INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

The Effects of Imbalanced Datasets on Machine Learning Algorithms in Predicting Student Performance

Khaled Mahmud Sujon [a], Rohayanti Hassan [a,*], Alif Ridzuan Khairudin [a], Sim Hiew Moi [a], Muhammad Luqman Mohd Shafie [a], Zainuri Saringat [b], Aldo Erianda [c]

[a] *Software Engineering Research Group, Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor, Malaysia*
[b] *Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Malaysia*
[c] *Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia*

*Corresponding author: *rohayanti@utm.my*

***Abstract*— Predictive analytics technologies are becoming increasingly popular in higher education institutions. Students' grades are one of the most critical performance indicators educators can use to predict their academic achievement. Academics have developed numerous techniques and machine-learning approaches for predicting student grades over the last several decades. Although much work has been done, a practical model is still lacking, mainly when dealing with imbalanced datasets. This study examines the impact of imbalanced datasets on machine learning models' accuracy and reliability in predicting student performance. This study compares the performance of two popular machine learning algorithms, Logistic Regression and Random Forest, in predicting student grades. Secondly, the study examines the impact of imbalanced datasets on these algorithms' performance metrics and generalization capabilities. Results indicate that the Random Forest (RF) algorithm, with an accuracy of 98%, outperforms Logistic Regression (LR), which achieved 91% accuracy. Furthermore, the performance of both models is significantly impacted by imbalanced datasets. In particular, LR struggles to accurately predict minor classes, while RF also faces difficulties, though to a lesser extent. Addressing class imbalance is crucial, notably affecting model bias and prediction accuracy. This is especially important for higher education institutes aiming to enhance the accuracy of student grade predictions, emphasizing the need for balanced datasets to achieve robust predictive models.**

***Keywords*—Imbalanced dataset; machine learning; higher education institute; multi-class prediction.**

Manuscript received 11 Mar. 2024; revised 17 Jun. 2024; accepted 22 Sep. 2024. Date of publication 30 Nov. 2024.

I. INTRODUCTION

Higher education institutions are now prioritizing the use of predictive analytics applications. Predictive analytics incorporates advanced analytics, including machine learning implementation, to extract valuable data and achieve high-quality performance across all educational levels. A teacher's grade is one of the most critical performance indicators that can be used to monitor their students' academic progress [1].

Students' performance fluctuates throughout the year, and because of a decrease in overall performance caused by a variety of circumstances, eventually, the percentage of students who fail increases [2]. Data mining techniques make it possible to predict students' failures in the courses at an early stage. Consequently, we recognize that predicting student grades could be an effective strategy for raising student academic achievement. All higher education institutions have grade-keeping systems and data management systems through which they collect information about their students. The educators would be able to anticipate the results of their students early based on previous results, which would be a revolutionary system. They can make many decisions to improve their students' performance levels. The use of predictive analytics for predicting students' academic achievement has increased over time [5]. As a result, machine learning techniques offer many options for building predictive models based on historical data. The presence of a class imbalance in datasets, in which one class significantly outnumbers the others, poses a severe challenge to the effectiveness of these machine-learning models. Studies have demonstrated that the popular machine learning algorithms Random Forest (RF) and Logistic Regression (LR) can be used more effectively to predict student grades, thereby promoting academic success in students while addressing the

challenges associated with managing imbalanced datasets. Besides predicting student grades accurately, we explore the potential impact of imbalanced datasets on machine learning models. Our analysis will be based on Logistic Regression (LR) and Random Forest (RF). Our study aims to investigate the implications of data-driven decision-making for educational strategies and improve academic performance.

This study investigates how imbalanced datasets affect machine learning models' ability to predict student performance. The problem statement focuses on the inherent bias and errors that occur when models are trained on unbalanced data, which impacts predictions and underrepresents underrepresented classes. Machine learning algorithms are evaluated in different class imbalance scenarios, techniques are identified to reduce the impact of imbalance, and the importance of balanced datasets in educational predictive modeling is emphasized. This research can improve predictive analytics in education because it can improve efficiency and equity. Educators, policymakers, and machine learning practitioners must understand how class imbalance impacts model performance to make well-informed decisions on feature engineering, data preprocessing approaches, and model selection.

