A Hybrid Approach Using Decision Tree and Multiple Linear Regression for Predicting Students' Performance

Huu Huong Xuan Nguyen[1], Tran Khanh Dang[2(✉)], and Ngoc Duy Nguyen[3]

[1] The Coca-Cola Company, Ho Chi Minh City Branch, Vietnam
[2] Ho Chi Minh City University of Technology, VNU-HCM, Ho Chi Minh City, Vietnam
khanh@hcmut.edu.vn
[3] Deakin University, Geelong, Australia
n.nguyen@deakin.edu.au

Abstract. A high percentage of graduates reflect the quality of training followed by an integral demand for human resource management. Traditionally, Grade Point Average (GPA) is the number indicating student performance at the end of each semester or a study program. If educational organizations engage in student performance at an early stage, they can provide proactive guidance to aid students to escalate the chance of completing a study program and eventually achieve a better GPA. On the other hand, there are few professional tools that allow universities to evaluate their students on a large scale by using different non-academic criteria/dimensions. Therefore, this study develops a Decision Support System (DSS) to assist educators to evaluate their students. Specifically, the system consists of three sub-modules and two steps of prediction. In the first step, DSS applies an attribute-based ranking method to assess the influence of each performance variable in the student learning program. Secondly, the system uses a decision tree and multiple linear regression algorithms to find out groups of students that have a higher potential of graduation. As a result, educators can provide appropriate plans and guidance for each group of students.

Keywords: Decision support system · Decision tree · Multiple linear regression · Student performance prediction

1 Introduction

Existing studies indicated that if educators gain insights into the performance of students in their classes; they can take proactive plans to improve the learning program [8]. Besides, student capability is one of crucial factors that affect learning performance, especially in an academic environment. Factors such as economics, student demographics, and educational backgrounds have a great influence on student completion rates [19]. Hence, different perspectives are required to predict student performance.

On another hand, finding a method that accurately predicts student performance is a challenging task owning to a wide array of issues. For instance, one main issue is

T. K. Dang et al. (Eds.): FDSE 2021, CCIS 1500, pp. 23–35, 2021.
https://doi.org/10.1007/978-981-16-8062-5_2

that performance prediction methods are normally inefficient because of an improper use of attributes/variables [5]. Teachers, administrators, and policy makers increasingly rely upon automated technologies for pedagogical decision-making [1, 2, 21]. Besides, collecting data from untrusted sources creates a ubiquitous surveillance. Moreover, people may frequently change data, evaluation mechanisms, and records used to assess and represent student achievements, academic credits, and intellectual mastery. Judgments in these systems are often opaque and inadvertent; but they contribute important consequences on achievements in academic and employment environments [25].

Motivated by existing studies and [23], we developed a proof of concept, i.e., a prototype, for a Decision Support System (DSS). The prototype consists of two components. The first component is a friendly interface that allows users to enter data and collect desired outcomes. Secondly, the prototype provides a decision tree classification model and a multi-linear regression model.

The final aim of this study is to create an automated prediction software to analyze student performance and generate recommendation reports for educators; and thus, aid teachers to carry out appropriate actions to improve the performance of their students.

2 Related Work

The implementation of intelligent methods is essential for extracting data patterns and analytical knowledge that are inherently hidden inside student databases [10]. This new research direction has become popular and attracted research interest, especially in the new era of modern education. This is due to its great means of enhancing the quality of education systems [4].

In [9], the author worked with 206 student records with three models including Artificial Neural Network (ANN) a decision tree model and a linear regression model. Based on the validation results, the smallest accuracy of 0.1714 is attained by the ANN model while the decision tree model attains an accuracy of 0.1769 and the linear regression model achieves an accuracy of 0.1848.

In [6], student performance is assessed using an association-rule mining algorithm. The study has been done by assessing student performance with various attributes (e.g., attendance, assignment, and unit test). The results include rules that measure the correlations between various attributes. The rules provide appropriate guidance to improve student academic performance. The study also inferred that student performance is poor in a unit test if either the attendance or assignment is limited or both. Therefore, the authors suggested to utilize the use of unit test, i.e., the higher the score, the higher chance of graduation.

