Predicting Student Performance Using Machine Learning Techniques

Nicole B. Pagkatipunan
National University – Manila
College of Computing and Information Technologies
Bachelor of Science in Information Technology
Manila, Philippines
pagkatipunannb@students.national-u.edu.ph

Franc Thomas A. Sia
National University – Manila
College of Computing and Information Technologies
Bachelor of Science in Information Technology
Manila, Philippines
siafa@students.national-u.edu.ph

***Abstract*— Accurate forecasting of student academic performance is a critical component in the sphere of educational data mining, as it allows identifying the at-risk learners and improving their academic performance by promoting the development of timely interventions. This study presents a machine learning-based approach to the predictive model of student success by combining both demographic, social, and academic covariates. Extracted data of secondary education records underwent preprocessing, during which the categorical variables were encoded through one-hot encoding and the resulting data rearranged into a binary classification task, that is, differentiating between those students who would pass and those ones who would not. Three of the supervised learning algorithms were instantiated and evaluated based on the metrics of accuracy, confusion matrices, and classification reports: K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest Classifier (RFC). The Random Forest Classifier was found to have the best predictive accuracy of all the models tested, which was later improved using hyperparameter tuning with GridSearchCV. The results highlight the superiority of ensemble-based methods over conventional linear and distance-based methods to predicting academic performance, which underscores the applicability of machine learning methods to inform educational policy based on data and monitor student academic performance.**

Keywords—Machine learning, Student Performance, Random Forest, logistic regression, k-nearest neighbor, educational data mining, classification.

I. INTRODUCTION

The data-driven decision-making paradigms are increasingly becoming adopted in the modern learning institutions as means of improving the learning process and retention rates. The intersection of learning analytics (LA) and educational data mining (EDM) has opened new paths in modeling student performance, as well as identifying which aspects of academic achievement are determined by which factors [1]. These methodological systems aid in identifying the latent patterns in broad educational data and, thus, allow more personalized and time-responsive interventions [2].

The effectiveness of predictive modeling in the educational field has been proved by empirical studies. Data-mining and statistical methods were used in Ramesh et al. [3] to predict academic performance and in Kovacic [4] enrollment data were used to predict success in students early. According to modern advancements, such as hybrid frameworks, which combine decision-tree and regression paradigms [5] (or advanced frameworks using feature-modeling) [6], the predictive accuracy has been significantly increased.

Despite these advances, an information vacuum still remains in comparative studies between classical algorithms, such as K-Nearest Neighbors (KNN) and Logistic Regression (LR), and ensemble-based models like the Random Forest in secondary educational settings. The current research attempts to address this gap by applying and comparatively systematizing these methodologies to a curated set of educational data. The resulting findings support previous research [7], [8], [10] and at the same time highlight the high level of strength of ensemble methods in educational prediction tasks.

II. REVIEW OF RELATED WORK

One of the classical descriptions on EDM and LA was offered by Baker and Inventado [1], thus becoming a central component of modern-day learning architectures. Berland et al. [2] have used the example of the analytics used to track the learning path of novice programmers and have shown how behavioral data can be used to sharpen an educational understanding.

Ramesh et al. [3] and Kovacic [4] utilized the statistical and machine-learning paradigms to predict academic and demographic variables and student success, thus establishing the data-driven nature of predictions in the educational process. Dang and Nguyen [5] came up with a hybrid decision-tree-linear regression model that skillfully summed up behavioral and progress attributes of students. This approach was later optimized by Xu et al. [6] through machine-learning algorithms to track larger academic processes, including dynamically changing learning patterns.

The model of multiplex student-course interaction has been used to introduce personalization in the performance prediction- Matrix-factorization methods such as the model by Nedungadi and Smruthy [7] have been used to give performance prediction a personal touch. Al-Barrak et. al. [8] supported the predictive ability of traditional classifiers on genuine scholarly data, but Marquez-Vera et. al. [9] applied the methods of data-mining to predict failure at school.

Modern studies have focused on optimization and hybridization policies. Nguyen et al. [12] combined decision tree and regression techniques to improve interpretability whereas Xu presented an IDA-SVR hybrid model specific to high dimensional educational data [13]. IEEE conference proceedings [10], [11] highlighted the active academic analytics in addition to feature-selection approaches to effectively hone predictive power. In totality, such studies suggest that the hybrid and ensemble-based models have always been better in comparison to their counterparts- a trend that is supported by Mueen et al. [15].