This study utilized several crucial techniques to improve the machine learning model's performance. These include the encoding of categorical features, the normalization of data, etc. Several performance measurement metrics evaluate the model's performance, including accuracy, precision, recall, and the F-1 score. Furthermore, the confusion matrix provides a detailed breakdown of the model's performance, handy when dealing with imbalanced datasets. Throughout the rest of the paper, three sections are presented: Section 2 summarizes the existing work in the same domain, Section 3 presents a detailed description of the research methodology, and Section 4 summarizes the experiment's findings. Lastly, Section 5 concludes the paper.

II. Materials and Method

A. Background of Student Grade Prediction

As institutions adopt more data-driven methods to improve academic outcomes, the literature review on student grade prediction and the use of predictive analytics in the education industry is growing rapidly. Several studies have examined the application of machine learning algorithms in higher education institutions. In this section, we summarize existing research, emphasizing the evolution of predictive models, their impact on decision-making in higher education institutions (HEIs), and the most used algorithms and datasets.

According to Jishan et al. [6], the final grade is based on a student's CGPA, quiz, midterm, lab, and attendance. Based on Naive Bayes, Decision Trees, and Neural Network Backpropagation with oversampling, the optimized Naive Bayes model achieves an accuracy of 75.28%. This approach can increase accuracy by resolving imbalances in the datasets and resolving their effects. In higher education institutions, the accuracy of academic performance prediction has increased since the data have been balanced, and the imbalanced data significantly impact the model's performance. Another study by Khan et al. [7] evaluated characteristics such as Test1_marks, CGPA, attendance, major, gender, and year for a higher education institution (HEI). Several machine learning methods were employed, including Naive Bayes, Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Lazy (IBK), and rules-based approaches such as Decision Tree (DT) and Random Forest (RF).

Based on the imbalanced datasets, the study determined that Decision Tree (J48) performed the best, with an accuracy of 88%, using feature selection and the Synthetic Minority Oversampling Technique (SMOTE). The results of this study support the effectiveness of various pre-processing strategies in enhancing predictive models for HEI variables.

Barrak et al. [8] analysis includes student information such as name, I.D., final GPA, graduation semester, major, nationality, campus, courses, and grades. In this study, the Decision Tree algorithm, namely J48, is used to determine which algorithm performs the best on the provided dataset. According to this study's findings, Decision Trees (J48) effectively capture complex interactions between variables and indicate their suitability for predictive modeling in our educational institution. A study by Mustafa Agaoglu [9] used several classification algorithms to predict instructors' performance. These algorithms include decision trees, support machine vectors, artificial neural networks, and discriminant analysis. The C5.0 classification demonstrated the highest and best accuracy among all models. After the students completed the questionnaires, the instructor's performance was assessed.

The study by Ismail et al. [10] examines the relationship between school details, student demographics, family factors, and academic performance. A variety of machine learning algorithms are employed by the authors, including Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), and Random Forests (RF), to analyze and predict academic outcomes. RF suggests an ensemble learning approach, leveraging multiple decision trees to improve prediction accuracy. The study will likely provide insight into the intricate interplay between diverse factors influencing academic success, utilizing machine learning techniques for comprehensive analysis and prediction in the educational field. Flanagan et al. [11] investigate the Undergraduate Introduction to Informatics course using data from 233 students, featuring an open-book assessment and the Book Roll digital learning system. The authors use Support Vector Machine (SVM) methods to forecast early warnings while focusing on evaluation metrics like Area Under Curve (AUC). Precision, recall, F1 score, accuracy, and AUC are the evaluation metrics for both baseline and optimized models. The proposed approach's correctness and effectiveness are evaluated using Student's t-tests. The study aims to improve predictive models for early detection of at-risk pupils, thus contributing to the field of education. Polyzou et al. [12] The research employs Linear Regression (LinReg) and Matrix Factorization (MF) approaches on its datasets. Linear regression is used to model the connection between variables in the datasets. The study will most likely examine previous student course grade data to get insight into predictive modeling and trends, with regression techniques utilized to analyze and forecast the complex dynamics of student performance over time.