In [22], the authors aimed to explore a relationship between contextual background characteristics and academic performance of students to identify factors that associate to a chance of achieving a ‘good degree’. Studies reveal that grades at school are not the only causal factor in academics. This research is evidenced by the fact that all variables, such as Index of Multiple Deprivation (IMD) school type, school performance, neighborhood participation, sex, and ethnicity, are significantly associated with entry grades and demographics have affected to the scores.

To the best of our knowledge, there are many studies comparing the effectiveness among models but few of them explains related variables. Therefore, this paper proposes

a hybrid approach that includes an ability of classification based on a decision tree and a prediction function based on multiple linear regression. By using a modern approach, the more classes are classified, the more appropriate guidance is suggested.

3 Methodology

Researchers aim to develop a decision support system to support universities to estimate the student performance at the end of each semester. Predicting students' academic performance initiates a formula between input variables (x) and an output variable (Y), where (x) include multiple factors of students and the output (y) represents the graduation performance of students. The goal of this supervised learning is to find the best equation that can predict the output. This section describes a step-by-step implementation of the proposed method including data exploration, feature selection, a decision tree, and multiple linear regression algorithms.

3.1 Dataset

In this paper, we use a dataset provided by Delahoz-Dominguez et al. [7]. The dataset was obtained by orderly crossing through the databases of the Colombian Institute for the Evaluation of Education (ICFES). This dataset has more than 12,412 records and 33 columns which are dedicated to technical engineering students. The variables describe students' personal information (categorical) and scores (numerical).

Student demographics include 20 attributes such as Gender, Edu_Father, Edu_Mother, Occ_Father, Occ_Mother, People_House, Internet Tv, Computer, Washing_Mch, Mic_Oven, Car, Dvd, Fresh, Phone, Mobile, Revenue, Job, School_Nat, and School_Type. Edu_Father and Edu_Mother present an education level of an individual's father and mother, respectively. Occ_Father and Occ_Mother indicate an occupation of parents. Internet Tv, Computer, Washing_Mch, Mic_Oven, Car, Dvd, Fresh, Phone, Mobile, Revenue and Job stand for living conditions.

Students' scores include 12 attributes. MAT_S11, CR_S11, CC_S11, BIO_S11, ENG_S11 describe the ability of a student when starting studying. These scores present for national standardized test at the final year of the high school. Five generic academic scores at the final year of the professional career on Engineering are QR_PRO, CR_PRO, CC_PRO, ENG_PRO, WC_PRO. Quantitative reasoning (QR_PRO) assesses the ability to understand and manipulate quantitative data such as tables, graphs, or diagrams. Critical Reading (CR_PRO) assesses understanding capability. Citizen competency (CC_PRO) assesses the concept of citizenship by the Colombian constitution. English (ENG_PRO) assesses an ability of communication in English. Written communication (WC_PRO) assesses an ability of describing a topic or subject.

Engineering Project Formulation variable (FEP_PRO) is average of project in academic program. G_SC score is stated as the "average score of academic programs" in the description of the dataset [7]. These scores do not relate to generic academic scores. Hence, the FEP_PRO and G_SC are considered as attributes in the dataset.

Besides, the label in this dataset is Quartile. Quartile is a student ranking methodology in the dataset description. Quartile 1 includes 25% of total students who have the lowest

score to the median of the dataset. Quartile 2 is 25% of the total number of students whose scores coincide with the median of the dataset. Quartile 3 includes students who have score in the range from 51% to 75% of the dataset. Lastly, 25% of students who have the highest score are belong to the quartile 4. Therefore, when running regression equation, the model's prediction result is quartile from 1 to 4.

In short, the granularity level of the dataset is high and it includes many factors that may affect students' achievement such as teacher, student, school, and family factors [12]. However, the dataset has a limitation that the number of records is skewed. In a total of 12,411 records, there are 11,387 records that have a professional evaluation score of (or above) 130 and 1024 records have a score that is lower than 130. Consequently, researchers are required to select records with a professional evaluation score of 130 (or above) as model inputs.

