different classification

methods. Classifiers such as Linear, Random Forest, k-Nearest Neighbour, Neural Networks, XG Boost, Support Vector Machine and Decision Trees are analysed with the sampling methods such as Borderline SMOTE, SMOTE, SVM-SMOTE, SMOTE-ENN and SMOTE-Tomek in imbalanced dataset.

In proposed method Random Forest proves that it achieves higher accuracy with 84.15% and SVM-SMOTE accuracy is 94.76%.

The results obtained shows that different evaluation metrics provides better performance with machine learning models. Hence, the SVM-SMOTE outperforms well among all other methods in predicting student's performance in imbalanced Portugal dataset and Kaggle dataset. Different classifiers are applied on validation techniques on different imbalanced datasets.

Future work is suggested to implement hybrid classifiers with large scale multiple datasets. In order to make more efficient early prediction, more features (attributes) would be considered. To enhance the performance of the prediction model, feature extraction techniques would be adapted. More semester's data and significant features would be added for better accuracy.

AVAILABILITY OF DATA AND MATERIAL - The data that support the findings of this study are openly available in Kaggle dataset at https://www.kaggle.com/datasets/ishandutta/student-performance-data-set, https://www.kaggle.com/datasets/dipam7/student-grade-prediction and available in UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/student%2Bperformance. The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

REFERENCES

[1] D. Solomon, "Predicting Performance and Potential Difficulties of University Student using Classification: Survey Paper," 2018.

[2] S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," IEEE Access, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[3] P. M. Moreno-Marcos, T. C. Pong, P. J. Munoz-Merino, and C. D. Kloos, "Analysis of the Factors Influencing Learners' Performance Prediction with Learning Analytics," IEEE Access, vol. 8, pp. 5264–5282, 2020, doi: 10.1109/ACCESS.2019.2963503.

[4] V. L. Miguéis, A. Freitas, P. J. v Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," Decis Support Syst, vol. 115, pp. 36–51, 2018, doi: 10.1016/j.dss.2018.09.001.

[5] A. E. Tatar and D. Düştegör, "Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?," Applied Sciences, vol. 10, no. 14, 2020, doi: 10.3390/app10144967.

[6] K. L.-M. Ang, F. Ge, and K. P. Seng, "Big Educational Data & Analytics: Survey, Architecture and Challenges," IEEE Access, vol. 8, pp. 116392–116414, 2020.

[7] A. Hellas et al., "Predicting Academic Performance: A Systematic Literature Review," in Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, in ITiCSE 2018 Companion. New York, NY, USA: Association for Computing Machinery, 2018, pp. 175–199. doi: 10.1145/3293881.3295783.

[8] L. Mahmoud Abu Zohair, "Prediction of Student's performance by modelling small dataset size," vol. 16, p. 18, Dec. 2019, doi: 10.1186/s41239-019-0160-3.

[9] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, "Grade Prediction of Student Academic Performance with Multiple Classification Models," in 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2018, pp. 1086–1090. doi: 10.1109/FSKD.2018.8687286.

[10] D. Berrar, "Cross-Validation," Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, vol. 1–3, pp. 542–545, Jan. 2019, doi: 10.1016/B978-0-12-809633-8.20349-X.

[11] A. Hellas et al., "Predicting academic performance: a systematic literature review," in Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, New York, NY, USA: ACM, Jul. 2018, pp. 175–199. doi: 10.1145/3293881.3295783.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[13] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[14] U. Ashfaq, Dr. B. P. M., and R. Mafas, "Managing Student Performance: A Predictive Analytics using Imbalanced Data," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 2277–2283, Mar. 2020, doi: 10.35940/ijrte.E7008.038620.

[15] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, "Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study," Sustainability, vol. 11, no. 10, p. 2833, May 2019, doi: 10.3390/su11102833.

[16] C. Jalota and R. Agrawal, "Feature Selection Algorithms and Student Academic Performance: A Study," 2021, pp. 317–328. doi: 10.1007/978-981-15-5113-0_23.

[17] S. Chinna Gopi, B. Suvarna, and T. Maruthi Padmaja, “High Dimensional Unbalanced Data Classification Vs SVM Feature Selection,” Indian J Sci Technol, vol. 9, no. 30, Aug. 2016, doi: 10.17485/ijst/2016/v9i30/98729.
[18] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” ACM Trans Intell Syst Technol, vol. 8, no. 3, pp. 1–19, May 2017, doi: 10.1145/2990508.
[19] P. Nair and I. Kashyap, “Optimization of kNN Classifier Using Hybrid Preprocessing Model for Handling Imbalanced Data,” 2019.
[20] H. Hang, Y. Cai, H. Yang, and Z. Lin, “Under-bagging Nearest Neighbors for Imbalanced Classification,” Journal of machine learning research, vol. 23, pp. 1–63, Apr. 2022.
[21] B. Lu and J. Hardin, “A Unified Framework for Random Forest Prediction Error Estimation,” J. Mach. Learn. Res., vol. 22, no. 1, Jan. 2021.
[22] U. Ashfaq, Dr. B. P. M., and R. Mafas, “Managing Student Performance: A Predictive Analytics using Imbalanced Data,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, pp. 2277–2283, Mar. 2020, doi: 10.35940/ijrte.E7008.038620..