Leveraging Regression-Based Machine Learning for Predicting Middle School Student Passing Grades

1st Fiona Maharani Nugraha
Data Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
fiona.nugraha@binus.ac.id

2nd Kimberly Kayla Dewi
Data Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
kimberly.dewi@binus.ac.id

3rd Alexander Agung Santoso Gunawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
aagung@binus.edu

4th Jeffrey Junior Tedjasulaksana
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
jeffrey.t@binus.ac.id

***Abstract*— Traditional methods of assessing student performance may be time-consuming and lack accuracy, therefore a modern method is needed since education is important for students and lecturers to enhance their academic quality. This research uses an ML model (KNN, Decision Tree, Linear Regression, and Random Forest) for predicting student passing grades using data on student demographics and academic performance. These models used regression techniques to predict the passing grades. The aim on this research is to compare and find which is the best ML model to predict the passing grade in regression schema. The final result shows that the Linear Regression models performs the best, achieving the lowest error rates (MSE = 4.8801, RMSE = 2.2091, MAE = 1.3436, MAPE = 3.29%) and the highest explanatory power (R^2 = 0.7620). Therefore, Linear Regression can be selected as the most effective model for predicting student passing grades based on academic performance data.**

Keywords—machine learning, student performance, passing grade, academic success, predictive analytics in education, educational data mining.

I. INTRODUCTION

Education is the backbone in the development of individuals and society, imparting skills, knowledge, and opportunities for personal and professional growth [1]. Academic achievement in secondary education is usually measured by the grades a student receives, which are key indicators of the level of understanding and mastery of subjects. Predicting student performance has been an important area of educational research, especially for enhancing teaching strategies, personalizing learning, and enabling early intervention for at-risk students [2]. The conventional ways of evaluating student performance, through standardized tests or teachers' evaluations, are usually plagued by subjective biases and often do not take complete consideration of all the complex variables that come into play for determining academic success [2]. It is due to this that modern approaches such as machine learning have become popular: they can handle huge volumes of data and make predictions about student performance more accurately.

Machine learning in education has promoted data-driven insights to enhance decision-making [3]. Students' data could be used to enhance machine learning models that identify patterns and relationships that may not be immediately obvious by the manual analysis of data. Many research studies suggest that, other than academic performance data, including social and demographic data with behavioral information leads to substantial improvements in the accuracy of predicted final grades of students. For instance, the models that factor in data from parental education [4], student attendance, and extra-curricular activities have shown great strength of predictions of students' achievement. This underlines the usefulness of using a wide range of variables other than academic scores to understand and predict student success. Probably the biggest challenge in conducting educational research relates to the multi-dimensionality of the factors affecting the performance of the student [5].

Family background, socio-economic status, and personal habits are the contributing elements towards the academic results of students. This research offers a comprehensive approach to consolidating previous research on the application of machine learning for predicting academic success. It synthesizes findings from multiple studies, giving a broader understanding of the field, identifying research gaps, and highlighting best practices in applying machine learning for educational purposes [6]. Therefore, this enable educators and policymakers to design better interventions aimed at helping such students and generally enhancing educational outcomes. Predicting student success in secondary education has especially long-term implications for consequences in the future course of higher education and career opportunities for students. For example, the work of Cortez and Silva [7] on the performance of students in Portuguese secondary schools upholds the need to consider both academic and social-demographic variables in predicting academic success.

This research justifies the role of predictive models in education, which will be very beneficial in predicting the outcomes of the performances of students academically. This

research brings together the current literature on the application of machine learning to predict student passing grades in secondary education. It seeks to answer a major research question: Which machine learning algorithms are most effective in predicting student passing grades in secondary education? By synthesizing the current state of research, this research will analyze methods, datasets, and variables used in prior studies. The performance of machine learning models in predicting student performance will also be reviewed to extend the current understanding of how data-driven predictions can benefit educational systems.