According to Flanagan et al. [11], an open-book assessment and the Book Roll digital learning system were used to study 233 students enrolled in the Undergraduate Introduction to Informatics course. In their study, the authors employ Support Vector Machine (SVM) methods to predict early warnings while they focus on evaluation metrics such as Area Under Curve (AUC). The evaluation metrics for baseline and optimized models include precision, recall, F1 score, accuracy, and AUC. Student's t-tests are used to evaluate the correctness and effectiveness of the proposed approach. The study aims to improve predictive models for early detection of at-risk pupils, thus contributing to the field of education. Using Linear Regression (LinReg) and Matrix Factorization (MF) approaches, Polyzou et al. [12] examined their datasets using Linear Regression (LinReg) and Matrix Factorization (MF). Linear regression is used to model the relationships between variables in a dataset. This study will examine previous student course grade data to gain insight into predictive modeling and trends. Regression techniques are employed to analyze and forecast the dynamics of student performance over time. It has been shown that Higher Education Institutions (HEI) use characteristics such as Research Method (R.M.) grade, Research Project (R.P.) grade, gender, backlog, and programming proficiency to predict student grades, as demonstrated by Abana's [15] study. Random Tree (R.T.) performed the best, with a 75.2% accuracy rate, alongside machine learning approaches such as RepTree and Decision Tree (J48). Thus, it emphasizes Random Tree's ability to predict academic grades based on the variables contained in the datasets. It also emphasizes its application to higher education institutions when assessing academic grades and related qualities.

In their study, Khan et al. [16] evaluated the performance of various machine learning algorithms for predicting student success based on multiple characteristics, such as scores on the first test, major, gender, year, and CGPA. There were four algorithms used in this study: decision trees (J48), random forests (RF), simple classification and regression trees (SimpleCART), naive Bayes (NB), multilayer perceptron's (MLP), support vector machines (SVM), instance-based k-nearest neighbor (IBK) and rules-based approaches such as decision tables, JRIP, OneR, PART, and ZeroR. Using feature selection and synthetic minority oversampling technique (SMOTE) for data balancing, the decision tree algorithm (J48) showed the highest accuracy of 88%. According to Wakelam et al. [17], quiz scores, classroom environment, lecture attendance, and intermediate evaluations were all associated with student success. The algorithms used in their analysis were random forest (RF), k-nearest neighbors (KNN), and decision tree (DT). Random forest algorithm (RF) was able to predict student performance based on the variables mentioned above, with a 75% accuracy rate.

Pristyanto et al. [18] conducted a study based on information from the data structure course at the University of Indonesia. A combination of naive Bayes (NB), support vector machine (SVM), and k-nearest neighbors (KNN) algorithms was used in conjunction with oversampling, followed by undersampling (OSS) and synthetic minority oversampling method (SMOTE). Combining NB (SMOTE + OSS) with SVM led to an outstanding accuracy of 96.5%. Zhang et al. [19] employed a variety of machine learning techniques, including random forest (RF), naive Bayes (NB), support vector machines (SVM), decision trees (DT), and multilayer perceptron (MLP). The total accuracy of these algorithms was 62.04%. A study conducted by Saifudin et al. [20] examined data from a UCI (University of California, Irvine) math course. Also, they used the naive Bayes (NB) algorithm with forward selection as a feature selection technique for class imbalances by utilizing machine learning algorithms. Their method was 85.6% accurate. A comprehensive literature review indicates that there has been significant research on the use of machine learning algorithms to predict student performance. Even so, several unanswered questions and areas may require further research, especially about how imbalanced datasets affect prediction models. Though several studies have examined predictive modeling in higher education, few have examined the impact of imbalanced datasets on the model's performance. Most studies cited in the literature review have addressed imbalanced data using SMOTE, oversampling, and under-sampling methods. Nevertheless, there is still a lack of attention paid to this matter.

B. Proposed Methodology

This section provides a detailed overview of the proposed methodology. The main objective is to predict student grades and examine how an imbalanced data set affects the ability to predict student grades. The method begins with preparing the data, which includes collecting and analyzing the information. The next step is to pre-process the data to make a successful prediction. To prepare data for analysis, it must be cleaned and prepared. It consists of removing null values, encoding categorical variables, and scaling features. During this phase, data exploration and imbalance analysis are also performed, and the study represents class imbalances within the dataset. Once the data has been analyzed, the data analysis will be performed using two well-known machine learning algorithms, Random Forest (RF) and Logistic Regression (LR). In addition to analyzing the models for their exceptional ability to predict imbalanced datasets, this phase also separates the dataset into a training and testing set so that we may observe how the model performs when exposed to unknown data sets. For a model to be effective and accurate, evaluating it after it has been built is critical. Therefore, the next phase is to evaluate the model based on essential evaluation metrics such as accuracy, precision, recall, F-1 score, and confusion. All phases of the proposed methodology are depicted in Figure 1. A detailed explanation of these phases is provided in the following subsection.

