Predicting the Subject Grades of Students with Reduced Error Pruning Tree and Cost Sensitive Learning

Pyae Myat Noe1, a), Keng-Hoong Ng1, b) and Gee-Kok Tong2, c)

1Lee Kong Chian Faculty of Science and Technology, Universiti Tunku Abdul Rahman, Bandar Sungai Long, Kajang, 43000, Malaysia

2Center for Natural Language Processing (CNLP), CoE for Artificial Intelligence, Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor 63100, Malaysia

c) Corresponding author: gktong@mmu.edu.my

a) pyaemyatnoe@1utar.my

b) nkhoong@utar.edu.my

**Abstract.** Predicting student grades is one of the important tasks in education, as for the early identification of at-risk students and for enhancing learning outcomes. This study focused on predicting student performance by applying supervised learning and feature selection to a dataset of Portuguese language students from two secondary education schools in Portugal. Firstly, data preprocessing and feature selection were conducted on the dataset. Then, a few widely used base classifiers as well as ensemble methods were applied and evaluated to determine the best-performing model. Among the classifiers, the Decision Tree (REPTree) model with eight selected features achieved the highest accuracy of 76.58%. In addition, we attempted to improve further the accuracy by applying a cost-sensitive classifier to address the class imbalance problem in the dataset. As a result, the accuracy improved to 76.89%. These achieved results outperform the best result reported (76.1%) in the earlier published work that utilized the same dataset. This concludes that an efficient forecast model for student grades could be built by using the REPTree model and coupled with cost-sensitive learning.

# INtroduction

Today, education is not only for acquiring basic knowledge, but also a foundation for personal growth and economic development. Academic achievement is regarded as a measure of students’ knowledge and skills, and it is one of the important key factors for employment opportunities [1]. This achievement is widely measured through one’s Grade Point Average (GPA) in the educational sector [2]. In this context, forecasting student grades has become one of the critical focuses for the educational sector. This is because with the predicted outcomes, institutions can identify at-risk students to prevent dropouts and students could improve their overall performances.

Machine Learning (ML) has become a powerful tool for predictive analysis in various sectors, including business and education [3]. ML is used in the education field to improve the learning process of students [4]. Educational sectors leverage ML techniques such as classification, regression, and clustering to analyze student data and forecast their performance. In this study, classification algorithms were applied to analyse students’ performance in Portuguese Language course for the final period (G3) grade from two secondary public schools in Portugal. This research started with data preprocessing, and then followed by training and evaluating the machine learning models. Single and ensemble classifiers were tested after data transformation, preprocessing, and feature selection processes.

# Related work

In order to predict student performance accurately, it is important to select relevant attributes when applying machine learning (ML) techniques [5]. These attributes can be categorized into various groups, including GPA and grades, demographics, psychological profiles, cultural background, academic progress, and educational history. For example, a study by [6] achieved 73% accuracy rate in predicting final exam grades by using only academic data which is mid-term grade. It contributed to the reliability of academic factors in predicting student performance. However, non-academic factors also significantly influence students' academic performance. A study by [7] demonstrated that family, economic, and social status and environment significantly affect the student performance with the mother’s educational level as a key predictor. Additionally, [8] revealed that personal, social, and geographical factors directly influence the teaching-learning process. For example, variables such as 'neighborhood' and 'school' were identified as the major factors affecting student failure rates while ‘grade’ was demonstrated as the most important predictor in predicting student performance. Among these, academic data such as grades and GPA have been shown to be the most significant predictors of student performance [9].

In addition, supervised ML approaches are commonly applied in predicting student performance because of their accuracy and consistency. Ersozlu et al.[10] noted that classification is among the most widely used methods for predicting performance and classifiers such as Decision Trees (DT), Artificial Neural Networks (ANNs), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Linear Regression (LR), and Naïve Bayes (NB) are commonly employed. Another study by [11] reported that ANN outperformed other models with a 98.3% accuracy rate, followed by DT at 98.2%. However, many studies applied only one classifier without comparing it with other methods. In this work, we will compare four base classifiers and ensemble methods to identify the best-performing model for the dataset.