3.2 Feature Selection

A feature selection process is necessary to eliminate irrelevant, noisy, or redundant attributes within the dataset, which essentially improves accuracy and interpretability of the final classifier [16]. A data sample often includes many different features, but the use of all features cannot provide a good result. Thus, a feature selection process is used to select relevant features of a corresponding problem.

Feature selection uses a ranking method to select features. This method then explores correlations among attributes to filter out independent ones. Feature selection has shown its efficacy in various areas including medical diagnosis, computer vision/image processing, text mining, bioinformatics, and industrial applications [18].

Therefore, we investigate a feature selection method toward predicting student performance. There are different ways of ranking features including the use of information entropy, correlations, Chi squared and Gini index. Entropy is a method to measure the uncertainty of a predictive variable. Let γ be a discrete random variable with two possible outcomes. The binary entropy function H, expressed in Eq. 1:

$$Gain(S, A) = Entropy(S) - \sum v \in Value(A) \frac{|Sv|}{|S|} Entropy(Sv) \tag{1}$$

where:

- Value(A) is the set of possible values for attribute A.
- S(v) is a subset of S where A takes the value v.

The dataset has 44 columns but the influence of each variable to final results is very different; hence, the researcher applied feature selection method, an effective dimension reduction technique removes noisy features [18]. Once a phenomenon of interest has been identified, a researcher should consider the phenomenon in question comprehensively and determine which features are likely to be salient before defining the constructs that represent these features. Choices about data collection and methods flow from these decisions [20]. The less variable, the more accuracy. Hence, instead of using all columns in the database, authors select specific variables that affect results. There is no single

work that carries out a comprehensive evaluation of the various multi-label classification techniques coupled with feature selection methods over data sets from different domains [20]. In this work, we have chosen the information gain measure, which is one of the most well-known measures for feature selection.

3.3 Decision Tree

Decision-tree is one of the most popular learning algorithms, due to various attractive features: simplicity, comprehensibility, and an ability to handle mixed-type data [11, 16]. This model as an intuitive form of data description that is easy to understand. The prediction consists of whether something will happen, or whether an item belongs in a category or not. The purpose of building a decision tree is to discover a set of rules that can be used to predict output values from input variables [13].

There are many classification algorithms such as ID3, J48, C4.5, CART (Classification and Regression Tree). Choosing an algorithm that has a high classification efficiency depends on many factors. For example, ID3 and CART algorithms generate high classification efficiency for numeric data fields (quantitative value) while algorithms like J48, C4.5 are more effective for qualitative data (ordinal, Binary, nominal) [3, 13].

ID3 is a simple decision tree algorithm introduced by Ross Quinlan in 1986 [11]. It is based on Hunt's algorithm. The idea of the ID3 algorithm is to explore set of rules by top-down through the given sets to test each attribute at each tree node. The tree is constructed in two phases: tree building and pruning. ID3 uses information gain measures to select splitting attributes. ID3 algorithm works only categorical attributes. It does not have a good result when there is noise within the data. To remove the noise, data processing technique has to be used. C4.5 algorithm is developed by Quinlan Ross that generates decision trees which can be used for classification problems [12]. It is the improvement of the ID3 algorithm by dealing with both categorical and continuous attributes to build a decision tree. It is also based on Hunt's algorithm. To deal with continuous attributes, C4.5 calculates the attribute values into two partitions based on a predefined threshold. It also handles missing attribute values. It uses Gain Ratio as an attribute selection measure to build a decision tree. C4.5 removes the biases of information gain when there are many outcome values of an attribute.

Multiple-Linear Regression

A multiple-linear regression model provides a continuous response variable through linear combinations of predictor variables. The result of the multiple-linear regression method is an equation. The equation for the multiple-linear regression method has the same form as that for a simple linear regression but has more terms, as shown below:

$$y_i = \beta_0 + \beta_{1.x_1} + \beta_{2.x_2} + \beta_{3.x_3} + \beta_{i.x_i} + \varepsilon \tag{2}$$

In a simple case, $\beta 0$ is a constant – which is a predicted value of y. In a model, $\beta 1$, $\beta 2$, $\beta 3$, … βn are coefficients, while $x1$, $x2$, $x3$ … xn are observed variables [22].

