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

Affiliation
Department of Data Science, School of Computer Science, BINUS University, Jakarta, Indonesia

BINUS UNIVERSITY
School of Computer Science

Authors

Fiona Maharani Nugraha
2602199582
fiona.nugraha@binus.ac.id

Kimberly Kayla Dewi
2602190816
kimberly.dewi@binus.ac.id

Alexander Agung Santoso Gunawan
D3579
aagung@binus.ac.id

Jeffrey Junior Tedjasulaksana
D7063
jeffrey.t@binus.ac.id

AIMS 2025
3rd IEEE International Conference on Artificial Intelligence and Mechatronics Systems

Table of Contents

Abstract

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

- **Problem:** Traditional methods for assessing student performance are time-consuming and lack accuracy.

- **Objective:** Apply machine learning (KNN, Decision Tree, Linear Regression, Random Forest) to predict student passing grades using demographics and academic performance data.

- **Methodology:** Use regression techniques to predict passing grades and compare the performance of different ML models.

- **Findings:**
  - Best Model: Linear Regression.
  - Performance Metrics:
    - MSE = 4.8801
    - RMSE = 2.2091
    - MAE = 1.3436
    - MAPE = 3.29%
    - R^2 = 0.762 (highest explanatory power).

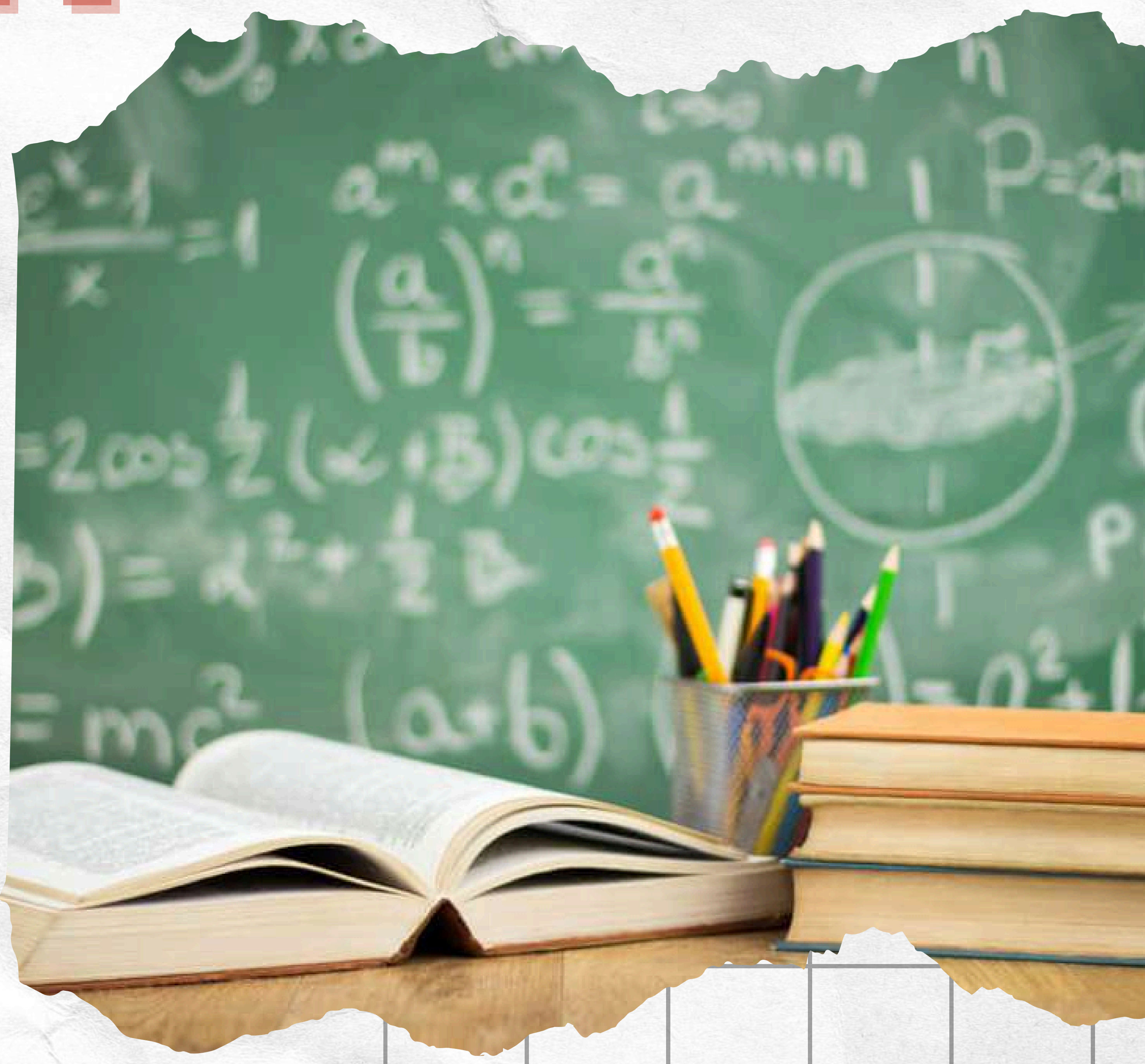
- **Conclusion:** Linear Regression provides the most accurate predictions with the lowest error rates.

Introduction

Role of Education: Essential for personal and societal growth through skill and knowledge development.

Importance in Secondary Education: Predicting student performance aids in enhancing teaching, personalizing learning, and supporting at-risk students.

Limitations of Traditional Methods: Often overlook complex factors influencing success.



Introduction

Machine Learning (ML):

- Utilizes academic, demographic, and behavioral data for accurate predictions.
- Identifies patterns to provide actionable insights on student outcomes.