Cortez et al*.* [12] have explored the Portuguese and Mathematics student datasets using various supervised learning approaches. They categorized the final period (G3) column into binary classification, 5-level classification, and regression. Other research works that utilized the same dataset tried to improve the model accuracy rate. For example, [13] and [14] applied binary classification to the Portuguese dataset where the former achieved 91.56% accuracy using an ensemble method of GBDT and XGBoost, while the latter obtained 96.31% accuracy by Random Forest that outperformed the accuracy of 93% achieved by the original work. For the 5-level classification, the study by [12] yielded the highest accuracy with Decision Tree (76.1%), while [15] managed to acquire 72.57% of accuracy rate. The 5-level classification results indicated that there is still room for improvement. In light of this, this study focused on the Portuguese dataset with a 5-level classification approach to predict the final student grades (G3).

# Research design and methodology

## Data Understanding and Preprocessing

The dataset in Table 1 used in this study is from the two Portuguese secondary schools that consisting of 649 student records with 33 attributes of academic, demographic, and socio-economic factors. [12] The grading system used in Portuguese high school and secondary education ranges from 0 to 20, where 0 represents the lowest grade and 20 the highest.

Data preprocessing was implemented using the Waikato Environment for Knowledge Analysis (WEKA) open-source software tool. First, the final grade (G3) column was discretized into a 5-level classification: A, B, C, D, and F (Figure 1). For easier reference in the following sections, this discretized dataset is labeled as Discretized-dataset.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

**FIGURE 1.** (a) Before data discretization, (b) After discretization based on the 5-level classification

After discretization, several preprocessing techniques were applied to the dataset before feeding it to the machine learning models. It started with the feature selection using wrapper method with CfsSubsetEval as the attribute evaluator, and BestFirst as the forward search method. This process selected eight key attributes which are relevant to the target variable (G3) and they are: studytime, failures, schoolsup, paid, activities, internet, G1, and G2. The shortlisted eight attributes were compiled to form a dataset labeled as Wrapper-dataset. The filter method was also applied to the dataset using GainRatioAttributeEval and ranker search method. After removing the occurrences of zero gain ratio as they are considered as no relevance to the target, a total of 24 attributes was selected and then formed a dataset labeled as Filter-dataset.

**TABLE 1.** Description of the dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Description** | **Features** | **Description** |
| School | GP - Gabriel Pereira  MS - Mousinho de Silveira | Famsize  (family size) | LE3 - less or equal to 3  GT3 - greater than 3 |
| Age | 15 to 22 | Guardian | mother, father, other |
| Address | U – urban, R - rural | Sex | F – female, M - male |
| Pstatus (parent’s cohabitation status) | T - living together  A - apart | Mjob (mother's job)  Fjob (father’s job) | Teacher, health, services, at home, others |
| Medu (mother's education)  Fedu (father’s education) | 0 - none  1 - primary education  2 - 5th to 9th grade  3 - secondary education  4 - higher education | absences (number of school absences) | From 0 to 93 |
| Reason (reason to choose this school) | Home, reputation, course, other | failures | n if 1<=n<3, else 4 |
| Traveltime (home to school) | 1 - less than 15 min  2 - 15 to 30 min  3 - 30 min to 1 hour  4 - more than 1 hour | Studytime (weekly) | 1 - less than 2 hours  2 - 2 to 5 hours  3 - 5 to 10 hours  4 - more than10 hours |
| schoolsup (extra educational support)  famsup (family educational support)  paid (extra paid classes)  nursery (attended nursery school)  extra (curricular activities)  higher (wants to take higher education)  internet (Internet access at home)  romantic | Yes or No | famrel (quality of family relationships)  freetime (free time after school)  goout (going out with friends)  Dalc (workday alcohol consumption)  Walc (weekend alcohol consumption)  health (current health status) | 1 - very low to 5 - very high |
| G1,G2 | From 0 to 20 | G3 (final grade) | From 0 to 20 |

Many machine learning models struggle with directly handling categorical data in text form, as they typically require all inputs to be numerical through encoding methods. Hence, one-hot encoding transformed categorical values into binary columns, and it prevents the model from misinterpreting categorical values as ordinal. Thereupon one-hot encoding was applied to Mjob, Fjob, reason, and guardian attributes to handle categorical variables with non-binary values. However, one-hot encoding increases the number of attributes because each categorical attribute is expanded into multiple columns. The total attributes expanded from 33 to 46 variables with the additional categorized columns. Due to the fact that higher dimensionality could cause noise and redundancy, hence we applied the wrapper method after one-hot encoding. This feature selection technique selected eight attributes to form a dataset labeled as One-hot-wrapper-dataset, it has attributes such as Mjob=at\_home, studytime, failures, schoolsup, paid, activities, G1, and G2.