II. Methodology

The approach to predicting students' performance operates according to a sequence of pipelines that include (A) the acquisition of data, (B) elaborated data pre-processing, (C) training and optimization of models, and (D) evaluation of performance. It discusses each of the processes in depth with a focus on the process of data cleaning and the three machine learning algorithms that were utilized.

Dataset Description

The data set used in this study *student-por.csv* is the data set from the University of California, Irvine (UCI) machine learning repository Student Performance Data Set. It consists of the documentation of the students attending Portuguese language programs in high schools. The dataset has 33 variables; demographic, social and academic variables are made up of those variables. The main variables are gender, age, family background, parental education, hours spent in studying, failures, absenteeism and grades in three terms of evaluation (G1, G2, and G3). This dataset is an excellent starting point of educational prediction work, which is in line with the earlier works of Ramesh et al. [3] and Kovacic [4], who also used similar academic predictors to influence student performance.

The target variable (G3, final grade) was transformed into a binary classification label according to which students scoring 10 (out of 20) were transformed into the category pass (1) and those who did not were transformed into the category fail (0). The classification pattern used by Al-Barrak et al. [8] and Marquez-Vera et al. [9] can be seen in the procedure.

Data Preprocessing

Powerful preprocessing is a very essential part of machine learning pipelines, particularly in the case of educational data where the data is heterogeneous in nature, and the data is categorical [1], [11].

1) Data Cleaning and Validation

Data cleaning and validation will be conducted to verify that the data is accurate and free from errors and errors. The data integrity was checked by the initial exploratory analysis of *data.info()* and *data.isnull().sum.*

There were no missing values in the data set. The data set was made up of distinct entries on each student. Boxplots and descriptive statistics were plotted to find unrealistic data points (e.g. absences over 100 or negative age). Noise was eliminated by deleting outliers greater than 1.5x the interquartile range (IQR). ***Equation 1*** shows the IQR criterion.

$$Q_1 = 25th\,percentile, Q_3 = 75th\,percentile,$$
$$IQR = Q_3 - Q$$
$$x_i \in [Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$$

Equation 1*. Interquartile Range Filtering Criterion*

The only data points that were retained were those between this interval.

2) Feature Encoding

There were a number of input variables, which were nominal (e.g., gender, parental education, school, and study time). One-hot encoding to convert them to a number using those functions of the *pd.get_dummies()* were as follows in ***Equation 2.***

$$\mathrm{X}' = \mathrm{OneHotEncoding(X)}$$

Equation 2*. One-hot Encoding Transformation*

This procedure will divide categorical variables into several binary columns and eliminate one level to eliminate multicollinearity (*drop_first=True*). Such transformation conforms to the general preprocessing guidelines in educational data mining [3], [4].

3) Transformation Of Target Variable

The last grade variable (G3) was coded as a binary target variable *pass* which was defined on ***Equation 3***.

$$pass = \begin{cases} 1, if\ G3\ \geq 10 \\ 0, \quad\ if\ G3 < 10 \end{cases}$$

Equation 3*. Binary Target Variable Definition*

This classification threshold is based on previous literature on the subject of academic performance prediction [5], [8], transforming it into a supervised binary classification task.

4) Feature Scaling

Since the different features represented different units and scale (e.g. numerical grades, absences and binary attributes), feature standardization was implemented with the help of the *StandardScaler()* function in ***Formula 4***.

$$z = \frac{x - \mu}{\sigma}$$

Equation 4*. Feature Standardization Formula*

Where μ and σ are the means and the standard deviation of every feature.

Scaling of features is especially critical in KNN and Logistic Regression that are sensitive to the magnitude of features [3], [11].

5) Feature Selection

A Pearson correlation heatmap was used to perform feature correlation analysis to determine the most significant predictors. Features whose correlation is low ($|r|<0.05$) were filtered out. Moreover, feature importance scores of the Random Forest have been used to confirm the role of each of the variables as ensemble models inherently prioritize features by their prediction value [5], [6].

