

Comparative Analysis to Predict Reading Literacy based on PISA 2022 using Gradient Boosted Decision Trees and Extreme Gradient Boosting

Hary Susanto¹, Ema Utami²

¹ Magister of Informatics, Universitas Amikom Yogyakarta, Indonesia

² Doctoral Program of Informatics, Universitas Amikom Yogyakarta, Indonesia

✉ **Corresponding Author** : Hary Susanto (email : harygading@gmail.com)

Article Information	ABSTRACT
Article History Received : December 23, 2024 Revised : January 5, 2025 Published : January 16, 2025 Keywords: <i>GBDT;</i> <i>Literacy;</i> <i>PISA;</i> <i>XGBoost</i>	Reading is a fundamental skill essential for interdisciplinary understanding and serves as a crucial indikator of a nation's educational quality. PISA provides an international evaluation of students' reading literacy across various countries, including Indonesia. This study compares the performance of Gradient Boosting Decision Trees (GBDT) and Extreme Gradient Boosting (XGBoost), two widely recognized machine learning algorithms for predicting reading literacy, utilizing PISA 2022 data from 12.853 Indonesian students and 59 variables from the Student Questionnaire Data File. GBDT achieved R^2 of 0.5106, with optimal parameters ($n_estimators = 150$, $learning_rate = 0.2$, $max_depth = 3$, $subsample = 0.9$). XGBoost reached a higher R^2 of 0.5247, with parameters ($n_estimators = 1000$, $learning_rate = 0.01$, $max_depth = 7$, $colsample_bytree = 0.3$, $min_child_weight = 20$, $gamma = 1$, $alpha = 0$), indicating XGBoost's superior performance in predicting reading literacy. Further analysis revealed that the most significant variables in the GBDT model included students' access to technology at home, extracurricular creative activities, socioeconomic status, school involvement in sustainable development, and problem-solving skills. In contrast, significant variables in the XGBoost model included family support, socioeconomic status, school belongingness, family environment's effectiveness in fostering creativity, and student imagination.

INTRODUCTION

The process through which a person reads, applies, assesses, considers, and interacts with a text in order to grow in knowledge, potential, and social participation is known as reading literacy (Dilekçi, 2022). Reading literacy has become one of the key indicators of a country's educational quality. The Organization for Economic Cooperation and Development (OECD) implements the Programme for International Student Assessment (PISA), which measures students' abilities in reading literacy, numeracy, science literacy, and creative thinking across various countries (State, 2022). PISA 2022 also includes an assessment of financial literacy for the younger generation, which is optional for countries and economies (State, 2022). Reading literacy skills are of paramount importance for a nation in determining its educational policy direction. Predicting reading literacy is also crucial for educators to identify steps for improvement in the learning process (Tağa, 2023). Because machine learning techniques do not rely on presumptions and can investigate trends in large, complicated datasets in a data-driven way, they can offer education researchers useful insights. In recent years, machine learning has gained popularity in educational research (Liu et al., 2022). Unlike statistical modeling, which often depends on predefined

hypotheses, machine learning relies more on data, highlighting a key difference between these two strategies. In general, statistical models are more suited for analyzing relationships between specific components, while machine learning is better suited for making predictions (Bu & Chen, 2023).

Previous research in predicting reading literacy using machine learning has shown the following results: Support Vector Regression (SVR) achieved $R^2 = 0.4660$ with RMSE = 63.4415; Random Forest Regression (RFR) resulted in $R^2 = 0.4372$ with RMSE = 63.1013; and eXtreme Gradient Boosting (XGBoost) reached $R^2 = 0.4802$ with RMSE = 62.6765 (Liu et al., 2022). Additional investigation utilizing PISA 2018 3,775 fifteen-year-old students from 45 Macau schools participated in a Random Forest Regression study with 41 independent variables. Thirteen factors were evaluated at each split point in the 1,000 decision trees that the algorithm produced. The regression model showed a prediction performance of 4,366.32 (RMSE = 66.28) and explained 43.50% ($R^2 = 0.44$) of the variance in reading achievement (Y. Wang et al., 2023).