Overall, we prepared four variants of the dataset for modeling, i.e. Discretized-dataset, Wrapper-dataset, Filter-dataset, and One-hot-wrapper-dataset. Each dataset variant was used to train and test with different machine learning models to compare their performances and then to determine the best performing model.

## Modeling and Evaluation

Several machine learning algorithms are applied to forecast the students’ grades: Logistic regression, Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (J48 and REPTree). These algorithms were selected due to their predictive capabilities and effectiveness in identifying learning patterns in educational data [16]. In the modelling phase, default algorithms’ parameters were applied. In order to assess the performance of each model, a 10-fold cross validation technique was used [17]. This approach helps to prevent overfitting and guarantees an impartial evaluation, as every fold serves as the test set once. The repeated iterations are designed to measure the model's ability to generalize effectively. The discretized dataset before any feature selection was used as the baseline for comparing the accuracy rate with other variant datasets that use different preprocessing methods.

On top of using the base classifiers, several ensemble methods with default parameters were also performed on the datasets with the aim to improve the accuracy rate. The ensemble methods combined multiple base models to create a more accurate model. However, based on the dataset and its nature, the ensemble models can be more effective or sometimes give lower predictive performance than single classifiers. Several commonly used ensemble methods such as Random Forest [18], Bagging, Boosting, and Stacking were used in this study.

# Results and discussion

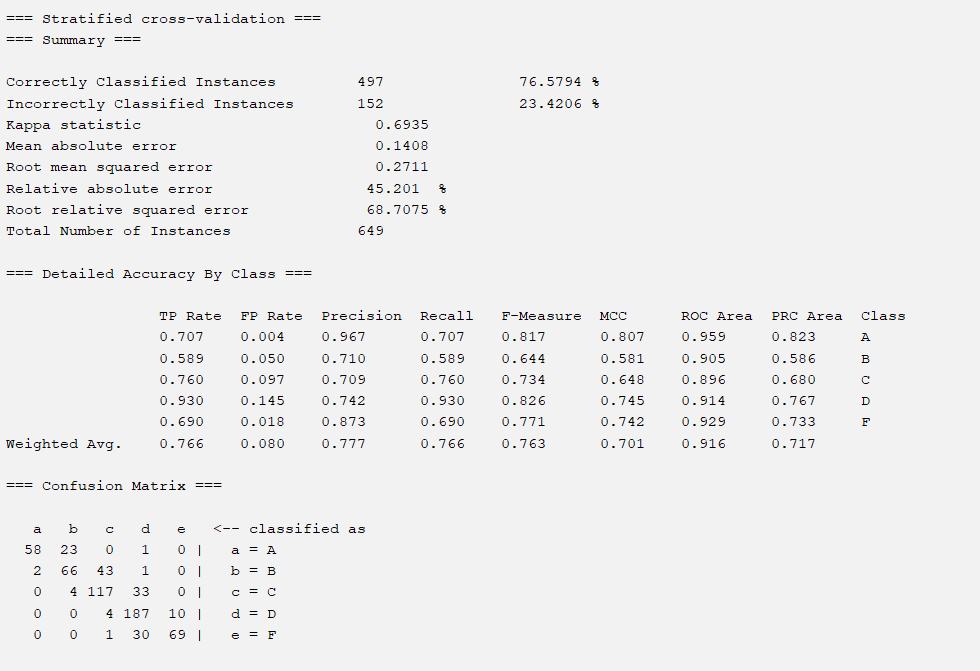
Table 2 shows the accuracy rates obtained by different classifiers. Generally, the REPTree model achieved the highest accuracy rate among the base classifiers. The discretized dataset before any feature selection showed that REPTree achieved the highest accuracy of 75.35% while SVM performed the worst at 57.47%. The wrapper method contributed significant improvements across most of the classifiers. SVM improved its accuracy by 12.64%, suggesting that the original feature set contained irrelevant features that were hindering SVM's performance. Despite the improvement, SVM is still the worst performer at 70.11% while REPTree has the highest accuracy rate, achieving 76.12%. The Filter method also improved the performance but generally less than the Wrapper method. The result showed that REPTree's accuracy performance dropped slightly with the filter method, this indicated that some discarded features were useful for this classifier. In the Filter-dataset, REPTree remained at the highest accuracy rate, 74.73%, while SVM stayed as the worst performer with 59.01%. We obtained the highest accuracy rate of 76.58% with REPTree model, in the dataset preprocessed with one-hot encoding followed by the wrapper method. This is the best achieved accuracy rate among the tested base classifiers.