3.4 Model Evaluation

The performance of classification algorithms was determined in the study. Depending on the type of algorithms, we evaluate performance differently. Performance of the decision

tree algorithm was based on the four standard evaluation metrics including confusion matrix, accuracy, sensitivity, and specificity. About multiple linear regression, a combination of metrics, including but not limited to the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), are often required to assess model performance [15]. Therefore, RMSE and MAE are considered in this research.

Confusion Matrix
Confusion matrix is used to analyze outcomes of a supervised training, where each column of the matrix represents the number of instances in a predicted class, while each row represents the number of instances in an actual class, as shown in Table 1.

Table 1. Sample confusion matrix.

	Actually positive (1)	Actually negative (0)
Predicted positive (1)	True positive (TP)	False positive (FP)
Predicted negative (0)	False negative (FN)	True negative (TN)

Based on the confusion matrix, we have the following metrics:

Sensitivity
The recall (TP) is the proportion of positive cases that were correctly identified, as calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{d}}{(d + c).} \tag{3}$$

Specificity
The TN rate is the proportion of negatives cases that are classified correctly, as calculated by the following equation:

$$\text{Specificity} = \frac{\text{a}}{(a + b).} \tag{4}$$

Accuracy
The accuracy (AC) is the proportion of the total number of predictions that is correctly classified. It is determined by the following equation:

$$\text{AC} = \frac{(d + a)}{\text{d} + \text{a} + \text{b} + \text{c}a} \tag{5}$$

Root Mean Squared Error (RMSE)
Error evaluation indices include the root mean squared error (RMSE) and mean absolute

error (MAE). Regarding RMSE, the mean square of residuals of estimated and actual values is calculated and the outcome is taken a square root [22].

$$\text{RMSE} = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\left(y_i - \hat{\gamma}_i\right)^2} \tag{6}$$

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\gamma}_i\right)^2 \tag{7}$$

4 Results

In this section, we describe the results in more detail: (i) pruning input data based on the information gain method, (ii) applying a decision tree and a multiple-linear regression equation, (iii) deploying the prototype.

4.1 Feature Selection

A feature selection process is necessary to remove irrelevant, noisy, or redundant attributes in the dataset. This can improve accuracy and interpretability of the final classifier. In wrapper methods, the performance of a learning algorithm is used to assess the quality of a subset of features [18]. The information gain measure, which is one of the most well-known measures for feature selection, has been chosen for this research.

The dataset has 44 columns but the influence of each variable to the outcome is different; hence, researchers applied an attribute-ranking algorithm which is an effective dimensionality reduction technique during data pre-processing stage [18]. Feature selection aids researchers to retrieve important information of the dataset. The result of attribute ranking is described in Table 2.

Table 2. Attribute ranking.

Ranked	Attribute	Ranked	Attribute
0.41	G-SC	0.094	MAT-S11
0.40	CR-PRO	0.092	WC-PRO
0.146	QR-PRO	0.897	ENG-S11
0.143	CC-PRO	0.815	FEP-PRO
0.13	ENG-PRO	0.772	REVENUE
0.117	BIO-11	0.686	EDU-MOTHER
0.102	CC-11	0.015	EDU-FATHER

Tables 2 shows the top 14 attributes which affect final results. The top 7 attributes in the dataset that have the most influence toward the professional evaluation is the academic result of a student. The next 7 attributes are the standard living conditions and the demographics of a student, those attributes have certain influence toward the professional evaluation. The last 30 attributes have no correlations with professional evaluation have been discarded. As a result of the feature selection process, the 14 selected attributes will be used as inputs of the decision tree to classify students into two groups as well as inputs of the multiple regression model to predict student performance in each group.

4.2 Decision Tree and Multiple Linear Regression Model

Decision Tree

The decision tree model divided the dataset into two groups: “yes” and “no”. Based on historical data, the attributes will be calculated the probability of reaching professional evaluation from 130. The classification results showed that the G_SC attribute is the only attribute that determines the classification result. Students with an average score of more than 130 are generally considered to be students with high graduation possibility regardless of other features such as the student demographic characteristics or the previous educational background. Table 3 reveals the accuracy of the model.

Table 3. Accuracy of decision tree model.