II. Literature Review

It is a part of artificial intelligence that enables computers to learn from data and make decisions or predictions without necessarily being explicitly programmed [8]. The capability of machines to improve their performance automatically, based on past experiences, through the identification of patterns in data, has made it an influential tool in predictive modeling and data analysis. Machine Learning algorithms can be divided into three broad categories: supervised, unsupervised, and reinforcement learning [9]. The focus of the study is on supervised learning, which involves training a model on a labeled dataset such that the model predicts an outcome on the basis of an input feature. Some of the common supervised learning algorithms are Decision Tree and Random Forest, K-Nearest Neighbors (KNN), and Linear Regression [10]. Decision Tree is one of the common algorithms used in ML; it splits data into branches based on feature value. KNN predicts K-Nearest Neighbors by taking a look at the nearest neighbors in feature space, while Linear Regression maps the relationship between input variables with continuous output. These algorithms have proven useful in education for the effective prediction of students' performances, behaviors, and outcomes [11]. Machine learning obtains insights from the data on how students learn and thereby supports educational strategy enhancement. Educational Data Mining (EDM) [12] makes use of machine learning and statistical methodologies to analyze data in education; these practices provide insights into the behaviors and performance of students.

With the growth of technology in education, large datasets from learning management systems and academic records are being collected. EDM helps predict student outcomes, identifies at-risk students, and personalizes learning [12]. The Student Performance Dataset [7] contains demographic, performance, and behavior data that are useful in building predictive models to assess factors that influence academic success. EDM has been used in the identification of predictors of success and to inform tailored interventions [13]. In educational data analysis, regression models predict continuous variables, such as students' grades, from several input features [14]. Linear regression models the relationship between a dependent variable and one or more independent variables, commonly used for predictions related to, for example, exam scores, GPA, or course completion rate outcomes [15]. These metrics offer a complete overview of the performance by showing both prediction accuracy and model generalization.

III. Material and Methodology

The research will progress through six stages: data collection, exploratory data analysis, feature selection, data preprocessing, model training and evaluation.

A. Data Collection

The dataset used for this research is titled "Student Performance Dataset". This dataset is obtained from Kaggle, uploaded by Dev Ansodariya who is a student at San Jose State University and an experienced software engineer [7]. The dataset contains information on student achievement in secondary education from two Portuguese schools: 'GP' (Gabriel Pereira) and 'MS' (Mousinho da Silveria). The data covers performance in two subjects: Portuguese and Mathematics. It includes student grades, demographic, social, behavioral, and school-related attributes, collected through school reports and questionnaires [7]. Examples of features are age, gender, parental education, study time, and health status. Since this research is a regression study, the following table is a description of the dataset with numerical variables.

TABLE I – *Dataset description with numerical features*

Features	Description
age	Student's age (numeric: from 15 to 22).
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education).
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education).
traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour).
studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours).
failures	Number of past class failures (numeric: n if 1<=n<3, else 4).
famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent).
freetime	Free time after school (numeric: from 1 - very low to 5 - very high).
goout	Going out with friends (numeric: from 1 - very low to 5 - very high).
Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high).
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high).
health	Current health status (numeric: from 1 - very bad to 5 - very good).
absences	Number of school absences (numeric: from 0 to 93).
G1	First period grade (numeric: from 0 to 20).
G2	Second period grade (numeric: from 0 to 20).
G3	Final grade (numeric: from 0 to 20, output target).

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to explore the relationships between features, summary statistics, and detect anomalies [16] which in this research, contain 395 rows and 33 columns. This dataset includes demographic, social, behavioral, school-related, and academic performance data for two subjects: Mathematics and Portuguese [7]. We will use only the numerical features since this research conduct a regression model with the total of 16 numerical columns. One of the goals is finding the strong correlation between the final year grade (G3) besides earlier period grades (G1 and G2), as G1 and G2 have a big influence on the final grade. The graph below shows the relationship between G2 and G3 by average.

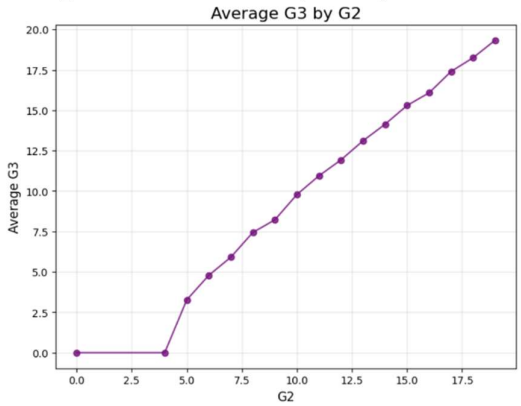

Fig. 1. Relationship between G2 and G3 by average.

Fig. 1 indicates that the higher the G2 score, the higher G3 score average. This relationship is nearly linear. Because of this finding, G2 is shows a strong correlation with G3 and serves a great predictor of the final grade (G3).