Study Focus: Reviews research to identify

- Effective ML algorithms.
- Key factors influencing academic success.
- Best practices to improve educational interventions and outcomes.

Literature Review

✱ Machine Learning

Definition:

- Subset of artificial intelligence.
- Enables data-driven predictions without explicit programming.

Applications in Education:

- Effectively predicts student performance.
- Provides insights to enhance learning strategies and understanding.

Focus on Supervised Learning Algorithms:

- Decision Tree (DT): Splits data into branches based on feature values.
- K-Nearest Neighbors (KNN): Predicts outcomes using nearby data points.
- Linear Regression: Models relationships between variables and continuous outcomes.
- Random Forest: Combines multiple decision trees for improved prediction accuracy and robustness.

Literature Review

✱ Educational Data Mining

Educational Data Mining (EDM):

- Applies machine learning to analyze student behaviors, performance, and learning patterns.
- Utilizes data from academic records and digital platforms.

Key Functions of EDM:

- Predicts student outcomes.
- Identifies at-risk students.
- Personalizes learning experiences.

Student Performance Dataset:

- Includes demographics, academic performance, and behavioral data.
- Used to build predictive models for identifying success predictors.

Impact of EDM:

- Helps institutions improve support and interventions for students.

Literature Review

✱ Regression and Evaluation Metrics

Regression Models in Education:

- Predict continuous variables (e.g., student grades, GPA, exam scores).
- Linear Regression models the relationship between dependent and independent variables.

Application in Education:

- Used to forecast academic outcomes like exam scores and course completion rates.
- Helps identify factors influencing academic performance and guide interventions.

Evaluation Metrics for Model Performance:

- MSE & RMSE: Measure average squared differences between predicted and actual values (error magnitude).
- MAE: Quantifies average error magnitude without direction.
- R-squared (R^2): Explains variance in the dependent variable, indicating model explanatory power.
- MAPE: Normalized error measure, easier to compare models across different datasets.

Literature Review

✱ Related Works

Suzan et al. (2021):

- Focus: Student adaptability in online education using machine learning.
- Algorithms used: Random Forest, Decision Tree.
- Key Finding: Random Forest achieved 89.63% accuracy, outperforming other models.
- Emphasis: Machine learning improves prediction accuracy, especially for adaptability in online learning.

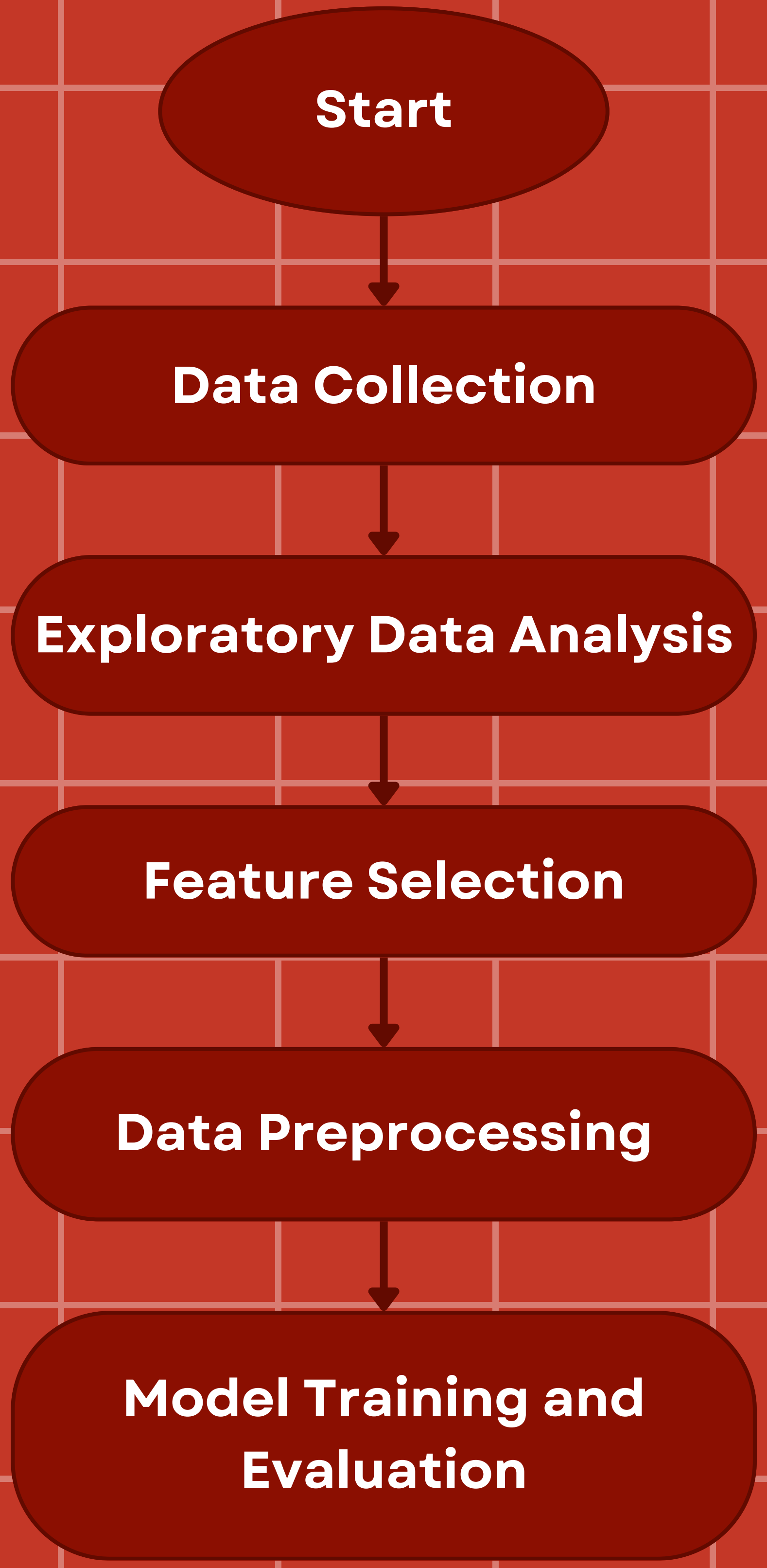
M. Wu et al. (2024):