**TABLE 2.** Accuracy comparison of different single classifiers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset Variant** | **Accuracy Rate (%)** | | | | |
| **Naïve Bayes** | **Logistic Regression** | **SVM** | **DT (J48)** | **DT (REPTree)** |
| Discretized-dataset | 68.26 | 66.72 | 57.47 | 67.80 | **75.35** |
| Wrapper-dataset | 71.50  (+3.24) | 73.96  (+7.24) | 70.11  (+12.64) | 75.35  (+7.55) | **76.12**  (+0.77) |
| Filter-dataset | 69.49  (+1.23) | 68.72  (+2.00) | 59.01  (+1.54) | 71.50  (+3.70) | **74.73**  (-0.62) |
| One-hot-wrapper-dataset | 71.03  (+2.77) | 72.88  (+6.16) | 69.03  (+11.56) | 75.04  (+7.24) | **76.58**  (+1.23) |

Figure 2 shows the detailed class-wise performance of the REPTree model on the dataset with one-hot encoding followed by the wrapper-based feature selection. The confusion matrix shows that out of 112 grade B students, only 66 were correctly classified as actual grade B while 43 students were misclassified as grade C. This suggested that the model found it difficult to distinguish between grade B and C. The model for class C has a moderate precision and recall rate, 0.709 and 0.760 respectively, while the model has a very high recall rate for class D with 93%. Finally, when the model predicts class F with 87.3%, it is usually correct in predicting, but some actual grade F students are misclassified as grade D. Overall, the model is highly precise in predicting grade A and F, but not so efficient in predicting grade B students.

Since class B is the worst-performing class, we tried to apply cost-sensitive learning to class B to improve its performance by adjusting the misclassification penalties. After applying a double penalty to the class C misclassifications with class B, the accuracy improved slightly from 76.58% to 76.89%. The correctly predicted class increased to 69 while misclassified grade C students decreased to 40, suggesting that the cost-sensitive classifier reduced the misclassification errors. The trade-off resulted in minor changes for other classes, but overall, the model performed moderately across different grades. The highest accuracy rate (76.58%) was observed when the dataset was treated with one-hot encoding and feature selection. Then, the accuracy improved slightly to 76.89% after the cost-sensitive classification was applied.



**FIGURE 2**. Detailed accuracy by class and confusion matrix by REPTree model

Table 3 shows the performance of accuracy rates for the ensemble methods on the four dataset variants. The Bagging model achieved the highest accuracy of 75.19% on the Discretized-dataset (prior to any feature selection), while Stacking being the worst performer, recorded the accuracy of 72.42%. The Wrapper method improved the accuracy of all ensemble classifiers except Random Forest which is the worst performer (71.03%). Bagging improved by 0.31% to score the highest accuracy rate of 75.5%. Meanwhile, with the Filter method, the performance of ensemble models declined in general except Bagging which achieved the highest accuracy of 75.89%. Lastly, the Bagging method gained the highest accuracy of 75.65% in the One-hot-wrapper-dataset. Overall, Bagging consistently outperformed other ensembles across different dataset variants. Overall, the ensemble methods did not significantly perform better than the base classifiers in this study. The accuracy rates from ensemble methods are only within the range of 70 to 75 percentage points.

**TABLE 3**. Accuracy Comparison of Ensemble Methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset**  **Variant** | **Accuracy Rate (%)** | | | |
| **Random Forest** | **AdaBoost (REPTree)** | **Stacking**  **(base- REPTree, meta-NB)** | **Bagging (REPTree)** |
| Discretized-dataset | 73.65 | 73.96 | 72.42 | **75.19** |
| Wrapper-dataset | 71.03  (-2.62) | 75.19  (+1.23) | 73.81  (+1.39) | **75.50**  (+0.31) |
| Filter-dataset | 70.26  (-3.39) | 71.65  (-2.31) | 68.26  (-4.16) | **75.81**  (+0.62) |
| One-hot-wrapper-dataset | 70.42  (-3.24) | **75.65**  (+1.69) | 71.96  (-0.46) | **75.65**  (+0.46) |

All in all, after modeling with different classifiers, the Decision Tree (REPTree) model achieved the highest accuracy of 76.58% and it slightly improved to 76.89% after applying the cost-sensitive learning. Both results surpassed the highest accuracy of 76.1%, which was reported from the earlier published research by [12] that employed the same dataset.