1) Dataset Preparation: The analysis collects data from 410 students who completed an artificial intelligence (AI) course at the Universiti Teknologi Malaysia (UTM) faculty of computing. Here, the study focuses on some essential features of our datasets, for example, Student ID, Quiz_1, Midterm_1, Assignment_1, Assignment_2, Assignment_3, PRO_10, AD_10, Final_Exam_Marks, Total, Grade, and the target variable is Categories. Focusing on these specific features provides crucial data points for predicting grades. Student ID allows for individual tracking, while project and total marks directly relate to academic achievements.

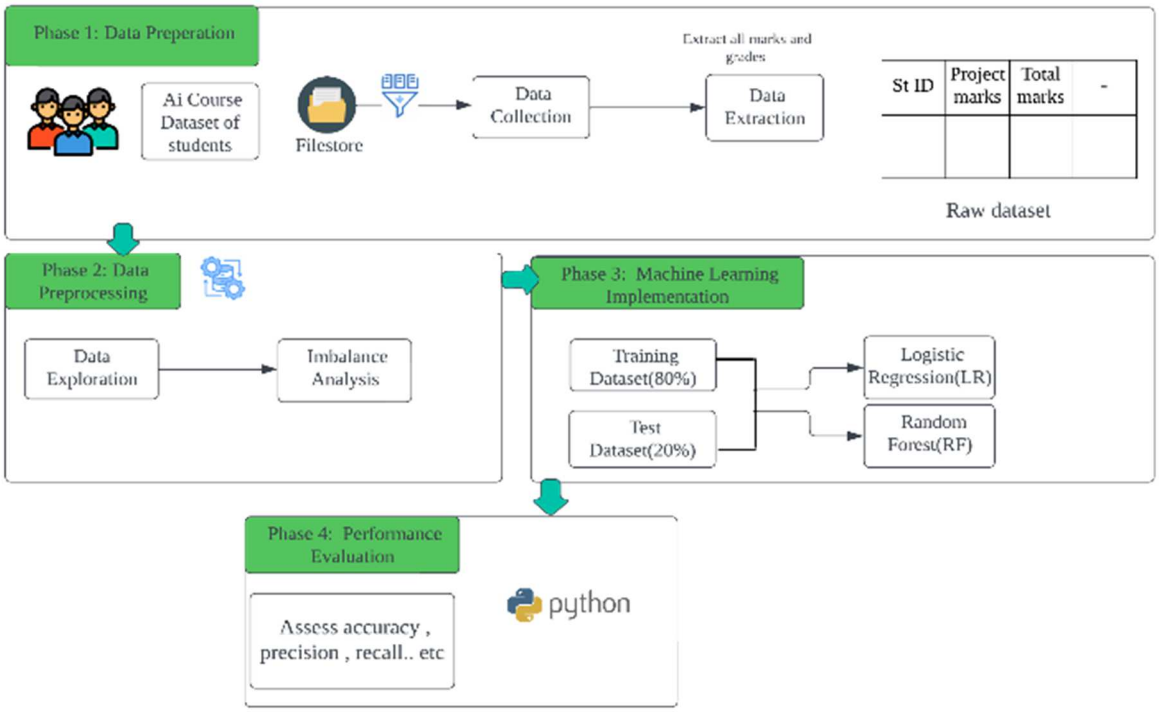

Fig. 1 Proposed Methodology

Then, the analysis calculated the total marks, and then, by applying the grade formula, it obtained the grades of specific students (A+, A, A-, B+, B, B-, C+, etc.). Finally, the analysis categorizes those grades into five categories: Excellent, Exceptional, Distinction, Pass, and Fail.

2) Data Preprocessing: Data preprocessing is one of the most important phases in any machine learning model, and it includes some crucial steps to prepare the data for a successful analysis.