- Focus: Predicting academic performance using machine learning.
- Key Finding: Ensemble learning methods performed best (87.67% accuracy), followed by SVM (84.30%).
- Emphasis: Importance of demographic, academic, and behavioral factors, and the need for early intervention.

Common Themes:

- Feature selection and algorithm performance are crucial for improving prediction accuracy.
- Machine learning enhances adaptive learning and early identification of at-risk students.

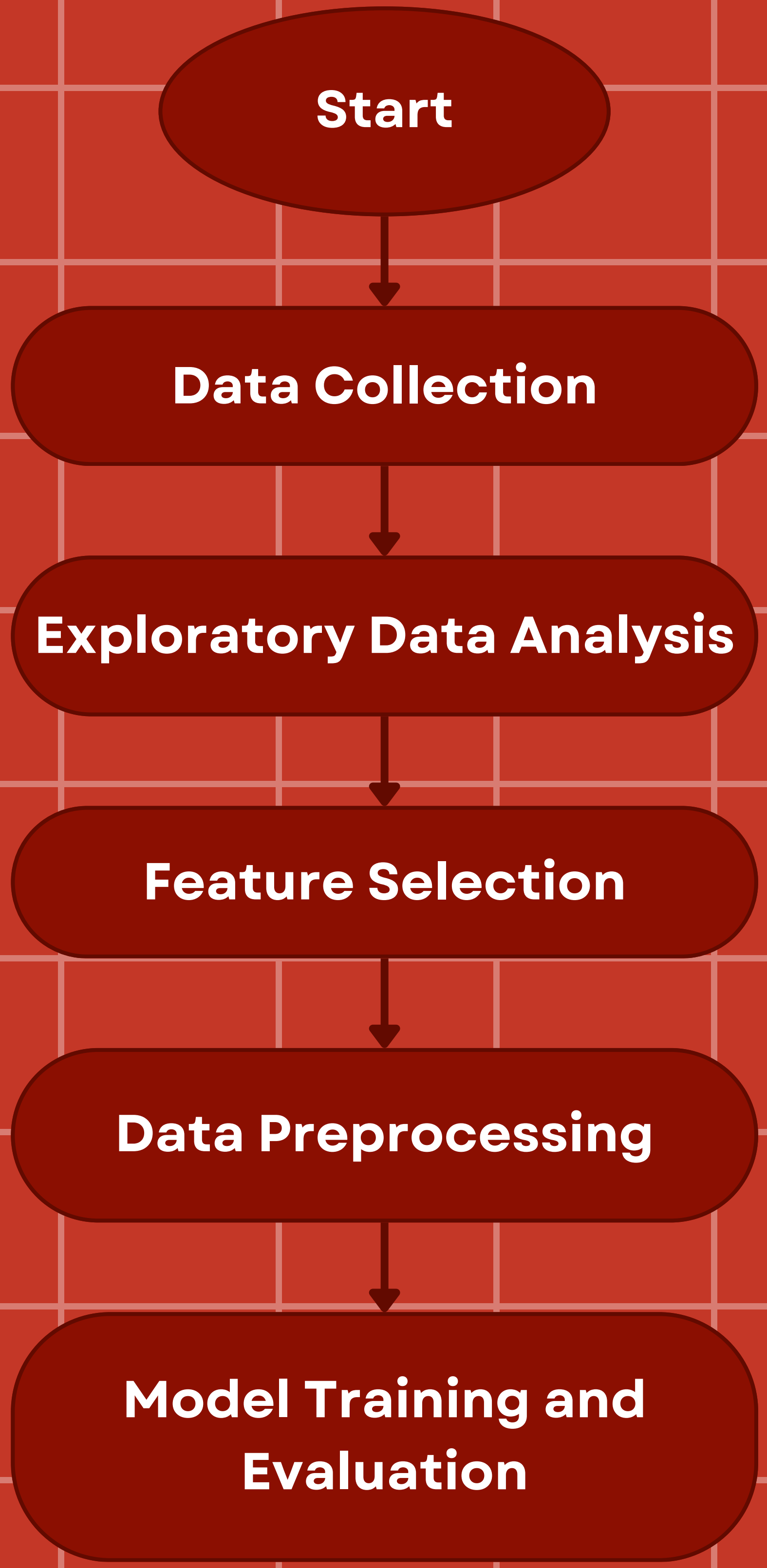
Material and Methodology



✱ Data Collection

- Dataset Title: Student Performance Dataset
- Source: Kaggle, uploaded by Dev Ansodariya, a student and software engineer at San Jose State University.
- Location: Data from two Portuguese schools: ‘GP’ (Gabriel Pereira) and ‘MS’ (Mousinho da Silveria).
- Subjects: Portuguese and Mathematics.
- Performance Metrics:
  - Student grades, demographics, social, behavioral, and school-related attributes.
  - Examples: Age, gender, parental education, study time, health status.
- Purpose: Used for regression analysis to predict student performance.