Additional examination of 12,058 Chinese students from the PISA 2018 database (also known as B-S-J-Z, or Beijing, Shanghai, Jiangsu, and Zhejiang), 6,676 students from Singapore, and 3,775 students from Macau, compared with the overall PISA 2018 data (612,004 students) using GBDT. This study revealed that a combination of a maximum depth of 3 and a maximum number of model parameters ($n_{\text{estimators}}$) set to 500 resulted in the minimum error. By utilizing optimal hyperparameters, the study further developed and evaluated the final GBDT model, which accounted for 49.90% of the variance (i.e., $R^2 = 0.499$) (Bu & Chen, 2023).

This study aims to compare the performance of GBDT and XGBoost in predicting reading literacy based on PISA 2022 data. This approach will not only identify the superior model but also reveal significant factors in predicting Indonesian students' reading. By using parameter optimization through GridSearchCV, this research aims to maximize the prediction accuracy of both models.

RESEARCH METHODS

This section provides detailed information on the dataset and methodology used to predict reading literacy skills of Indonesian students using the PISA 2022 dataset with the GBDT and XGBoost models. Figure 1 illustrates the overall research flow, beginning with the acquisition of the PISA 2022 data, preprocessing steps, model testing, analytical approach, and concluding with analysis and conclusions.

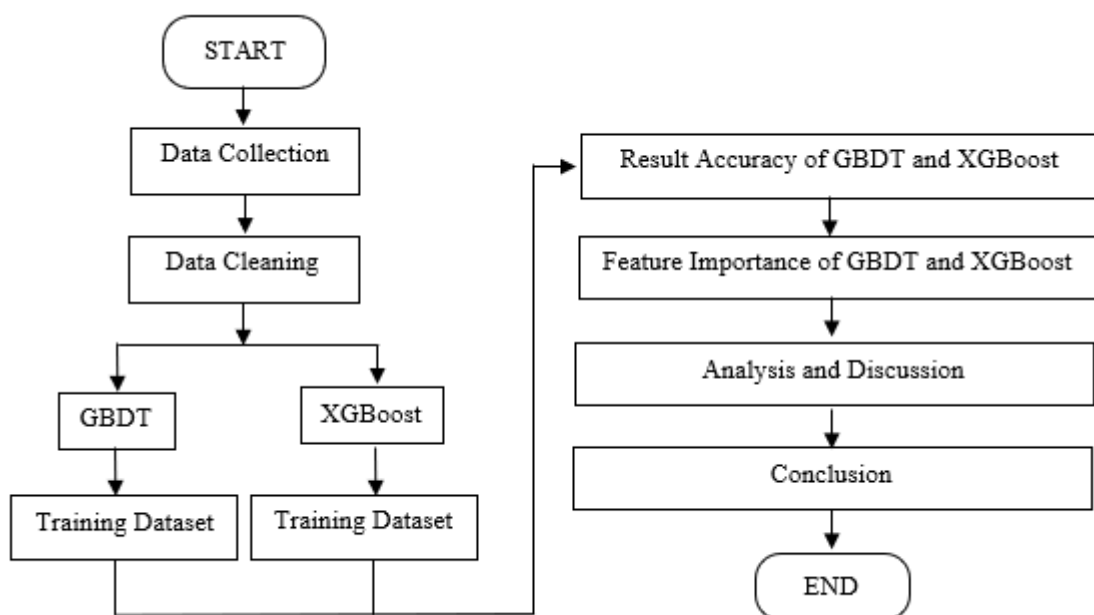


Figure 1. Research Process Flow

1. Data Collection

The PISA 2022 dataset from Indonesia was used in this study, which includes 13.439 students from 410 schools. This sample represents approximately 3.790.800 students aged 15, or around 85% of the total population in that age group (State, 2022). The PISA dataset provides comprehensive data from students, school principals, and parents, allowing for an in-depth analysis of the factors influencing reading literacy.

2. Data Preprocessing

In the data cleaning process, entries with missing values were eliminated to preserve the dataset's quality. Data from other countries was also excluded, resulting in a final dataset with 12.853 data points and 59 relevant variables related to reading literacy, out of an initial 1.278 variables.