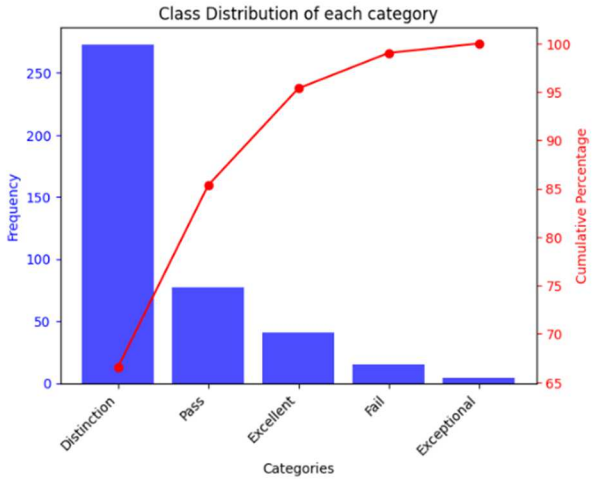

Fig. 2 Class distribution of each category

These steps include cleaning the data, converting non-numeric values into numeric format, and feature scaling, which will ensure that all features are brought to a similar scale. This proposed methodology for feature scaling standardization has been incorporated, and it is one of the most popular feature scaling techniques. The next step in data preprocessing is data exploration. The final state of the data preprocessing of our proposed methodology is imbalance analysis.

In our dataset, the analysis observed that the dataset is imbalanced, and the distribution of each class shows that the significant class is Distinction. A minor class is Exceptional, which can make our model biased toward the considerable class. The analysis shows a clear overview of the distribution of each class in Figure 2. The instances or frequencies of Distinction are 273, which will be considered a significant class, and the frequencies of Pass, Excellent, Fail, and Exceptional are 77, 41, 15, and 04 consecutively. Visualizing the frequencies can be done by focusing on the level of the bar chart of the mentioned Pareto chart, which demonstrates the characteristics of an imbalanced dataset for this AI dataset.

3) Machine Learning Implementation: In this phase, the analysis will split our dataset into test and training data; training data will be used to train our model, and test data will be used to evaluate our model. The data splitting ratio will be 80:20, meaning 80% of the data will be used for training, and the rest will be used for testing. Splitting training and test data proportions provides a robust predictive model for educational outcomes, improving any algorithm's ability to recognize patterns and provide accurate student grade projections. The next step in the study is to compare and select the best model. The analysis uses LR and RF to predict student grades. Of course, the question may arise: why is the analysis using these two ML algorithms? The reasons are that by using RF, the analysis can minimize the overfitting of our datasets, and by using LR, the analysis can improve accuracy as well. Besides that, the interpretability of LR will help investigate feature-grade correlations. These are the benefits of using LR and RF to predict student grades. Performance matrices like accuracy, precision, recall, and the F-1 score will be used to measure performance.

III. Results and Discussion

In this phase, the study presents the results obtained from the machine learning models and compares their efficiency. The interpretation of the result is presented using various diagrams. The study also demonstrates the impact of an imbalanced dataset on model performance. Finally, this phase decides the effect of an imbalanced dataset.

A. Experimental Results

In this section, our primary goal is to compare the predictive model based on its accuracy and performance. Here, the study trained the AI dataset using LR and RF algorithms, and the prediction accuracy of each approach was assessed. We evaluated the performance using various measures, such as f-1, recall, precision, and classification accuracy, to ensure the predictive model was well-fitted to produce reliable results. The prediction performance metrics of the LR and RF classifiers on the student dataset are compiled in Table 1. Table I presents the findings, which show that Random Forest predicts the best, with accuracy and recall values of 0.975, while the accuracy and recall of LR are 0.914, respectively. However, since the study focuses on an imbalanced dataset, we concluded that the prediction findings were insignificant due to overfitting and bias issues that may have arisen during dataset training because the classes in our dataset were significantly imbalanced.

TABLE I
COMPARISON OF PERFORMANCE FOR LR AND RF

Metrics	Logistic Regression	Random Forest
Accuracy	0.914634	0.975610
Precision	0.871021	0.955285
Recall	0.914634	0.975610
F1 Score	0.891272	0.964523

The study that follows in the next subsection describes how an imbalanced dataset affects the performance of the model. Figure 3(a) compares logistic regression and random forest based on their accuracy scores, which have been achieved in the experiment. Similarly, 3(b) shows the comparison for the F-1 score, 3(c) depicts the comparison for precision, and 3(d) shows the comparison for recall. Based on the representation in the following figures, the performance of the random forest is higher than the logistic regression for all evaluation metrics. Since random forest is an ensemble method, it can learn from imbalanced data more effectively [21] compared to logistic regression.