	TP	FP	Precision	Recall
	0.99	0.000	1	0.999
	1	0.000	0.985	1
Weighted AVG	0.99	0.000	0.999	0.999

With classification models, we use PRECISION and RECALL to measure the accuracy. PRECISION evaluates how many correct ones are taken out. RECALL evaluates how many records which are taken out are correct. These metrics are also known as coverage. Hence, with Precision $= 1$ and Recall $= 0.99$, the decision tree model has a high accuracy. Table 4 shows the confusion matrix of the decision-tree model.

The confusion matrix represents how many data points belong to a class, and how many points were incorrectly predicted. In Table 4, 11371 records were classified as “Passed” and the actual data also shared the same result. Only 16 records were labelled as “Failed” but in fact they are “Passed”. 1024 students were failed and all of them have been identified correctly by the Decision Tree algorithm. However, previous research addressed that the student completion possibility could be influenced by other factors [17]. Accordingly, the researcher applied the multiple linear regression to 11371

Table 4. Confusion matrix.

	Actually positive (1)	Actually negative (0)
Predicted positive (1)	11371	0
Predicted negative (0)	16	1024

records to reinforce the point of view that information about demographics and previous education background also have an important influence on students' performance.

Multiple Linear Regression

In predicting student outcomes, the problem usually has only two sets of outcomes (i.e. "good" and "bad") by using binary logistic regression. Multiple linear regression refers giving more than 2 outcomes (i.e. "good", "near-good", "potential"). The result is predictive quartile from 1 to 4. Students are predicted in the quartile 4 classifying as "good". Students having the results from 3 to under 4 are "near-good". Students having the results from 2 to 3 are "potential" students.

Using the multiple linear regression, we find out those students having a car gain higher the professional evaluation, QR_PRO (quantitative reasoning) score and ENG_PRO (English in engineering) score is very likely to achieve high professional evaluation at the end of the study program. Students with high CR_PRO (critical reading) score, CC_PRO (citizen competencies) score and WC_PRO (written communication) score slightly impact high professional evaluation at the end of the study program. Meanwhile, Mathematic, English, and critical reading have negative effects toward professional evaluation. The specific equation is described as follows:

$$\begin{aligned} Y = {} & 0.0118 * CAR'Yes - 0.0032 * MAT_S11 - 0.0011 * CR_S11 - 0.0058 \\ & * ENG\_S11 + 0.0111 * QR\_PRO + 0.0085 * CR\_PRO \\ & + 0.0084 * CC\_PRO + 0.0106 * ENG\_PRO + 0.00071 \\ & * WC\_PRO + 0.0046 * G\_SC - 0.0007 * FEP\_PRO \\ & + 0.1909 \end{aligned} \tag{8}$$

The relative impact indicates that for every unit increase in having a car there is a 0,0118 increase in student achievement. The model summary shows that the simultaneous multiple linear regression was conducted. The QR-PRO, CR-PRO, CC- PRO, ENG-PRO, WC-PRO AND G-SC score has a significant impact on student achievement. On the other hand, The MAT-S11, CR-S11 and Eng-S11 scores have a negative impact on student academic. Simply put, the better living condition (having a car) gives student chance to have higher scores.

The relative abilities of 2 dimensioned statistics - the root mean square error (RMSE) and the mean absolute error (MAE) to describe average model-performance error are examined [14] and [21]. Table 5 shows Multiple linear regression accuracy model.

Table 5. Multiple linear regression accuracy.

Correlation coefficient	0.912
Mean absolute error	0.3126
Root mean squared error	0.3878
Relative absolute error	0.378631
Root relative square error	0.396151
Total number of instances	12411

Table 5 shows multiple linear regression accuracy. RSME is 0.38 and MAE is 0.31. The RMSE and MAE is a measure how well our model performed. It is commonly accepted that the lower the RMSE the better the model performance. Hence, RSME and MAE value have proved multiple linear regression model is acceptable.

4.3 Decision Support System

In this section, we present a prototype of DSS for predicting student performance (Fig. 1). The tool has been developed in Python and deployed as an opensource package. The prototype has been going through three stages of the development cycle including: design stage, build and testing stage, and deployment phase.