Additionally, this dataset is imbalanced, because some columns such as school, sex, address, and schoolsup have uneven frequency distributions. For example, in the school column, the amount of data for GP (349) is much larger than MS (46). This condition can affect model performance, especially in algorithms that are sensitive to class distribution. The target variable in this research is G3, which represents the final grade of students and serves as an indicator of students' academic success.

C. Feature Selection

Feature selection was first conducted using heatmap correlation between the features and the target label and selected the feature that has the absolute value from correlation score of <=1. Below is the correlation score with G3.

TABLE II – *Correlation score with G3*

Feature	Score	Abs_Score	Description
G2	0.9049	0.9049	Selected G2 reflects students' progress during the course
G1	0.8015	0.8015	Selected. Provides early student's progress into their performance
Medu	0.2171	0.2171	Selected. Resources and support a student from their mother
higher	0.1824	0.1824	Selected. Wants to take higher education
Fedu	0.1524	0.1524	Selected. Similar to Medu, but from their father
reason	0.122	0.122	Selected. Reason to choose the school
address	0.1057	0.1057	Selected. Student's home address type
sex	0.1034	0.1034	Selected. Student's sex
Mjob	0.1020	0.1020	Selected. Mother's job
paid	0.102	0.102	Selected. Extra paid classes within the course subject
internet	0.0985	0.0985	Not selected.
studytim	0.0978	0.0978	Not selected.
famsize	0.0814	0.0814	Not selected.
nursery	0.0517	0.0517	Not selected.
famrel	0.0513	0.0513	Not selected.
Fjob	0.0423	0.0423	Not selected.
absences	0.0424	0.0424	Not selected.
activities	0.0161	0.0161	Not selected.
freetime	0.0113	0.0113	Not selected.
famsup	-0.0391	0.0391	Not selected.
school	-0.045	0.045	Not selected.
Walc	-0.052	0.052	Not selected.
Dalc	-0.0547	0.0547	Not selected.
Pstatus	-0.058	0.058	Not selected.
health	-0.0613	0.0613	Not selected.
guardian	-0.0701	0.0701	Not selected.
schoolsup	-0.0828	0.0828	Not selected.
traveltime	-0.1171	0.1171	Selected. Home to school travel time
romantic	-0.13	0.13	Selected. Whether the student is in a romantic relationship
goout	-0.1328	0.1328	Selected. Going out with friends
age	-0.1616	0.1616	Selected. Student's age
failures	-0.3604	0.3604	Selected. Number of student's past class failures

From the scores shown in the table II, G1 and G2 (grades) have a strong correlation with G3 (final grades). Meaning that students that performs well in first and second period grades are more likely to have a good understanding of the material and srong study habits, which carry through to the final grades. Other numeric features, such as parental education levels (Medu, Fedu) are correlated with G3 although lower than G2, and some behavioral features (Dalc, Walc), show weaker correlation with G3.

The features with an absolute score of correlation values that are lower than or equal to 0.1 are removed from the model as it is not significantly important to the target (G3). In total, we will use 15 features in the dataset as shown in the table II. These features are G2, G1, Medu, higher, Fedu, reason, address, sex, Mjob, paid, traveltime, romantic, gout, age, failures. By focusing on these selected features, the model's complexity is reduced, improving both its

interpretability and performance. This guarantees that the data is properly structured for use with machine learning algorithms.

D. Data Preprocessing

Scaling and feature selection are crucial steps in preparing data for machine learning models to ensure optimal performance and accuracy [17]. In this research, numerical data were standardized using the StandardScaler from sklearn. This step ensures that all features are on the same scale, minimizing bias in models that are sensitive to differences in feature magnitude, such as Linear Regression and K-Nearest Neighbors. Standardization transforms the data to have a mean of 0 and a standard deviation of 1, making it easier for the models to converge during training.

There are no missing values in the dataset. However, outliers were identified in the columns age, G1, G2, and G3, which led to a reduction in the dataset from 395 rows to 341 rows after outlier removal. After preprocessing, the dataset was split into training and testing sets using an 80:20 ratio with train_test_split.