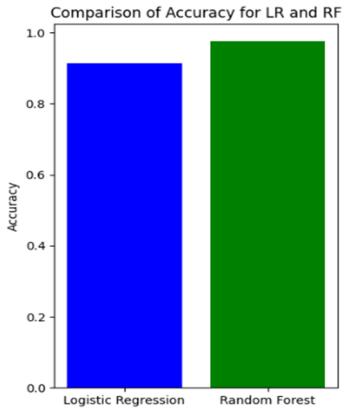

Fig. 3(a) Accuracy

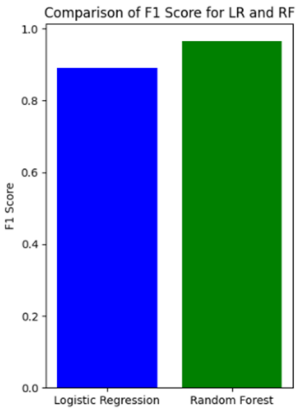

Fig. 3(b) F1

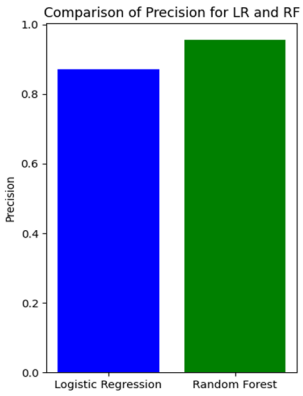

Fig. 3(c) Precision

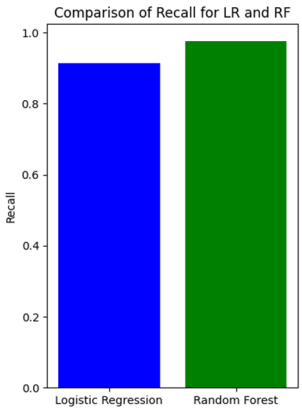

Fig. 3(d) Recall

B. Impact of Imbalanced Dataset on Performance

Imbalanced datasets have a considerable impact on a machine learning model's performance. The classification report in Table II shows that the LR model performed very well in predicting the "Distinction" class, which is a significant class in our dataset. The precision, recall, and F-1 scores are around 1.0, indicating that the Logistic Regression model can effectively predict instances of the "Distinction" class or significant class.

TABLE II
COMPARISON OF PERFORMANCE FOR LR AND RF

Predicted class	Precision	Recall	F1-score
Distinction	0.92	1.00	0.96
Excellent	1.00	0.90	0.95
Exceptional	**0.00**	**0.00**	**0.00**
Fail	**0.00**	**0.00**	**0.00**
Pass	0.86	0.86	0.86

However, when it comes to predicting minor classes like "Exceptional" and "Fail," the logistic regression model completely fails to predict those classes correctly, as indicated by the precision, recall, and F-1 score value, which is 0.0.

Similarly, the random forest model performed exceptionally well in predicting the primary class, which is "Distinction," where it achieved perfect precision, recall, and F-1 score. However, the random forest also performed exceptionally well for the "Fail" class, a minor class in our dataset. By looking at the classification report in Table III, it achieved perfect precision, recall, and F-1 scores, which are 1.0 for this minor class, which indicates the ability of RF to learn from an imbalanced dataset for this minor class.

TABLE III
COMPARISON OF PERFORMANCE FOR LR AND RF

Predicted class	Precision	Recall	F1-score
Distinction	1.00	1.00	1.00
Excellent	0.83	1.00	0.91
Exceptional	**0.00**	**0.00**	**0.00**
Fail	1.00	1.00	1.00
Pass	1.00	1.00	1.00

The random forest (RF) technique is a type of ensemble learning where many decision trees are aggregated and used to reduce variance [22]. Since RF is an ensemble method, RF showed better generalization and performance than LR. The impact of the imbalanced dataset for LR and RF can be visualized more precisely in Figure 4, where the study compares the classification reports for those two popular machine learning algorithms. Looking at Figure 4, it is clear that both algorithms failed to predict the "Exceptional" class, which is the most minor class in our dataset. The confusion matrix allows the visualization of the performance of a model, and it plays a crucial role in understanding the performance of the classification model.