In the design stage, we create an architecture of the DSS system including:

- A Data entry UI allows users (students) to enter learning performance data and demographics data.
- A module stores algorithms and rules to score applications (submitted by users).
- A UI to present final outcomes (recommendations).

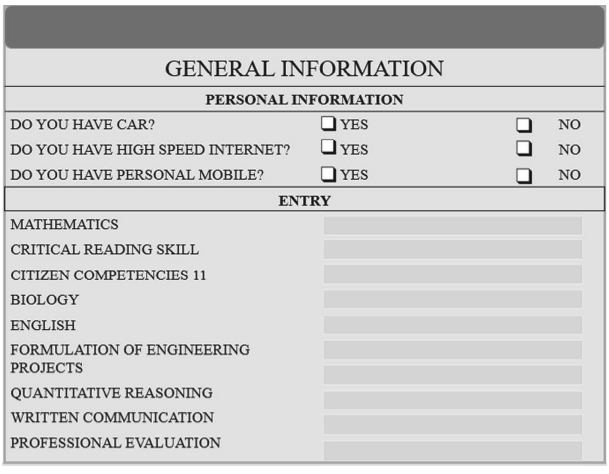

Fig. 1. Interface of our proposed prototype.

In the build and testing stage, we focused on designing a complete business workflow starting from the moment that users enter data into the system until the system returns desired results. Based on the workflow, we removed unnecessary components before proceeding to develop the solution (coding). Finally, we performed unit test to ensure the prototype to operate as expected.

In the deployment stage, we installed the prototype in a Window 10 (64 bit) environment and we embedded a csv file that stores student access information to the tool. In the production scenario, the tool should be integrated with the database of the university in order to authenticate the users. After users have been logging in into the tool, they will be able to enter study performance data and demographic data then click submit. We assumed that a student must completed at least 8 courses before applying for scholarship otherwise the tool cannot score properly. All the data will be grouped and treated as an application that apply for scholarship.

The data flows from the front-end to the module that stores algorithms to score and generate the final outcomes. At the first stage of the scoring process, the decision tree algorithm is applied to classify the student into two groups: potential group and none potential group. If a student belongs to a potential group, the system will run multiple linear regression to give exactly score. The results could be used as a reference for the school to consider whether to grant the scholarship to that student or not.

In the future with additional development resources, the researcher hopes to upgrade this tool by adding more advanced modules and functions such Administrative UI for university's staffs, interfaces to integrate this tool with university databases to retrieve necessary data and so on.

5 Conclusions

In this paper, we focused on building a model to predict the final grade of students. Firstly, understanding factors that affect student learning performance is crucial, especially those factors are different between datasets. The present study provides insight into educators among different standards of living and academic outcomes. In recent years, scholars often receive subsidiary scholarships or sponsorships, unfortunately these funds are no longer sufficient. According to Maslow hierarchy, people need to be satisfied physiological needs in order to pursue higher levels. In short, students can achieve higher results if they are comfortable while studying. Secondly, we have built a prototype for predicting student performance, based on demographics, living conditions and scores. The prototype is scalable and can adapt with various contexts with different types of input data.

The research, however, has limitations. Firstly, the training dataset is collected at Columbia Institute and hence the results are not well-generalized in a different environment. Secondly, the dataset is based on a technical and engineering program. Finally, the dataset is skewed due to a lack of data on students who are unlikely to graduate. This suggests a future direction of this study.

The use of different machine learning methods can be an extended research potential. For example, clustering and artificial neural networks can be used as alternative methods to analyze student performance. This is also of great interest in our future work.

Acknowledgment. The study is supported by a project lead by Prof. Tran Khanh Dang at the Department of Science and Technology, Ho Chi Minh City, Vietnam (contract with HCMUT No. 42/2019/HD-QPTKHCN, 11/7/2019). We also thank Mai Tan Ha (National Taiwan University, Taiwan) and all members of the AC Lab and D-STAR Lab (HCMUT) for their valuable support and comments during the preparation of this paper.