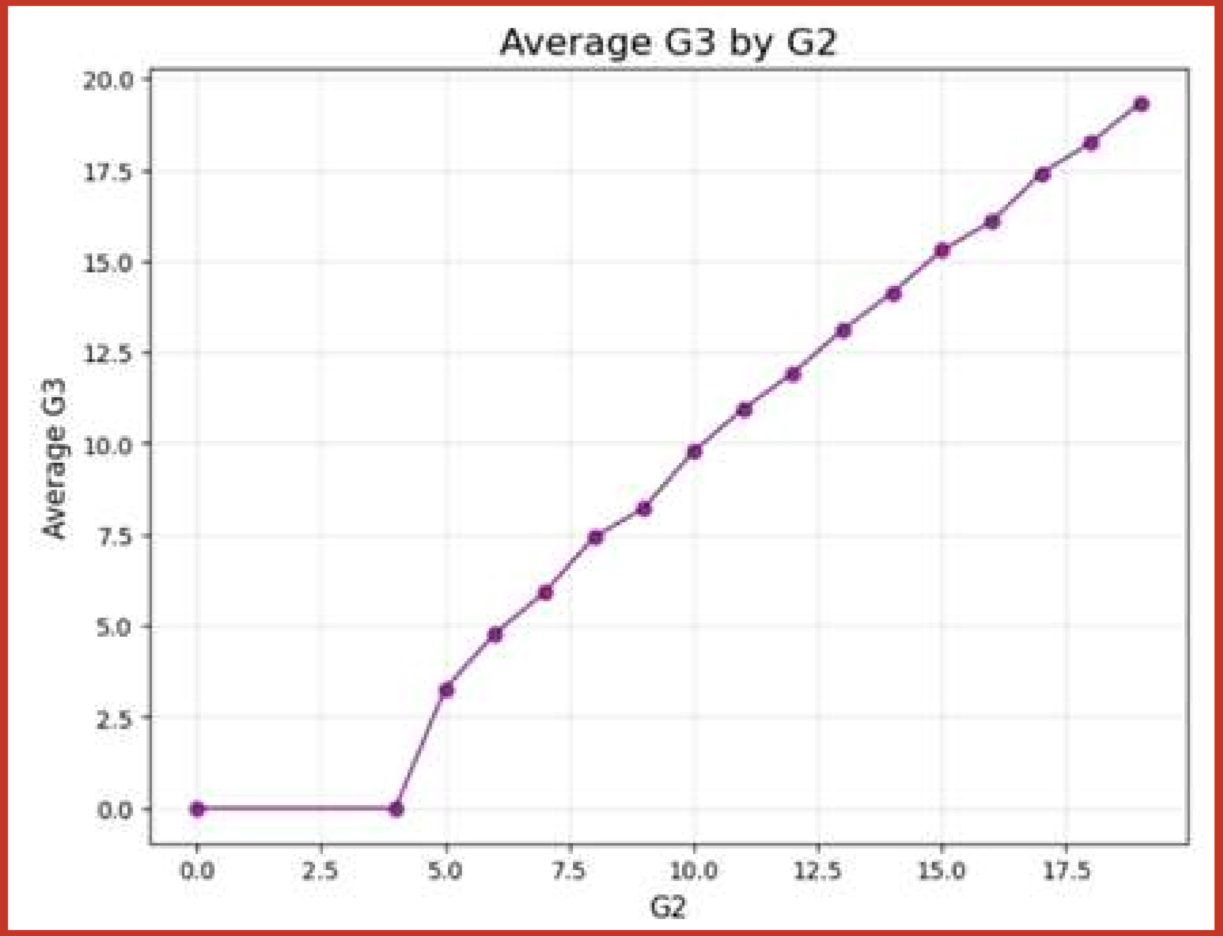
Material and Methodology



✱ Exploratory Data Analysis (EDA)

Dataset Overview:

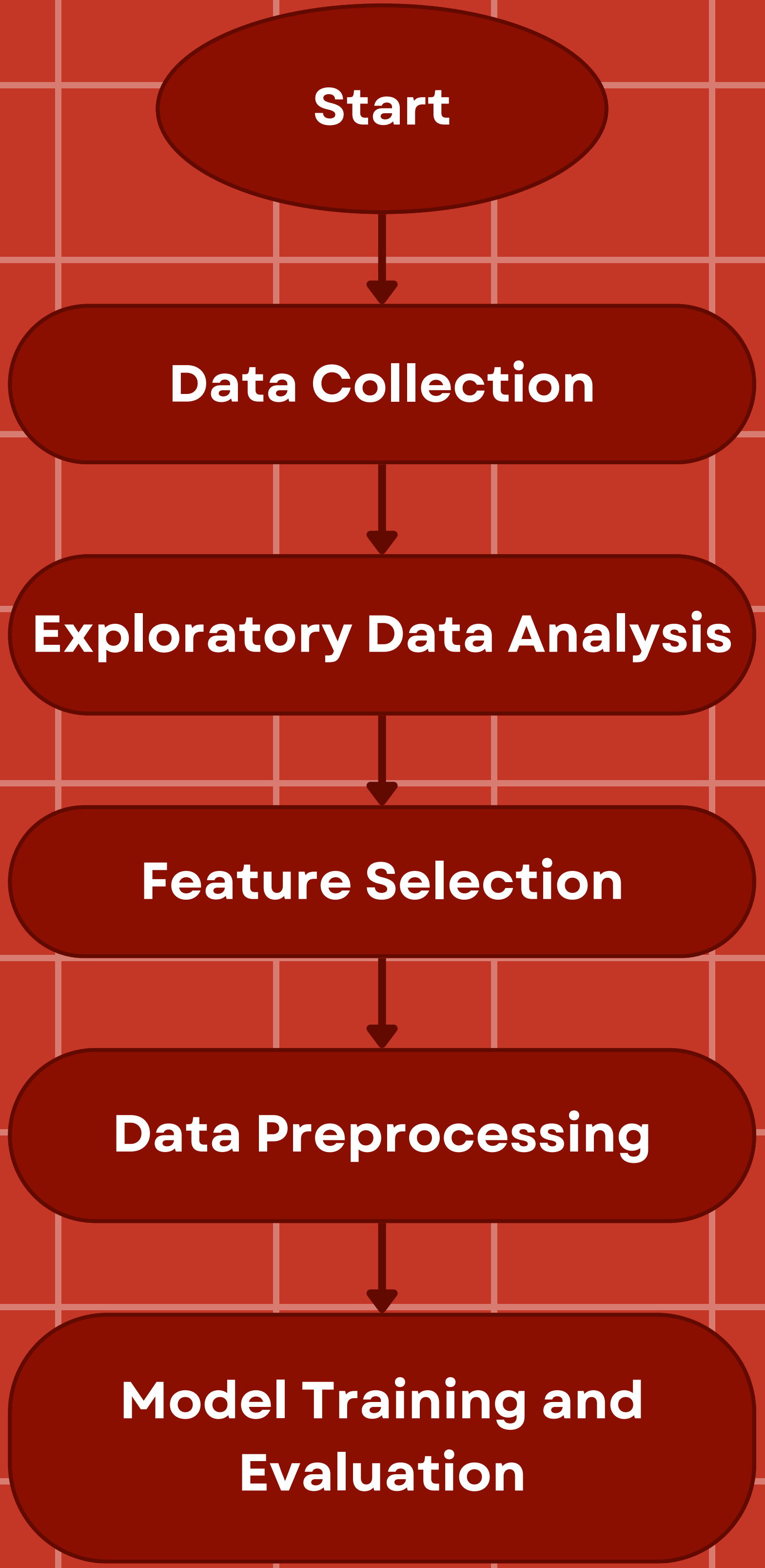
- 395 rows, 33 columns.
- Includes demographic, social, behavioral, school-related, and academic performance data for Portuguese and Mathematics subjects.
- Focused on 16 numerical features for regression analysis.



Key Findings:

- Strong correlation between G2 (previous grade) and G3 (final grade). Higher G2 correlates with higher G3 scores (near-linear relationship).
- G2 is a strong predictor of G3.

Material and Methodology



Exploratory Data Analysis (EDA)

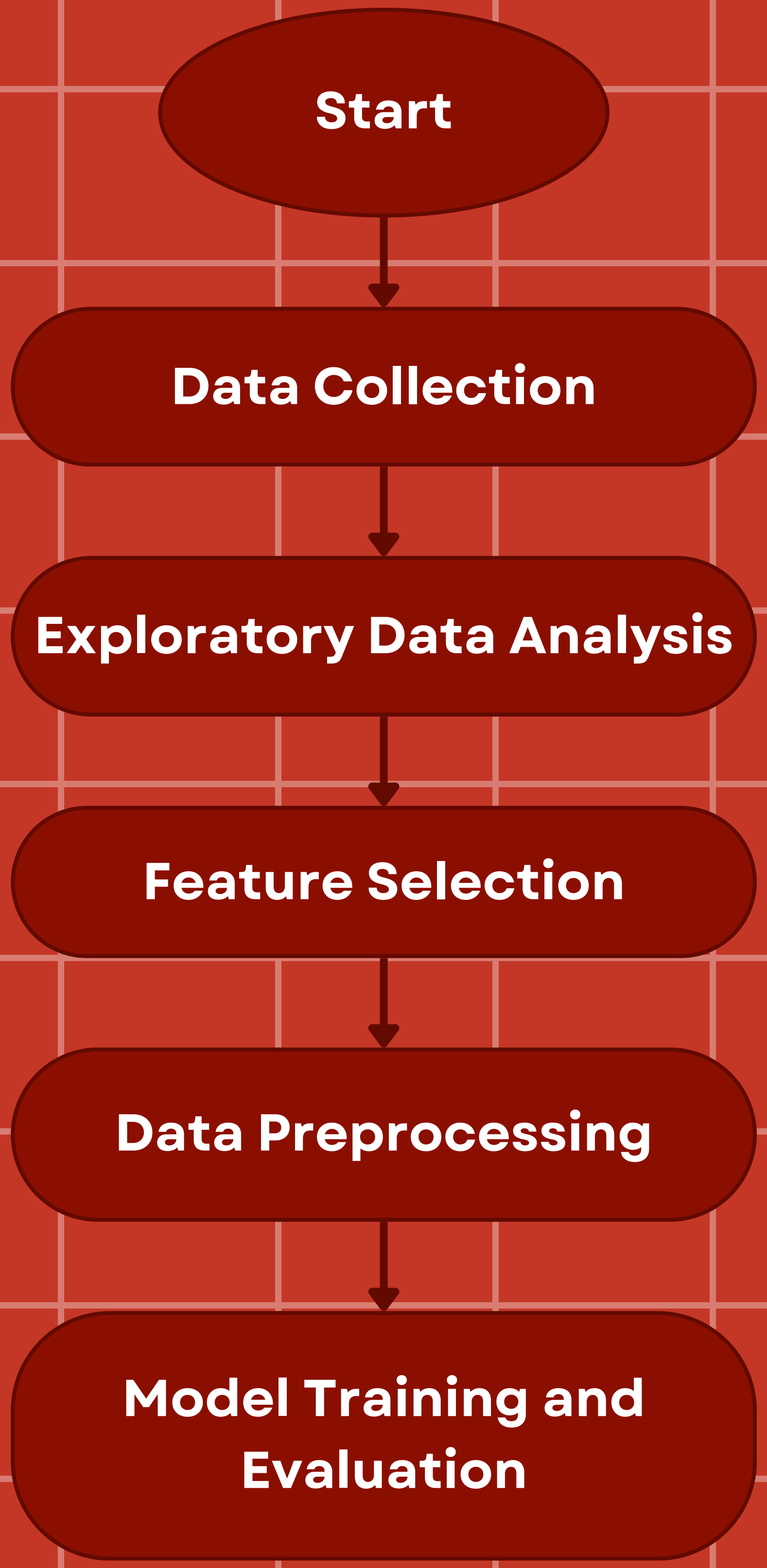
Imbalance Issues:

- Dataset is imbalanced, e.g., the "school" feature has more data for GP (349) compard to MS (46).
- Imbalance may affect model performance, especially in algorithms sensitive to class distribution.

Target Variable:

- G3 (final grade) is the target variable to predict student academic success.

Material and Methodology



✱ Feature Selection

Correlation Analysis:

- Strong correlation between G1, G2 (grades) and G3 (final grade).
- Parental education (Medu, Fedu) and behavioral features (goout, Dalc, Walc) show weaker correlation with G3.

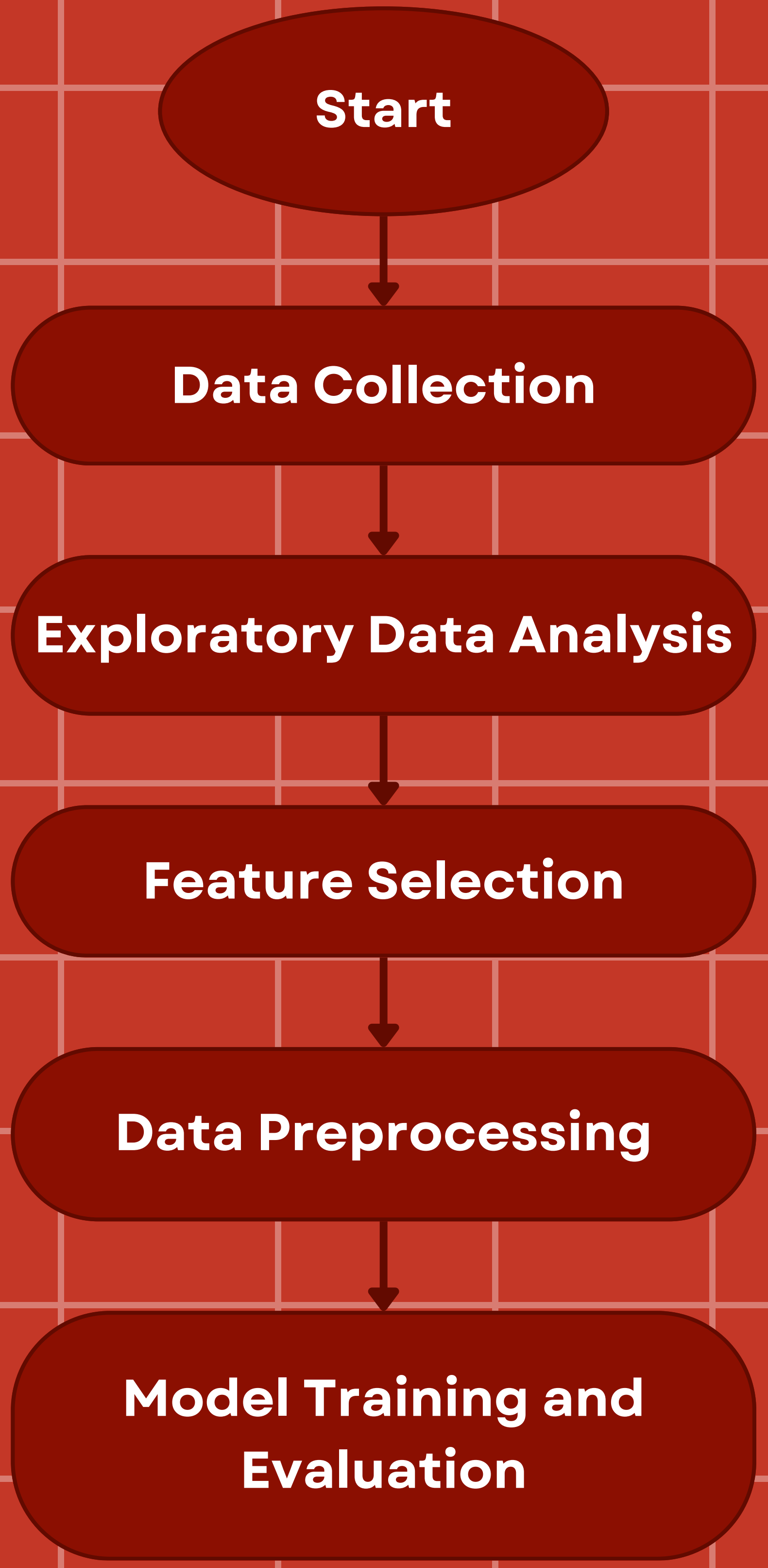
Feature Removal:

- Features with correlation < 0.1 were removed as they are less significant for predicting G3.
- The "failures" feature (past class failures) has strong negative correlation with G3, so it is retained.

Result:

- Reduced model complexity, improving interpretability and performance.
- Data is prepared for machine learning algorithms.