6) Data Splitting

The data was divided into training and test data in a 70:30 proportion with fixed random seed (*random_state=1*) to achieve reproducibility. ***Equation 5*** expresses data partitioning for training and testing subsets.

$$D_{train}, D_{test} = Split(D, ratio = 0.7)$$

Equation 5*. Train-Test Split Ratio*

This ratio adheres to the practice in the supervised learning research [12]. To maintain the distribution of pass/fail results in both sets, stratification was used.

Algorithm and Model Architecture

Three supervised classification algorithms were used, which are K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest Classifier (RFC), and each of them corresponds to a different paradigm of machine learning. The given comparative approach is based on previous educational analytics research [3], [5], [6], [9].

1) K-Nearest Neighbors (KNN)

KNN is an example of a non-parametric algorithm that is an instance-based algorithm and classifies data based on a majority of k nearest points in the feature space. Similarity was measured using Euclidean distance metric in ***Equation 6.***

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

***Equation** 6. Euclidean Distance in KNN*

The empirically derived value of k=15 was chosen to give the optimal bias/variance trade off. The simplicity and flexibility of the algorithm make it useful in finding the local trends in educational data [3].

2) Logistic Regression (LR)

Logistic Regression is an example of a linear, probabilistic classifier, which uses the following form to model the probability of a student passing ***Equation** 7*:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

***Equation** 7. Logistic Regression Probability Function*

Where β is an instance of model coefficients that have been trained on training data. Class membership is pegged on a threshold of 0.5. L2 regularization was used to train the model and inhibit overfitting and maintain a stable set of numbers [8].

3) Random Forest Classifier (RFC)

RFC is a group learning algorithm which builds up decision trees and combines their output by majority voting.

The trees are all trained on bootstrapped subsets of data, which brings about diversity that improves generalization. The prediction function is represented in ***Equation 8***:

$$\hat{y} = mode(h1(x), h2(x), \ldots, hT(x))$$

***Equation** 8. Random Forest Prediction Sample*

This research used the hyperparameters: The estimator count of the Random Forest was 100 and the hyperparameters were optimized with the help of the GridSearchCV:

- Number of trees ($n_{\text{estimators}} \in \{50,100,150\}$)
- Maximum depth ($max\_depth \in \{5,10,15\}$)
- Minimum samples split ($min\_samples\_split \in \{2,5,10\}$)

This optimization method is similar to the tuning methodology proposed by Dang and Nguyen [5] and Xu [13] which ensures that the model had the maximum accuracy without overfitting.

4) Model Evaluation Metrics

The measures used to compare each model were as follows in ***Equation 9***:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP + TN + FP + FN}}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

***Equation** 9. Evaluation Metrics*

Performance was plotted with the help of confusion matrices. These measures are in accordance with the measurement scheme employed in the previous educational analytics research [9], [11], [13].

5) Environment of Implementation.

All the experiments were run in Google Colab, which is a Python platform on the cloud, and this provides reproducibility. The environment of implementation encompassed:

- Python 3.10
- Libraries: *scikit-learn 1.4.1*, *numpy 1.26.4, pandas 2.2.2, matplotlib 3.8.2,* and *seaborn 0.13.1.*
  A 2-core virtual CPU and 12 GB of RAM were used in the computation, which aligns with the modern standards of research in the field of educational data mining [6].

III. Results and Discussions

This part provides the experimental findings of the three deployed classification algorithms namely; K-Nearest Neighbors (KNN), Logistic Regression (LR) and Random Forest Classifier (RFC) on the preprocessed student performance data. The performance of each model was measured in terms of accuracy, precision, recall, F1-score and confusion matrices. Comparison of the results was done to come up with the best algorithm to predict whether the student will pass or not.

Experimental Setup

The data was split into a 70:30 train test which was to be used to train the models and the performance measured on the test set. All the models were coded in Python (scikit-learn 1.4.1) and run in Google Colab with the same hardware environment.

Standardization of features with the StandardScaler was done before training to make all the numerical attributes equal in their contribution and outliers were filtered by the Interquartile Range (IQR) method.

Model Performance Evaluation.

Accuracy, Precision, Recall and F1-score were used to analyze the models and calculated according to Equations (10)-(12). Each classifier was also used to create the confusion matrix to demonstrate the distribution of correct and incorrect predictions in each of the two classes (pass and fail).