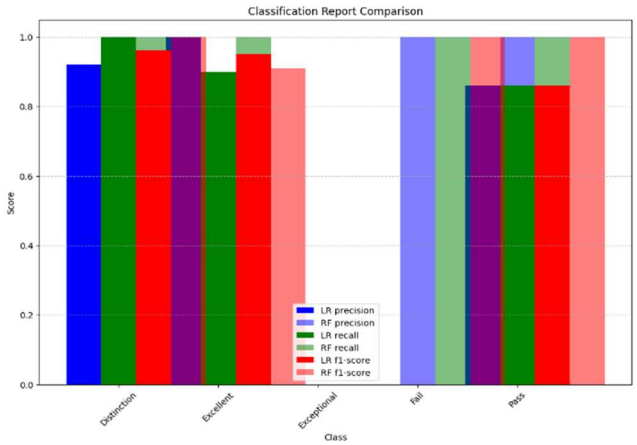

Fig. 4 Comparison of Performance for LR and RF

Looking at the confusion matrix in Figure 5, it is clear that RF performs better than LR, especially in predicting major classes (Distinction and Excellent). However, both algorithms struggle with predicting minor courses. LR failed to predict minor classes (Exceptional and Fail) from the confusion matrix.

In summary, in this section, the study can decide from the detailed investigation and visualization by classification report and confusion matrix. The imbalanced dataset significantly impacts model performance, mainly if it affects the ability of the model to predict minority classes. In our study, both the LR and RF models performed exceptionally well on the majority class. In contrast, RF performed better in the minority class due to its ability to capture complex relationships in the dataset.

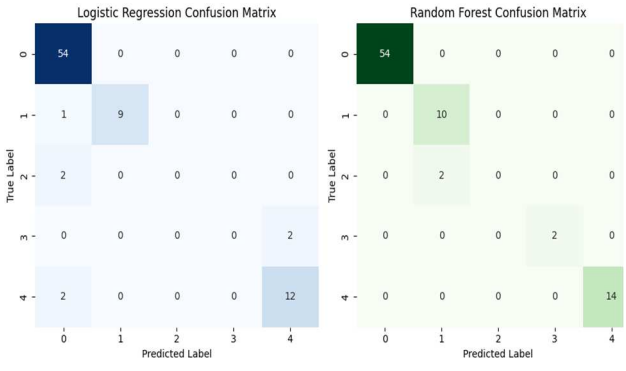

Fig. 5 Confusion matrix for LR and RF

C. Discussion

The findings of this research demonstrate the importance of machine learning, especially two popular machine learning algorithms, Logistic Regression (LR) and Random Forest (RF), to predict student grades based on an artificial intelligence course dataset from the Universiti Teknologi Malaysia (UTM). Not only that, but the study also evaluated the predictive model by using various evaluation metrics such as accuracy, precision, recall, and F1 score. The results prove that RF performs better accuracy and recall, achieving values of 0.975, or 98%, compared to LR, which is 0.914, or 91%. Additionally, RF showed exceptional performance in predicting both major and minor classes, achieving perfect precision, recall, and F1 scores for "Fail," a minor class in the dataset. In contrast, LR performed very well in predicting significant courses such as "Distinction" but struggled significantly with minor classes, incredibly "Exceptional" and "Fail." It ultimately failed to predict these classes, as indicated by the precision, recall, and F1 scores of 0.0 for these classes.

Another important aspect is also evident: imbalanced data sets impact model performance. especially the inability of LR to accurately predict the minor classes since LR can perform very well for binary classification [23]. While RF also faced massive challenges in predicting the "Exceptional" class, it was successful in predicting one minor class, "Fail", which means it showed better generalization and the ability to learn from imbalanced data effectively due to its ensemble nature.

IV. Conclusion

Finally, this study demonstrates the effectiveness of machine learning techniques, especially logistic regression and random forest, in predicting student grades in an imbalanced dataset. Through rigorous evaluation and detailed experimentation, RF emerges as the superior model by handling class imbalance and successfully predicting major and minor classes compared to LR. Moreover, the study also investigated the challenges associated with imbalanced datasets, particularly the limitations or failure of those algorithms in predicting minor courses in the dataset.