References

1. Adams Becker, S., Estrada, V., Freeman, A., Johnson, L.: NMC horizon report: 2017 Higher education edition. Austin, Texas: The New Media Consortium (2017)
2. Ahmed, M.H.E., Eltayeb, M.M.: Building Credit Risk Scoring Model for A Sudanese Bank Using Data Mining Techniques. Master's thesis, University of Science & Technology (2017)
3. Alexeyev, A., Solianyk, T.: Decision-making support system for experts of penal law. In: Ageyev, D., Radivilova, T., Kryvinska, N. (eds.) Data-centric business and applications. LNDECT, vol. 42, pp. 163–182. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-35649-1_8
4. Ashraf, A., Anwer, S., Khan, M.G.: A Comparative study of predicting student's performance by use of data mining techniques. Am. Sci. Res. J. Eng. Technol. Sci. (ASRJETS) **44**(1), 122–136 (2018)
5. Borkar, S., Rajeswari, K.: Predicting students' academic performance using education data mining. Int. J. Comput. Sci. Mob. Comput. **2**, 273–279 (2013)
6. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci. Model Dev. **7**(3), 1247–1250 (2014)
7. Delahoz-Dominguez, E., Zuluaga, R., Fontalvo-Herrera, T.: Dataset of academic performance evolution for engineering students. Data Brief **30**, 105537 (2020)
8. Sujatha, G., Sindhu, S., Savaridassan, P.: Predicting students performance using personalized analytics. Int. J. Pure Appl. Math. **119**, 229–238 (2018)
9. Ibrahim, Z., Rusli, D.: Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In: 21st Annual SAS Malaysia Forum (2007
10. Dasgupta, J.: Imparting hands-on industry 4.0 education at low cost using open source tools and python eco-system. In: Patnaik, S. (ed.) new paradigm of industry 4.0. SBD, vol. 64, pp. 37–47. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-25778-1_3
11. Jin, C., De-Lin, L., Fen-Xiang, M.: An improved ID3 decision tree algorithm. In: 4th International Conference on Computer Science & Education, pp. 127–130. IEEE (2009)
12. Killian, S.: Hattie's 2017 updated list of factors influencing student achievement (2017). http://www.Evidencebasedteaching.org. au/hatties-2017-updated-list
13. Dole, L., Rajurkar, J.: A decision support system for predicting student performance. Int. J. Innov. Res. Comput. Commun. Eng. **2**(12), 7232–7237 (2014)
14. Li, J.: Assessing the accuracy of predictive models for numerical data: not r nor r2, why not? then what? PLoS ONE **12**(8), e0183250 (2017). https://doi.org/10.1371/journal.pone.0183250
15. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**, 491–502 (2005)
16. Livieris, I.E., Mikropoulos, T.A., Pintelas, P.A.: Decision support system for predicting students' performance. Themes Sci. Technol. Educ. **9**(1), 43–57 (2016)
17. Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., Reber, S.: Descriptive Analysis in Education: A Guide for Researchers. National Center for Education Evaluation and Regional Assistance (NCEE), 4023 (2017)

18. OECD. How many students complete tertiary education? Education at a Glance 2019 (2019). https://doi.org/10.1787/62cab6af-en
19. Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.: Information gain feature selection for multi-label classification. J. Inf. Data Manag. **6**, 48–48 (2015)
20. Price, T.: Big data, dashboards, and data-driven educational decision making. IGI Global (2017). https://doi.org/10.4018/978-1-5225-1049-9.ch091
21. Thiele, T., Singleton, A., Pope, D., Stanistreet, D.: Predicting students' academic performance based on school and socio-demographic characteristics. Stud. High. Educ. **41**, 1424–1446 (2016)
22. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82 (2005)
23. Xuan, N.H.H.: Fintech in Education: Credit Scoring System for MIS Students. Master's thesis, HCMC University of Technology, VNU-HCM, Vietnam (2020)
24. Yang, S.J., Lu, O.H., Huang, A.Y., Huang, J.C., Ogata, H., Lin, A.J.: Predicting students' academic performance using multiple linear regression and principal component analysis. J. Inf. Process. **26**, 170–176 (2018)
25. Zeide, E.: The structural consequences of big data-driven education. Big Data **5**(2), 164–172 (2017)