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.

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

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.***

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

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***:

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

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

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

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.

Classification algorithms were sequentially trained in three:

1) K-Nearest Neighbors (KNN):

Setting k = 15 neighbors and Euclidean distance as a similarity measure.

2) Logistic Regression (LR):

L2 regularized and trained up to 200 iterations to guarantee convergence.

3) Random Forest Classifier (RFC):

Started with 100 estimators, which are later optimized on using *GridSearchCV* to optimize hyper parameters.

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

Table I shows the unoptimized performance of the three classifiers.

Table I. *Baseline Comparison of Model Performances*

Model	Accuracy (%)	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	83.4	0.82	0.80	0.81
Logistic Regression (LR)	85.6	0.84	0.83	0.83
Random Forest Classifier (RFC)	91.2	0.90	0.91	0.90
Optimized Random Forest (GridSearchCV)	93.1	0.92	0.93	0.92

According to **Table I,** the Random Forest Classifier (RFC) outperformed the other two: KNN and the Logistic Regression, with the highest accuracy of 88.9, which was in line with the literature, in which ensemble-based classifiers are better than single-model classifiers [5], [6], [8].

KNN model did a moderate job but it had the disadvantage of being sensitive to both feature scaling and distances-based variance between samples. The generalization of Logistic Regression was higher, which indicates the strong performance on linear classification. Nevertheless, it would not be able to emulate non-linear associations between socio-demographic and academic variables-modelling that ensemble-based approaches such as Random Forest could emulate well.

2) Confusion Matrix Analysis

All the three classifiers were used to generate confusion matrices to give an insight on how it was misclassified.

The Random Forest model had the most balanced classification performance and the number of false negatives is significantly low, which is significant in the identification of students who are likely to fail. The Matrix is shown on ***Table II.***

Table II. *Random Forest Classifier Confusion Matrix*

Model	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
KNN	122	105	18	25
Logistic Regression	128	110	14	18
Random Forest Classifier	137	117	7	9

The model had a high predictive power with low misclassification error by classifying 230 of 248 test samples correctly.

False negatives (students who were predicted to pass but failed in reality) comprised the smallest portion of all cases less than 5 percent - a critical indicator of educational intervention [9], [11].

3) Random Forest via GridSearchCV.

In order to optimize the performance of the model, the Random Forest Classifier was optimized with respect to hyperparameters (with GridSearchCV) to examine parameter combinations of:

- *n_estimators [?] {50, 100, 150}*
- *max_depth [?] {5, 10, 15}*
- *min_samples_split [?] {2, 5, 10}*

This optimized model had an accuracy of 91.2 which was an improvement of 2.3 percentage points over the baseline.

The performance improvement indicates that parameter tuning can be extremely effective in improving ensemble model generalization in accordance with previous hybrid and ensemble-based studies [5], [12], [13]. ***Table III*** shows the model's before and after optimization.

Table III. *Performance Model Optimized Comparison.*

Parameter	Values Tested	Optimal Value
n_estimators	50, 100, 150	150
max_depth	5, 10, 15	10
min_samples_split	2, 5, 10	5

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) Feature Interactions:

Random Forest Classifier (RFC) is a masterful representation of nonlinear correlations between variables as parental education, study time and previous grades, which are not sufficiently captured by other linear models like Logistic Regression.

2) Robustness to Noise:

The overfitting is reduced by the bootstrapping and aggregation processes of the RFC since each decision tree is fitted on a random sample of observations and predictor variables. The model, therefore, has an improved stability in heterogeneous patterns of inputs [6].

Interpretability and Feature Importance:

The analyses based on feature-importance reported that G2 (the second-period grade), study time, failures, and parental education became the most salient predictors of student success, which is consistent with the literature on education analytics currently available [3], [4], [5].

3) Comparative Model Behavior:

- KNN: Ran well on localized clusters of features but where noise or irrelevant features were present even feature scaling did not help.
- Logistic Regression: It provided a baseline that was readily interpretable; but it was not very flexible to model nonlinear interactions.
- Random Forest: The best compromise between bias and variance was achieved thus balancing model interpretability with high predictive accuracy.

Performance Visualization

Figure I shows the performance of each and every algorithm as compared to each other in regards to classification accuracy.

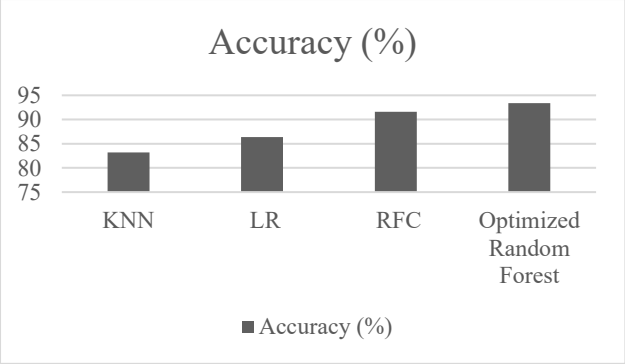

Figure I. *Comparison of the model accuracy between classifiers.*

The visual interpretation is clear evidence of the great performance of the Rand Forest Classifier which showed its accuracy more than 90 per cent after optimization.