These findings emphasize the importance of implementing vigorous techniques to mitigate class imbalance and enhance performance in educational prediction tasks. However, addressing class imbalances is still a crucial aspect of developing any model to ensure fair and accurate predictions across all classes. Recently, predictive analytics have become surprisingly popular in most HEIs (higher education institutions). These analytics utilize sophisticated analytics, which includes machine learning deployment, to generate high-quality performance and meaningful data for all educational levels. A dynamic change is happening with the incredible use of predictive analytics, especially to evaluate students' grades. [24] Predictive modeling can be used to identify vulnerable students in higher education, which will contribute a lot to HEI [25]. Several areas need further exploration, including investigating advanced feature engineering techniques for educational datasets, which can uncover valuable insights. Exploring the use of ensemble algorithms to optimize the model's performance may offer further improvement in handling imbalanced datasets. Increasing the number of features and instances in the dataset

will lead to better analysis and more effective predictions. Overall, future work should prioritize resolving the issues raised by imbalanced datasets and developing techniques to improve the accuracy and usefulness of prediction models in educational settings.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Malaysia (MOHE) under the Fundamental Research Grant Scheme (FRGS), FRGS/1/2023/ICT02/UTM/02/8.

REFERENCES

[1] 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.

[2] D. Solomon, "Predicting Performance and Potential Difficulties of University Student using Classification : Survey Paper," *Int. J. Pure Appl. Math.*, vol. 118, no. 18, pp. 2703–2707, 2018.

[3] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, 2017, doi: 10.1016/j.chb.2017.01.047.

[4] Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, "Graphs regularized robust matrix factorization and its application on student grade prediction," *Appl. Sci.*, vol. 10, no. 5, pp. 1–19, 2020, doi:10.3390/app10051755.

[5] 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.

[6] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, 2015, doi: 10.1186/s40165-014-0010-2.

[7] I. Khan, A. Al Sideiri, A. Ahmad, and N. Jabeur, "Tracking Student Performance in Introductory Programming by Means of Machine Learning," Feb. 2019, pp. 1–6. doi:10.1109/ICBDSC.2019.8645608.

[8] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[9] 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.

[10] L. Ismail, H. Materwala, and A. Hennebelle, "Comparative Analysis of Machine Learning Models for Students' Performance Prediction," in *Advances in Intelligent Systems and Computing*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 149–160. doi:10.1007/978-3-030-71782-7_14.

[11] B. Flanagan, R. Majumdar, and H. Ogata, "Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis," *Int. J. Educ. Technol. High. Educ.*, vol. 19, no. 1, Dec. 2022, doi: 10.1186/s41239-022-00348-4.

[12] A. Polyzou and G. Karypis, "Grade prediction with models specific to students and courses," *Int. J. Data Sci. Anal.*, vol. 2, no. 3–4, pp. 159–171, Dec. 2016, doi: 10.1007/s41060-016-0024-z.

[13] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," *Appl. Math. Sci.*, vol. 9, no. 129, pp. 6415–6426, 2015, doi:10.12988/ams.2015.53289.

[14] T. Anderson, "Applications of Machine Learning to Student Grade Prediction in Quantitative Business Courses," 2017.

[15] E. C. Abana, "A decision tree approach for predicting student grades in Research Project using Weka," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 285–289, 2019, doi:10.14569/ijacsa.2019.0100739.

[16] I. Khan, A. Al Sadiri, A. R. Ahmad, and N. Jabeur, "Tracking Student Performance in Introductory Programming by Means of Machine Learning," in *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 2019, pp. 1–6. doi:10.1109/ICBDSC.2019.8645608.

[17] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, "The potential for student performance prediction in small cohorts with minimal available attributes," *Br. J. Educ. Technol.*, vol. 51, no. 2, pp. 347–370, Mar. 2020, doi: 10.1111/bjet.12836.

[18] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, "Hybrid resampling to handle imbalanced class on classification of student performance in classroom," Feb. 2017, pp. 207–212. doi:10.1109/ICICOS.2017.8276363.

[19] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, "Grade Prediction of Student Academic Performance with Multiple Classification Models," Feb. 2018, pp. 1086–1090. doi:10.1109/FSKD.2018.8687286.

[20] A. Saifudin, Ekawati, Yulianti, and T. Desyani, "Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, 2020. doi:10.1088/1742-6596/1477/3/032007.

[21] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," *Discovery*, no. 1999, pp. 1–12.

[22] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–15, 2018, doi: 10.1186/s12859-018-2264-5.

[23] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[24] P. Brous and M. Janssen, "Trusted decision-making: Data governance for creating trust in data science decision outcomes," *Adm. Sci.*, vol. 10, no. 4, 2020, doi: 10.3390/admsci10040081.

[25] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing autoML in educational data mining for prediction tasks," *Appl. Sci.*, vol. 10, no. 1, pp. 1–27, 2020, doi:10.3390/app10010090.