3. Model Training and Parameter Optimization

The dataset used in this study was split into two parts, with 80% allocated as training data and 20% as testing data. This division ensures an accurate evaluation of the model's predictive ability. To optimize the model parameters, the GridSearchCV method was used to identify the optimal parameter combinations, including learning rate, maximum depth, and n_estimators, for the GBDT and XGBoost models. Additionally, to improve the reliability of the model's results, cross-validation was conducted using 5-fold and 10-fold methods, enabling a more in-depth evaluation of the stability and generalizability of the developed models.

4. Model Selection and Analysis

GBDT and XGBoost were chosen for their efficiency in handling large, structured data and their ability to perform gradient optimization with high speed. The model flow for both algorithms is visualized in Figure 2, which illustrates the iterative process of improving prediction accuracy.

5. Evaluation Metrics

The models were evaluated using several performance metrics, including the Coefficient of Determination (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). R^2 measures the extent to which the dependent variable can be explained by the independent variables in the model, with a better fit indicated by values near 1 (Katagiri & Fujii, 2022). The average squared discrepancies between anticipated and actual values are assessed using MSE and RMSE, better predictions are indicated by smaller MSE and RMSE values. On the other hand, Unlike MSE or RMSE, which are more impacted by extreme values, MAE provides a more straightforward understanding of the size of prediction mistakes by measuring the average absolute error between projected and actual values (Robeson & Willmott, 2023).

Gradient Boosting (GB) is a method commonly applied in regression model development (Natekin & Knoll, 2013). This method aims to minimize the loss function through a gradient-based approach, performed iteratively. A supervised learning model called Decision Tree is based on decision trees (Farhangi et al., 2021). This approach solves simple classification and regression issues by utilizing the information theory concept of entropy. The decision tree model's structure can be modified based on the dataset's properties (Breiman, 2019), Gradient Boosting has become a powerful boosting-based machine learning algorithm achieving cutting-edge accuracy in various real-world tasks (Moitra, 2024). Gradient Boosting has emerged as one of the most effective methods for creating predictive models. With gradient boosting at its core, the XGBoost algorithm offers the benefits of high prediction accuracy, good robustness, and quick operation (Shen, 2022). In this context, the XGBoost algorithm has demonstrated excellent performance as an efficient machine learning algorithm. It not only can handle largescale datasets, but also has strong generalization ability and robustness, and can discover potential patterns and patterns in the data (Su et al., 2024). One popular boosting algorithm among ensemble approaches that has been effectively used to solve imbalanced data difficulties in predictions is XGBoost (Smiti et al., 2024). XGBoost uses decision trees as base learners and employs the idea of integrating the concept of Gradient Boosting to form a strong learner by combining several decision tree models and synthesizing them together. The basic steps of its algorithm are as follows (Guoqing Chen, 2023):

Step 1: Constructiion of decision tree function

$$\hat{y} = \sum_{k=1}^K f_k(X_i) \quad (1)$$

where K contributes the quantity of trees, $\hat{y}$ to the anticipated result of example X_i :

Step 2: The error function is Obj θ which is calculated using the mean square error.

$$Obj \theta = \sum_{i=1}^n l(y_i + \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

Where $l(y_i + \hat{y}_i)$ is the training error of sample X_i , $\Omega(f_k)$ denotes the regular term for the K number of tress, $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{i=1}^T w_j^2$

Step 3: Iteratively include the regression tree to optimize the finas loss function. Obj θ can be regarded as:

$$Gj(\theta) = \sum_{i=1}^n [2(y_i - \hat{y}_i^{t-1})f_t(x_i) + f_t(x_i)^2] + Y^T + \frac{1}{2} \lambda \sum_{i=1}^T w_j^2 + C \quad (3)$$

In order to predict the reading literacy score PV1, both models are trained using 59 filtered variables. The hyperparameters of the two aforementioned models are tuned independently to prevent overfitting. GridSearchCV is used to find the ideal values for the hyperparameters of the model.. GridSearchCV is a critical feature in this model, as its performance improves significantly after selecting the right values for the hyperparameters (Ahmad et al., 2022). The best hyperparameter combination on the test set is chosen by adjusting the parameters using 5 fold and 10 fold cross-validation.