E. Model Training & Evaluation

Each selected machine learning model is trained on the training dataset to understand the relationships between the input features and the target variable (G3) in order to make predictions. The trained models are assesses using the testing set metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), Mean Absolute Percentage Error (MAPE) to evaluate the accuracy and reliability of their predictions. MSE and RMSE measure the average squared differences between predicted and actual values, providing insights into the magnitude of errors [18]. MAE quantifies the average magnitude of errors without considering their direction, offering a straightforward interpretation of prediction accuracy [18][20]. R-squared explains the proportion of variance in the dependent variable accounted for by the model, serving as an indicator of the model's overall explanatory power [18][20]. MAPE calculates the average percentage difference between predicted and actual values. This metric offers a normalized error measure, making it easier to compare models across datasets with different scales. However, it is less effective when the actual values are close to zero, as it can produce excessively high error percentages [19]. Together, these metrics provide a comprehensive framework for evaluating model performance, ensuring that predictions not only align closely with actual outcomes but also generalize effectively to new data [20]. The model training involved three machine learning algorithms: Linear Regression, K-Nearest Neighbors (KNN) with n_neighbors=5, Decision Tree Regressor, and Random Forest.

IV. RESULT

This research conduct experiments of implementing machine learning to a regression. Three different models were compared in this research to get the best performance in regression prediction. The four models were constructed by different machine learning models: Linear Regression, K-Nearest Neighbors (KNN), Decision Tree Regressor, and Random Forest. The following table shows the performance of the models.

TABLE III – *Regression Performance*

Model	Metrics				
	MSE	**RMSE**	**MAE**	**R^2**	**MAPE**
Linear Regression	4.8801	2.2091	1.3436	0.762	3.29%
KNN	8.3251	2.8853	1.9494	0.594	7.08%
Decision Tree	10.5696	3.2511	1.8101	0.4845	2.2%
Random forest	5.4219	2.3285	1.4828	0.7356	4.68%

As shown in the table III, Linear Regression model has the best performance than the other models in all metrics. It has lowest errors Mean Squared Error (MSE: 4.8801) and Root Mean Squared Error (RMSE: 2.2091) meaning that this model has the best fit to the dataset and minimal errors. This model has the lowest Mean Absolute Error (MAE: 1.3436), meaning its predictions deviate by only 1.34 points on average. Highest R-squared score (R^2 : 0.762) meaning that it explains 76.20% of the variance in the target variable, making it the most accurate model. Mean Absolute Percentage Error (MAPE: 3.29%) shows accurate predictions as a percentage of actual grades. K-Nearest Neighbors (KNN) has potential to improve as it is worse than Linear Regression and Random Forest but better than Decision Tree. This model has moderate MSE (8.3251) and RMSE (2.8853) that indicates greater errors. It has MAE of (1.9494) and relatively high MAPE (7.08%) shows that it is less precise compared to top model. This KNN model has lower R^2 score (0.5940). The Decision Tree model has the weakest performance as it has the highest errors with MSE (10.5696) and RMSE (3.2511), meaning there are significant prediction errors. Lowest R^2 score (0.4845) as it explains only 48.45% of the variance in final grade. Although it has MAE (1.8101) and MAPE (2.2%) that suggests occasional food predictions but overall less reliable performance. The Decision Tree model is the least suitable model for this research. The Random Forest model has a good performance, with MSE (5.4219) and RMSE (2.3285) that are reasonable but higher than Linear Regression's. This model has R^2 score of (0.7356) that explains 73.56% of the variance, MAE (1.4828), and MAPE (4.68%) as it shows decent predictive accuracy. Random Forest performs well but does not surpass the Linear Regression model.

Overall, Linear Regression has the best model due to its superior performance across all metrics. KNN and Random Forest performs reasonably well, with Random Forest is slightly better than Decision Tree. Decision Tree is the least effective model since it has the highest error rates and lowest overall accuracy.

V. CONLCUSION

This research helps to aid predict middle school student passing grades, using machine learning models with regression approach. Based on the regression performance metrics, Linear Regression has the best results with lowest errors (MSE, RMSE, MAE) but its MAPE is slightly lower than Decision Tree, and it has the highest R^2 compared to

other regression models, KNN, Decision Tree, and Random Forest.

For future work, we suggest approaching advanced regression techniques such as neural networks to capture complex patterns in the data, integrating additional data sources and fine-tuning hyperparameters to potentially achieve higher predictive accuracy.

AUTHOR'S CONTRIBUTION

K.K.D conceived the project, developed the key conceptual ideas, and outlined the proof. F.M.N lead and wrote the manuscript in consultation with A.A.S.G and J.J.T. K.K.D carried out the experimental work and implementation of the research. All authors have reviewed and approved the final version of the manuscript.