Material and Methodology

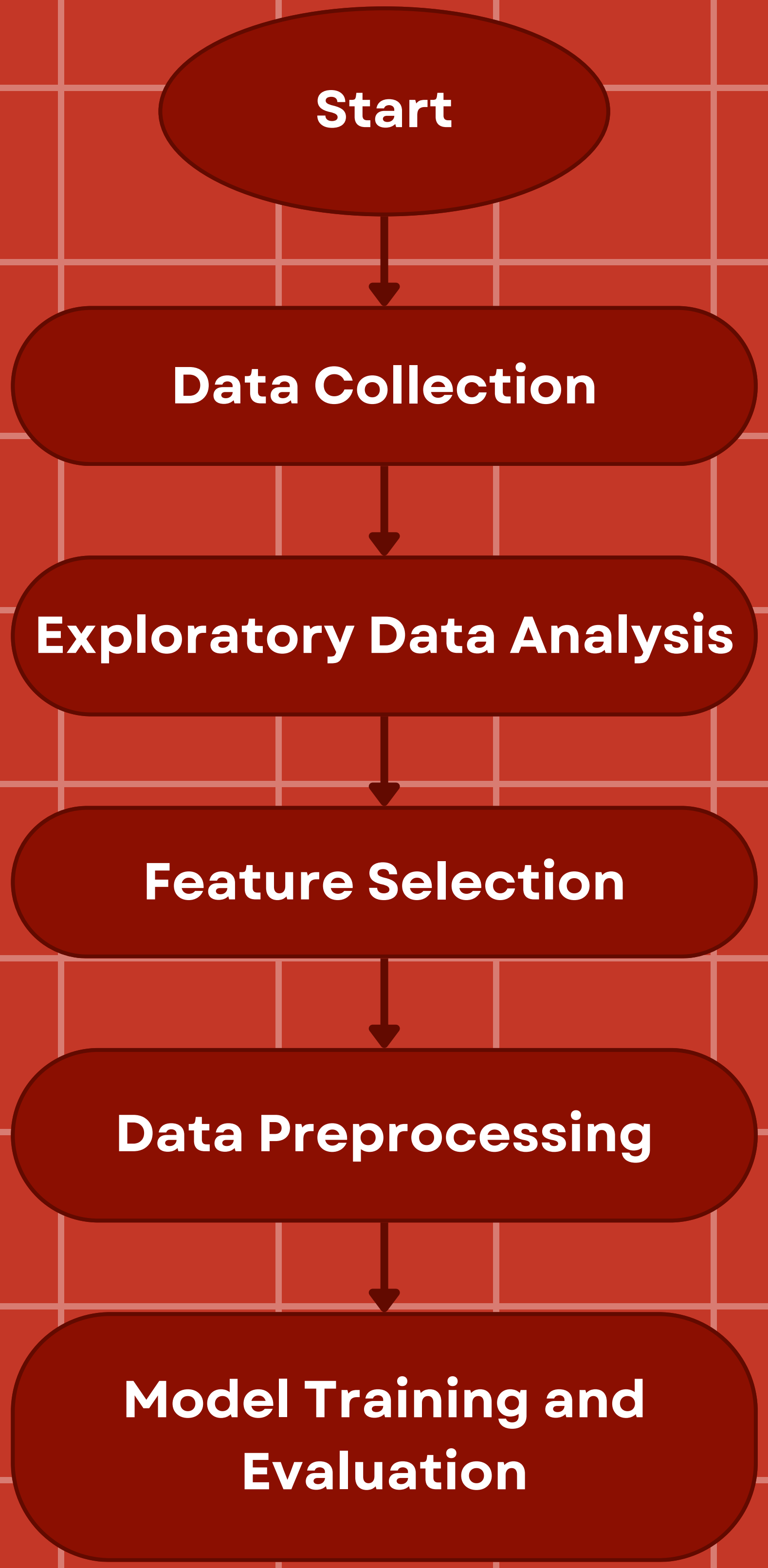


Feature Selection

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
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

Material and Methodology



✱ Data Preprocessing

Scaling:

- Numerical data standardized using StandardScaler to ensure features are on the same scale.
- Transforms data to mean = 0, standard deviation = 1 for improved model convergence.