Table 1 displays the predictive skills that were optimized from both models. R2 (Coefficient of Determination), MSE (Mean Squared Error), RMSE (Root Mean Square Error), and MAE (Mean Absolute Error) are used to quantify the predicting performance. The degree to which the dependent variable can be anticipated from the independent variables is shown by the coefficient of determination. Stated otherwise, this index displays the degree to which the regression line fits the measured data set. A greater R^2 number denotes a reduced error between the estimated and measured values. The coefficient of determination normally runs from $0 \leq R^2 \leq 1$ (Katagiri & Fujii, 2022).

In the field of AI and ML applications, calculating Mean Square (MS) and Mean Squared Difference (MSE), which measure the average squared difference between predicted and actual values, is crucial for evaluating model performance. In training algorithms, MSE also acts as a loss function, directing them to reduce the discrepancy between expected and actual results (Mahdavi, 2024). RMSE can be derived by calculating the standard deviation between the actual and predicted ratings of each user. The smaller the RMSE value, the higher the accuracy of the recommended algorithm (Kuan et al., 2022). The average absolute error between the actual and anticipated values is measured by MAE. Squared measurements like Mean Squared Error (MSE) or its square root, Root Mean Squared Error (RMSE), are frequently used to characterize dimensional mistakes when assessing the performance of quantitative models. Mean Absolute Error (MAE) and other absolute value-based measurements are easier to interpret than Mean Squared Error (MSE) or root mean square error (RMSE) when it comes to assessing average errors (Robeson & Willmott, 2023). Following weighting, each coefficient is computed and described in detail below.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

RESULTS AND DISCUSSION

Model testing was conducted using Python and the pandas, numpy, matplotlib, and scikit-learn libraries in Google Colab. The GridSearchCV method was chosen to find the best parameter combinations for both models, helping to efficiently optimize model accuracy and ensure that the selected parameters yield the best performance on the test data. The best hyperparameter combination on the test set was chosen by adjusting the parameters using 5 fold and 10 fold cross-validation assessment techniques. The best hyperparameter combinations for the two models above are displayed in Table 1. After conducting 5-fold and 10-fold cross-validation, the prediction results from both models were found to be similar, and XGBoost was slightly better than the GBDT model.

Table 1. Optimal Hyperparameter Combination for Both Models

No	Model	Hyperparameter Combinations	Optimal Hyperparameter Combination
1	GBDT	n_estimators=(50, 100, 150) learning_rate=(0.01, 0.1, 0.2) max_depth=(3, 5, 7) subsample=(0.6, 0.8, 0.9)	n_estimators=150 learning_rate=0.2 max_depth=3 subsample =0.9
2	XGBoost	n_estimators=(100, 500, 1000) learning_rate=(0.01, 0.1, 0.2) max_depth=(3, 5, 7) colsample_bytree=(0.3, 0.7) min_child_weight=(10,20,30) gamma=(0,1,5) alpha=(0, 0.1, 1)	n_estimators=1000 learning_rate=0.01 max_depth=7 colsample_bytree=0.3 min_child_weight=20 gamma=1 alpha=0

Table 2. Comparison of Predictive Performance and Parameter Tuning Results Using 5 Fold Cross-Validation for Both Models

No	Model	R ²	MSE	RMSE	MAE
1	GBDT	0.4875	2975.2897	54.5462	43.1856
2	XGBoost	0.5083	2876.8655	53.6364	42.4631

Table 3. Comparison of Predictive Performance and Parameter Tuning Results Using 10 Fold Cross-Validation for Both Models

No	Model	R ²	MSE	RMSE	MAE
1	GBDT	0.4634	15716.5256	125.3656	97.2856
2	XGBoost	0.5102	2864.1093	53.5174	42.3702

Table 4. Comparison of Test Data Evaluation Results for Predictive Performance and Parameter Tuning of Both Models

No	Model	R ²	MSE	RMSE	MAE
1	GBDT	0.5106	2975.2897	54.5462	42.1724
2	XGBoost	0.5247	2788.1808	52.8032	41.5819