AVAILABILITY DATA AND MATERIALS

The Student Performance dataset utilized in this research can be accessed publicly at https://www.kaggle.com/datasets/devansodariya/student-performance-data. All associated code has also been made publicly available on GitHub https://github.com/kimikayz/Leveraging-Regression-Based-Machine-Learning-for-Predicting-Middle-School-Student-Passing-Grades.

REFERENCES

[1] M.K. Shomirzayev and K.K. Yuldashov, "The Educational Importance of Teaching Knowledge to Secondary School Students," Current Research Journal of Pedagogics, vol. 2, no. 8, pp. 132-142, Aug. 2021, doi: https://doi.org/10.37547/pedagogics-crjp-02-08-28.

[2] Z. Ibrahim, N. Omar, and M. Aziz, "Student Review," TEM Journal, vol. 10, no. 4, pp. 1919-1927, Nov. 2021, doi: 10.18421/TEM104-56.

[3] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," in Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008), Porto, Portugal, 2008. Available: https://www.kaggle.com/datasets/devansodariya/student-performance-data

[4] I.H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," SN Comput. Sci., vol. 3, no. 2, p. 158, Feb. 2022, doi: 10.1007/s42979-022-01043-x.

[5] Z.-H. Zhou, *Machine Learning*, Springer Nature, 2021.

[6] J. Alzubi *et al.* "Machine Learning from Theory to Algorithms: An Overview". *J. Phys.: Conf. Ser.* 1142 012012. 2018. doi: 10.1088/1742-6596/1142/1/012012.

[7] Z. Ibrahim, N. Omar, and M. Aziz, "Student Performance Prediction Using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 10, no.8, pp. 1-5, 2019.

[8] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics. 2020.

[9] Y.J. Wang, C.L. Gao, and X.D. Ye, "A data-driven precision teaching intervention mechanism to improve secondary school students' learning effectiveness," *Educ. Inf. Technol.*, vol. 29, pp. 11645–11673, Jun. 2024, doi: 10.1007/s10639-023-12238-x.

[10] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A Multiple Linear Regression-Based Approach to Predict Student Performance," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, M. Ezziyyani, Ed., vol. 1102, Advances in Intelligent Systems and Computing, Springer, Cham, 2020, doi: 10.1007/978-3-030-36653-7_2.

[11] J. Han, J. Pei, and M. Kamber. "Data Mining: Concepts and Techniques". Elsevier. 2021.

[12] N. Ekbote, P. Dhanshetti, and S. Sakhrekar, "Techniques of Exploratory Data Analysis," *Madhya Pradesh Journal of Social Sciences*, vol. 28, no. 2(v), p. 10, Dec. 2023, doi: 10.13140/RG.2.2.13578.03522.

[13] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[14] B.J. Erickson and F. Kitamura, "Magician's Corner: 9. Performance Metrics for Machine Learning Models," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200126, 2021, doi: 10.1148/ryai.2021200126.

[15] D.S. Moore, G.P. McCabe, and B.A. Craig, "Introduction to the Practice of Statistics," 9th ed. New York: W.H Freeman, 2017.

[16] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean Absolute Percentage Error for Regression Models," *Neurocomputing*, vol. 192, pp. 38-48, 2016, doi: 10.1016/j.neucomPerformance Prediction Using Machine Learning Techniques," International Journal of Advanced Computer Science and Applications, vol. 10, no.8, pp. 1-5, 2019.

[17] D. Ansodariya and P. Pathak, "Exploring the Relationship between Students' Engagement and Self-Regulated Learning: A Case Study using OULAD Dataset and Machine Learning Techniques," ResearchGate, 2023.

[18] D Claudia, N Paun, "The Parental Impact on Education: Understanding the Correlation between the Parental Involvement and Academic Results," Acta Educationis Generalis, vol. 14, no.2, pp. 16-26, May 2024, doi: 10.2478/atd-2024-009.

[19] F.S. Alani and A.T. Hawas, "Factors Affecting Students Academic Performance: A Case Study of Sohar University," Psychology and Education, vol. 58, no. 5, pp. 4624-4635, 2021, ISSN: 1553-6939. Available: www.psychologyandeducation.net.

[20] L. S. Sandra, F. Lumbangaol, and T. Matsuo, "Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature

[21] .2015.12.114.