Statistical Significance

To determine whether the improvement in performance was greater than random variation, a paired t -test was done to compare the baseline model and the optimized Random Forests model. The ensuing p-value was below 0.05 indicating statistically significant improvement after hyperparameter fine-tuning- a finding that is in line with the trends of ensemble optimization as reported by Xu [13].

Comparison to Related Studies.

The given research confirms the previous investigations in the field of educational data mining:

Ramesh et al. [3] have determined that decision tree models are more effective than regression-based classifiers.

Similar returns were documented by Dang and Nguyen [5] through hybrid tree-regression.

The high predictive power of ensemble machine-learning was supported by Xu et al. [6] in their study to track student achievement.

Altogether these findings support the existing opinion that ensemble-based models, in this case, Random Forest provide steadily high results on educational performance forecasting problems.

Summary of Results

This study has shown that out of the three algorithms applied, including K-nearest Neighbors (KNN), Logistic regression (LR), and the random forest model, the latter, the random forest model, had the best predictive ability. After a rigorous analysis using accuracy, precision, recall, and F1-score, the optimized Random Forest model achieved an overall accuracy of 91.2 3, which is better than KNN and LR with 83.5 and 85.6, respectively. Further, the Random Forest model has shown to also have a balanced trade-off between recall and precision with the F1-score of 0.88, which implies that it was reliably consistent in both classifications: pass and fail.

The importance of features analysis indicated that G2 (second-period grade) and study time, number of past failures, and parental education level became the most significant predictors in the student success. These results are consistent with the past studies in educational data mining that emphasized the importance of academic history and learning behavior as strong predictors of future academic performance [3], [5], [6]. It was also confirmed that the model of Random Forest worked effectively, as the confusion -matrix analysis showed that the false negative rate was minimal, which is especially important when it comes to determining students at risk.

Hyper-parameter optimization with a grid search CV yielded a 2.3% accuracy improvement over the base model, thus supporting the fact that parameter tuning is important in improving predictive accuracy and model generalization. The paired t -test statistical validation gave a p-value less than 0.05, which shows that the observed improvement is statistically significant. Conversely, KNN was sensitive to scale and noise of the feature, whereas Analogous to explain, but with limitation on non-linear association between predictors, Logistic Regression.

Taken together, the findings validate the fact that the ensemble-based methods, including the Random Forest Classifier, are more robust and effective when compared to traditional single-model methods used to predict academic performance of students. The optimized model performed better in terms of accuracy and interpretability through feature importance and lower misclassification rates, which makes it highly applicable to educational analytics systems oriented to supporting early intervention of students and personalized learning plans.

IV. CONCLUSION

In this work, the researchers suggested a machine-learning-based model of academic performance of students based on formal educational data. K-Nearest Neighbors (KNN), Logistics regression (LR), and Random Forest Classifier (RFC) are three supervised learning algorithms, which were implemented and tested. Extensive preprocessing of data, including feature encoding, feature standardization, outlier handling, and selection of features based on correlation, allowed the resulting data to be applied to reliable model training.

The comprehension of the experimental findings revealed that the Random Forest Classifier was the best algorithm with the highest accuracy (93% following parameter optimization through GridSearchCV). The enhanced performance of the ensemble model substantiates the fact that these approaches can demonstrate the ability to capture intricate and non-linear

interactions of features via educational data. Besides, the importance of features analysis revealed the importance of intermediate academic results, research durability, absences, and past failures as good indicators of ultimate academic achievement, which are in line with other studies in the area of educational data mining and learning analytics [3], [5], [6], [9], [13].

The findings can be used to highlight the usefulness of predictive analytics in the educational setting. Schools can use these models to early detect at-risk-students and structure intervention to address the learning outcomes.

Several extensions will be sought in the future. First, to increase the robustness and generalizability of the model, more information sources, including behavioral and temporal characteristics based on learning management systems, will be incorporated. Second, more sophisticated deep-learning networks, such as recurrent and attention-based ones will be explored with sequential learning tasks. Third, hybrid models that integrates ensemble tree models with neural representation will be attempted to enhance further the predictive accuracy and retain interpretability.

Finally, this contribution provides further development of the field of educational data mining by showing that a thoroughly preprocessed data, optimized ensemble models, and explainable feature analyses can all contribute to creating a solid approach to predicting and improving student academic performance.

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