Figure 2 shows that at each value of n_estimators, it is seen that max_depth=3 (blue line) has the lowest RMSE compared to other max_depth values (max_depth=5 and max_depth=7). This indicates that the model with max_depth=3 may provide the best balance between bias and variance for this data, and more complex models (max_depth=5 and max_depth=7) are likely to

tend towards overfitting. For each value of `max_depth`, there is a slight decrease in RMSE when `n_estimators` increases from 50 to 100. However, this decrease is not significant when `n_estimators` is further increased to 150. This indicates that after a certain point (around `n_estimators`=150), adding more estimators does not provide a significant improvement in performance, and RMSE remains relatively stable. The combination of `max_depth`=3 and `n_estimators`=150 seems to provide the best performance (lowest RMSE) on this data, as it strikes a good balance between model complexity and prediction error. Increasing the number of estimators or maximum depth further may not provide significant benefits and could increase the risk of overfitting. With these results, `max_depth`=3 and `n_estimators`=150 are the optimal hyperparameters for the GBDT model, producing $R^2 = 0.5106$, $MSE = 2975.2897$, $RMSE = 54.5462$, $MAE = 97.2856$, compared to previous research using the PISA 2018 dataset with 12,058 students (Beijing, Shanghai, Jiangsu, Zhejiang, Singapore, and Macau) where the GBDT with `max_depth`=3 and `n_estimators`=500 yielded $R^2 = 0.499$, $MSE = 4670.3556$, $RMSE = 68.34$ (Bu & Chen, 2023). The subsequent research in Macau with 10-fold cross-validation tested a Random Forest regression model consisting of 41 independent variables. The algorithm was set to test 13 variables at each decision tree split and was instructed to grow 1,000 decision trees randomly. The regression model has a predictive performance of 4,366.32 ($RMSE = 66.28$) and can explain 43.50% ($R^2 = .44$) of the variance in performance. (Y. Wang et al., 2023).