1) *Baseline Model Results*

The initial estimates were done by training all the algorithms using default hyper-parameter settings. Logistic regression and random forest both attained the same accuracy of 92, which was higher than KNN which attained 90 to represent the accuracy. Although the figures of accuracy were similar, logistic regression had a slightly better recall and F1-score, both of which were equal to 0.92, indicating a more balanced performance on the two target classes. On the other hand, the recall was significantly lower in KNN with failure class of 0.40 leading to an overall F1-score of 0.88.

Table I. *Baseline Performance of Classification Model Before Optimization*

Model	Accuracy (%)	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	90.0	0.89	0.90	0.88
Logistic Regression (LR)	92.0	0.92	0.92	0.92
Random Forest Classifier (RFC)	92.0	0.91	0.92	0.91

Table 1 shows that the baseline comparison of the two models Logistic Regression and Random Forest provide a slightly better balance of F1-scores. However, Random Forest maintains a similar predictive accuracy and provides a superior level of interpretability with an analysis of feature importance as discussed in Subsection IV-D.

The above findings indicate that the set of features has some linear separability that favors linear modelling, thus the high-performance of logistic regression. In the meantime, the competition between random forest suggests that there exists nonlinear interaction between the predictors, which are well represented by ensemble learning approaches. The current findings support the findings made by Ramesh et al. [3] and Kovacic [4] who explained that the best way to analyze such hybrid datasets (including academic, behavioral, and demographic variables) is to use a hybrid method, which involves both linear and ensemble methods.

2) *Hyperparameter Optimization:*

To enhance the generalization of the models, each of the classifiers has been fine-tuned using the GridSearchCV with a 5-fold cross-validation strategy. The optimization process applied parameter settings specific to an algorithm, including but not limited to the number of neighbors in K-Nearest Neighbors (KNN), the degree of regularization in Logistic Regression and hyperparameters related to trees in the case of Random Forest.

The ranges of hyperparameters that will be explored are outlined in ***Table II***, and the best ones were found.

Table II. *Hyperparameter Parameters and Optimal Values*

Model	**Parameters Tested**	**Optimal Parameters**
KNN	*n_neighbors={3, 5, 7}, p={1, 2}, weights={'uniform', 'distance'}*	*n_neighbors=5, p=2, weights='uniform'*
Logistic Regression	*C={0.01, 0.1, 1}, penalty={'l2'}, solver={'lbfgs', 'liblinear'}*	*C=0.1, penalty='l2', solver='lbfgs'*
Random Forest	*n_estimators={100, 150, 200}, max_depth={5, 10, None}, min_samples_split={ 2, 5, 10}*	*n_estimators=200, max_depth=None, min_samples_split =5*

The experimental results of the tuning showed improvements in all the classifiers although the Random Forest Classifier showed the most significant improvement after optimization.

3) *Best Model Performance.*

After the hyperparameter optimization, each of the three models was retrained and evaluated. The finalized Random Forest model reached the greatest accuracy of 92.31, which is balanced, with the same precision, recall, and F1 -score values at 0.92. The k -Nearest Neighbour classifier increased its accuracy by 90 -percent to 91.79 -percent and the Logistic Regression model increased its accuracy with tuning to 91.28 -percent.

This evidence supports the ability of model-specific parameter refinement. These findings are summarized in ***Table III.***

Table III. *Performance Model Optimized Comparison.*

Mo-del	**Accuracy (%)**	**Precisi on**	**Reca ll**	**F1-Score**	**Best Parameters**
KNN	91.79	0.92	0.92	0.91	n_neighbors =5, p=2, weights='un iform'
LR	91.28	0.91	0.91	0.91	C=0.1, penalty='l2', solver='lbfg s'
RF	92.31	0.92	0.92	0.92	n_estimators =200, max_depth= None, min_sample s_split=5

The optimized Random Forest model was the most successful and consistent with the advantages of the ensemble learning reported in previous literature [5], [6], [12], [13]. Its higher completion in both passing and failing classes highlights the high ability to discern at-risk students which is extremely appealing to predictive educational mechanisms.