# conclusion and future reserach

This study aimed to predict student performance by applying supervised learning models, i.e. base classifiers and ensemble classifiers, with the combination of feature selection and cost-sensitive learning to a Portuguese language students’ dataset. In conclusion, the REPTree Decision Tree model is the best performer, it achieved the highest accuracy of 76.89%, outperforming all other tested models. The result also showed that feature selection and cost- sensitive learning play key roles in improving prediction accuracy. For future work, we may consider extending the research work to larger datasets, and developing a dashboard that incorporates recommendations such as tailored study plans and suggestions which can help to improve academic progress.

# References

1. K. Meepradit, T. Auamnoy, and N. Leelakanok, “Factors affecting GPA in higher education in Asia: A systematic review,” *International Journal of Early Childhood Special Education* **14**, 1581–1595 (2022).
2. R. Steinmayr, A. Meißner, A. F. Weidinger, and L. Wirthwein, and Wirthwein, L., “Academic achievement,” in *Education (Chula Vista)*, (Oxford University Press, 2014).
3. H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Computers and Electrical Engineering* **89**, (2021).
4. Z.Iqbal, J.Qadir, A.N.Mian, and F.Kamiran, “Machine learning based student grade prediction: A case study,” (2017).
5. Y. Z. Li, K. H. Ng, K. C. Khor, and Y. H. Lim, “Predicting undergraduate academic success with machine learning approaches,” in *International Conference on Soft Computing and Data Mining*, pp. 144–153, Springer Nature, Switzerland (2024).
6. M.Yağcı, “Educational data mining: Prediction of students’ academic performance using machine learning algorithms,” *Smart Learning Environments* **9**(1), (2022).
7. J. D. García-González and A. Skrita, “Predicting academic performance based on students’ family environment: Evidence for Colombia using classification trees,” *Psychology* **11**(3), 299–311 (2019).
8. E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. van Erven, “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil,” *Journal of Business Research* **94**, 335–343 (2019).
9. H. Altabrawee, O. Abdul, J. Ali, and Q. Ajmi, “Predicting students’ performance using machine learning techniques,” *Journal of University of Babylon for Pure and Applied Sciences* **27**(1), 194–205 (2019).
10. Z. Ersozlu, S. Taheri, and I. Koch, “A review of machine learning methods used for educational data,” *Education and Information Technologies*, (2024).
11. Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, “Assessment and evaluation of different machine learning algorithms for predicting student performance,” *Computational Intelligence and Neuroscience* **2022**, (2022).
12. P. Cortez and A. M. G. Silva, “Using data mining to predict secondary school student performance,” (2008).
13. S. Zhang, J. Chen, W. Zhang, Q. Xu, and J. Shi, “Education data mining application for predicting students’ achievements of Portuguese using ensemble model,” *Science Journal of Education* **9**(2), 58 (2021).
14. M. Singh, C. Verma, R. Kumar, and P. Juneja, “Towards enthusiasm prediction of Portuguese school’s students towards higher education in real-time,” in *Proceedings of the 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, pp. 421–425, IEEE (2020).
15. A. K. Shrivas and P. Tiwari, “Comparative analysis of models for student performance with data mining tools,” *International Journal of Computer Trends and Technology* **46**(1), (2017).
16. Z. Ersozlu, S. Taheri, and I. Koch, “A review of machine learning methods used for educational data,” *Education and Information Technologies*, (2024).
17. G. Czibula, G. Ciubotariu, M. I. Maier, and H. Lisei, “IntelliDaM: A machine learning-based framework for enhancing the performance of decision-making processes. A case study for educational data mining,” *IEEE Access* **10**, 80651–80666 (2022).
18. M. Pandey and S. Taruna, “A comparative study of ensemble methods for students’ performance modeling,” *International Journal of Computer Applications* **103**(8), (2014).