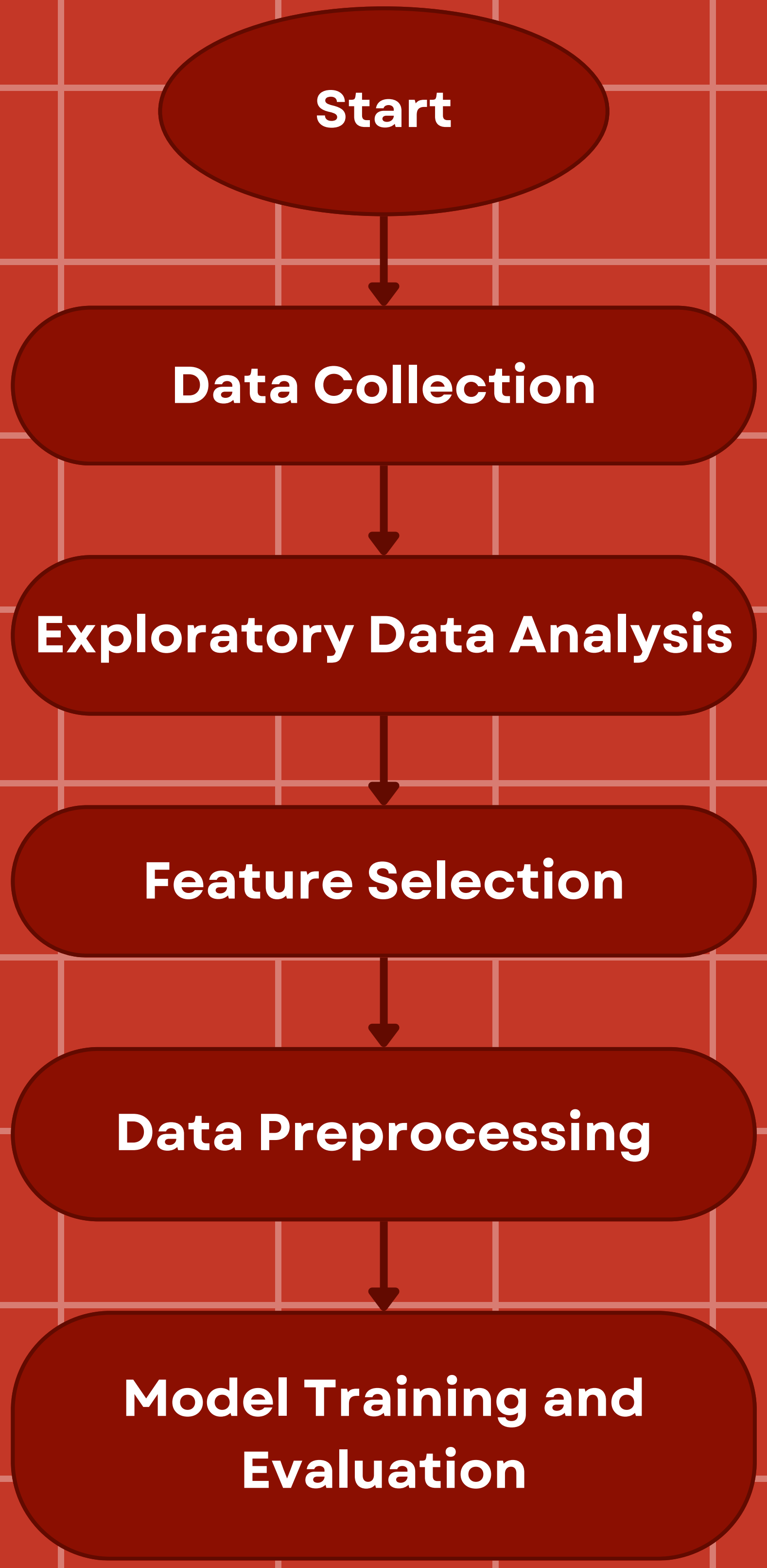
Outliers:

- Identified in age, absences, G1, G2, G3 columns.
- Dataset reduced from 395 to 341 rows after outlier removal.

Data Split:

- Dataset split into 80% training and 20% testing using train_test_split.

Material and Methodology



✱ Model Training and Evaluation

Training:

- Models trained on the training dataset to learn relationships between features and target variable (G3).

Models Used:

- Linear Regression
- K-Nearest Neighbors (n_neighbors=3)
- Decision Tree Regressor (max depth=5)
- Random Forest (n_estimators=100)

Evaluation Metrics:

- MSE, RMSE: Measure average squared differences between predicted and actual values.
- MAE: Quantifies average error magnitude, ignoring direction.
- R^2: Indicates the proportion of variance explained by the model.
- MAPE: Normalized error measure, useful for comparing models but ineffective when actual values are close to zero.

Result

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%

Result

Best Performing Model: Linear Regression

Strong Performance: Random Forest

Moderate Performance: K-Nearest Neighbors (KNN)

Weakest performance: Decision Tree

Conclusion

Research Objective

Predict middle school students' passing grades using regression-based machine learning models.

Best Performing Model: Linear Regression

- Lowest errors: MSE, RMSE, MAE.
- Highest R^2 among models.
- Slightly lower MAPE than Decision Tree.

Future Work Recommendations

- Explore advanced regression techniques (e.g., neural networks).
- Integrate additional data sources.
- Fine-tune hyperparameters for improved predictive accuracy.

Author’s Contribution

Fiona Maharani Nugraha
- Led manuscript writing.
- Collaborated with co-authors for content refinement.

Kimberly Kayla Dewi
- Devised the project and developed conceptual ideas.
- Conducted experimental work and implementation.

Alexander Agung Santoso Gunawan
- Provided consultation and insights for manuscript development.
- Reviewed and approved the final manuscript version.

Jeffrey Junior Tedjasulaksana
- Provided consultation and insights for manuscript development.
- Reviewed and approved the final manuscript version.

Availability Data and Materials

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 Performance Prediction Using Machine Learning Techniques,” International Journal of Advanced Computer Science and Applications, vol. 10, no.8, pp. 1-5, 2019.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] L. S. Sandra, F. Lumbangaol, and T. Matsuo, “Machine Learning Algorithm to Predict Student’s Performance: A Systematic Literature Review,” TEM Journal, vol. 10, no. 4, pp. 1919-1927, Nov. 2021, doi: 10.18421/TEM104-56.
- [7] 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.
- [8] 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.
- [9] Z.-H. Zhou, Machine Learning, Springer Nature, 2021.

References

- [10]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.
- [11] 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.
- [12]C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics. 2020.
- [13] 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.
- [14] 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.
- [15]J. Han, J. Pei, and M. Kamber. “Data Mining: Concepts and Techniques”. Elsevier. 2021.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] D.S. Moore, G.P. McCabe, and B.A. Craig, “Introduction to the Practice of Statistics,” 9th ed. New York: W.H Freeman, 2017.
- [20] 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.neucom.2015.12.114.

Thank You So Much!