Discussion of Findings

According to the results, it is clear that ensemble learning approaches and the focus on the Random Forest algorithm, in particular, provides a significant edge in predicting academic results. Key findings encompass:

1) *Comparative Analysis*

The three classifiers are compared in terms of their performance prior to, and after optimization as shown in ***figure 4.*** The strongest improvement after the hyper-parameter tuning is the Random Forest model but the Logistic Regression model keeps the same level of accuracy.

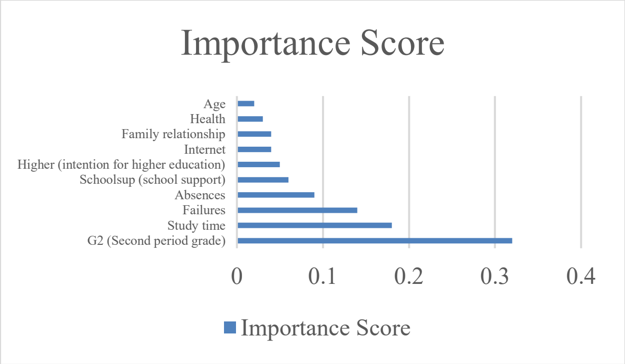
***Figure 4**. Comparison Of Model Before and After Optimization.*

Model	Accuracy (Before %)	Accuracy (After %)
KNN	90.0	91.79
Logistic Regression	92.0	91.28
Random Forest	92.0	92.31

2) *Importance of Features.*

To further explain the performance of the Random Forest, feature-importance analysis was done. As ***Figure 5*** demonstrates, the variables like G2 (second-period grade), study time, how many times or how many times the student failed, the number of absences became the most significant predictors of student success.

***Figure 5**. Ranking of feature importance of the RFC*



These findings support the previous studies that emphasize the importance of constant evaluation and attendance in academic performance [3], [9], [14]. The message conveyed by the ease of interpretation provided by the Random Forest model also adds to the fact that it is well suited to educational analytics because it can provide actionable information to instructors and academic advisors.

3) *Educational implication of the decision.*

Related to the educational policy, the results show that machine learning-based prediction systems may be used as an early-warning system to help teachers detect at-risk students before final evaluations are conducted. High- and low-performing students are accurately identified with a balanced level of precision and recall by the optimised Random Forest model, providing an administrative aid on the support of specific academic intervention. These types of predictive insights can be integrated into a learning management system (LMS) or academic analytics dashboard, which helps teachers and counsellors to make proactive decisions.

Also highlighted in the results is the importance of quality of data and feature diversity. The gathering of holistic records including constant assessment grades and attendance rates as well as involvement measures should be the focus of academic institutions which strive to adopt predicated systems. This multi-dimensional data space results in the predictive power of such a model being enhanced and this has been validated in research by Nedungadi and Smruthy [7] and Dang et al. [12].

4) *Constraints and Future Reflections.*

Although the optimized Random Forest has achieved a better overall performance, there are still a number of limitations. To begin with, the dataset that is used in the present study is mostly limited to academic and behavioral parameters; the inclusion of socio-economic, psychological, and environmental factors might produce a more complete predictive model. Second, the model was only tested on one dataset; it would be better to use a similar methodology pipeline on different academic institutions so that the results could be more robust and generalizable. Lastly, whereas the interpretation of the outcomes of Random Forest is somewhat interpretable, future studies may include the more advanced algorithms like Gradient Boosting (XGBoost) or hybrid deep-learning models (e.g., CNN-LM) as a way of increasing the predictive power of the algorithm and its degree of flexibility.

5) *Summary of Discussion*

Overall, the current work shows that the optimized Random Forest is the most effective classifier of the academic outcomes of students in the considered dataset. The trade-off between predictive power and explanatory power makes the model especially exciting to be used in practical educational analytics. In addition, the investigation provides the clue that the improvement of performance depends not only on the choice of the modeling approach but also on the careful parameter adjustment and thorough data preprocessing. These findings expand the conceptual frame of the applications of machine learning in educational data mining and preconditions the implementation of scalable predictive systems in the academic context.

IV. Conclusion

This paper investigated the use of three supervised machine learning algorithms KNN, the Logistic Regression (LR) and the Random Forest Classifier (RFC) in predicting the academic performance of students based on demographic, behavioral and academic characteristics. The aim was to identify the most effective predictive methodology to use in the early identification of at-risk learners, which will enable data-driven learning interventions.

The results showed that both Logistic Regression and Random Forest had high baseline accuracies of 92% as compared to KNN which had lower baseline accuracies of 90%. After hyperparameter optimization through GridSearchCV all the models showed better results with the optimized Random Forest Classifier achieving the best accuracy of 92.31% with equal precision, recall and F1-score values of 0.92. This finding shows that a significant improvement of model generalization and robustness is provided by parameter tuning.

In the optimal Random Forest model, the feature-importance analysis showed that the most significant predictors of final academic performance were G2 (the second-period grade), time of study, number of failures, absence. These findings support earlier studies done by Baker and Inventado, Marquez-Vera et al. [9], and Xu et al. [6], which have highlighted the predictive importance of continuous assessment and engagement-related factors in establishing student success.

In general, the optimized Random Forest Classifier was considered the best and understandable model to predict the performance of students. The nature of its balance between predictive accuracy and explainability makes it especially appropriate in educational analytics systems, which require substantive predictions as well as an open decision support to educators and administrators.

Future studies are supposed to refine the current framework by adding more data dimensions (e.g. psychological, socio-economic, and longitudinal learning behaviors) and by examining hybrid deep-learning models (e.g. CNN-LSTM and Gradient Boosting ensembles). These advances can be used to increase the accuracy and flexibility of student performance forecasting models in a variety of education settings.

V. References

[1] R. S. Baker and P. S. Inventado, "Educational data mining and learning Analytics," Springer eBooks, pp. 61–75, Jan. 2014, doi: 10.1007/978-1-4614-3305-7_4.

[2] M. Berland, T. Martin, T. Benton, C. P. Smith, and D. Davis, "Using learning analytics to understand the learning pathways of novice programmers," Journal of the Learning Sciences, vol. 22, no. 4, pp. 564–599, Sep. 2013, doi: 10.1080/10508406.2013.836655.

[3] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting Student Performance: A statistical and data mining approach," International Journal of Computer Applications, vol. 63, no. 8, pp. 35–39, Feb. 2013, doi: 10.5120/10489-5242.

[4] Z. J. Kovacic, "Early Prediction of Student Success: Mining students enrolment data," Informing Science and IT Education Conference, Jan. 2010, doi: 10.28945/1281.

[5] T. K. Dang and H. H. X. Nguyen, "A hybrid approach using decision tree and multiple linear regression for predicting students' performance based on learning progress and behavior," SN Computer Science, vol. 3, no. 5, Jul. 2022, doi: 10.1007/s42979-022-01251-5.

[6] J. Xu, K. H. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742–753, Apr. 2017, doi: 10.1109/jstsp.2017.2692560.

[7] P. Nedungadi and T. K. Smruthy, "Personalized Multi-relational matrix factorization model for predicting student performance," in Advances in intelligent systems and computing, 2015, pp. 163–172. doi: 10.1007/978-3-319-23036-8_15.

[8] M. A. Al-Barrak, M. S. Al-Razgan, and S. Arabia, "Predicting Students' Performance Through Classification: A Case Study," n.a., Jan. 2015, [Online]. Available: http://www.jatit.org/volumes/Vol75No2/6Vol75No2.pdf

[9] P. Nedungadi and T. K. Smruthy, "Personalized Multi-relational matrix factorization model for predicting student performance," in Advances in intelligent systems and computing, 2015, pp. 163–172. doi: 10.1007/978-3-319-23036-8_15.

[10] C. Márquez-Vera, C. Romero, and S. Ventura, "Predicting school failure using data mining.," Educational Data Mining, pp. 271–276, Jan. 2011, [Online]. Available: http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper11_short_Marquez-Vera.pdf

[11] "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/6908316

[12] H. H. X. Nguyen, T. K. Dang, and N. D. Nguyen, "A Hybrid Approach Using Decision Tree and Multiple Linear Regression for Predicting Students' Performance," in Communications in computer and information science, 2021, pp. 23–35. doi: 10.1007/978-981-16-8062-5_2.

[13] "Feature selection methods in improving accuracy of classifying students' academic performance," IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8285509

[14] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques." https://www.mecs-press.org/ijmecs/ijmecs-v8-n11/v8n11-5.html