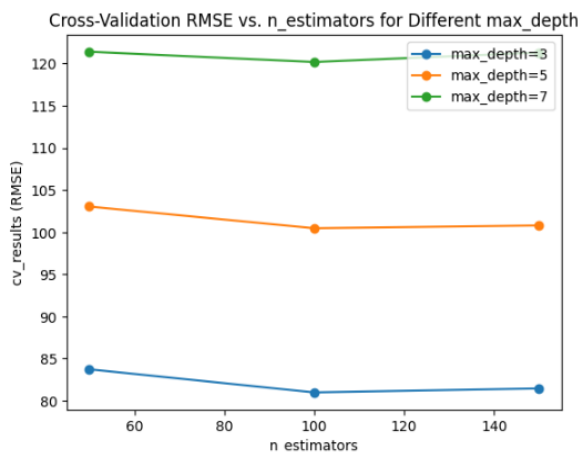


Figure 2. RMSE vs. `n_estimators` for `max_depth` in GBDT

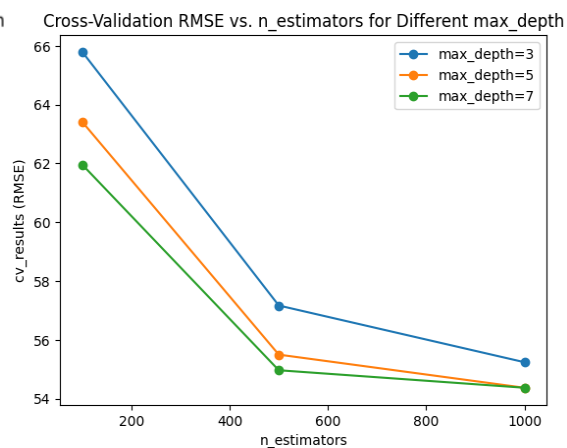


Figure 3. RMSE vs. `n_estimators` for `max_depth` in XGboost

Figure 3 shows that as `n_estimators` increases, RMSE continues to decrease for all `max_depth` values. This indicates that the model improves its prediction accuracy by increasing the number of trees, which helps to better capture patterns in the data. `Max_depth`=3 has the highest RMSE compared to `max_depth`=5 and `max_depth`=7, indicating that trees that are too shallow (low depth) are less capable of capturing data complexity, resulting in greater prediction errors. `Max_depth`=7 shows the lowest RMSE value, meaning the model is better at predicting because it can capture more complex patterns in the data. `Max_depth`=5 is in between, with slightly better performance than `max_depth`=3 but slightly worse than `max_depth`=7. It can be concluded that a higher `n_estimators` value (around 1000) generally provides lower RMSE results, indicating better model performance. Higher `max_depth` also tends to result in lower RMSE, making `max_depth`=7 the best among the three, as shown in Table 4. The optimal hyperparameter combination resulted in $R^2 = 0.524$, $MSE = 2788.1808$, $RMSE = 52.8032$, $MAE = 41.5819$. Previous research used XGBoost, SVR, RFR. In XGBoost with `n_estimators`=1800, `max_depth`=20, `min_child_weight`=3, `gamma`=0.2, `subsample`=0.7, `colsample_bytree`=0.8, `reg_alpha`=0.1, `reg_lambda`=0, `learning_rate`=1, it produced $R^2 = 0.4802$, $RMSE = 62.6765$. SFR with `n_estimators`=1000, `max_depth`=22, `min_samples_leaf`=3, `min_samples_split`=3, `max_features`=24 yields $R^2 = 0.4732$, $RMSE = 63.1013$. SVR with $C=1$, $gamma=0.00$ yields $R^2 = 0.4660$, $RMSE = 63.4415$ (Liu et al., 2022).

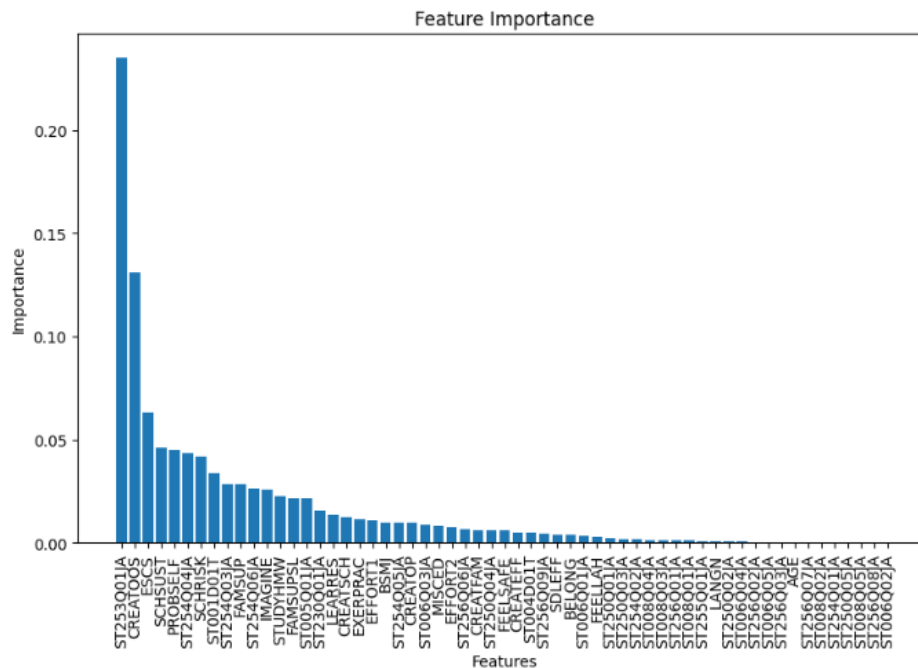


Figure 4. Features importance plot of GBDT

Figure 4 shows the variables that contribute the most to making predictions in the GBDT model. The first variable is ST253Q01JA (student access to technology at home), followed by CREATOOS (creativity outside the school context), ESCS (economic, social, and cultural status index), SCHSUST (School Sustainability), which refers to the school's efforts to educate students about the importance of environmental sustainability, adopt eco-friendly practices, and integrate sustainability education into the curriculum, PROBSSELF (Problem Solving Self-Efficacy), which measures students' self-belief in their ability to solve problems. This variable is designed to assess how confident students feel in facing and solving complex problems independently. The sixth variable is ST254Q04JA (students' perceptions of their literacy skills or learning environment), followed by SCHRISK (School Risk), which measures students' or teachers' perceptions of risks in the school environment. As the list continues, the importance values of the variables decrease, approaching zero or indicating no significant contribution to predictions.

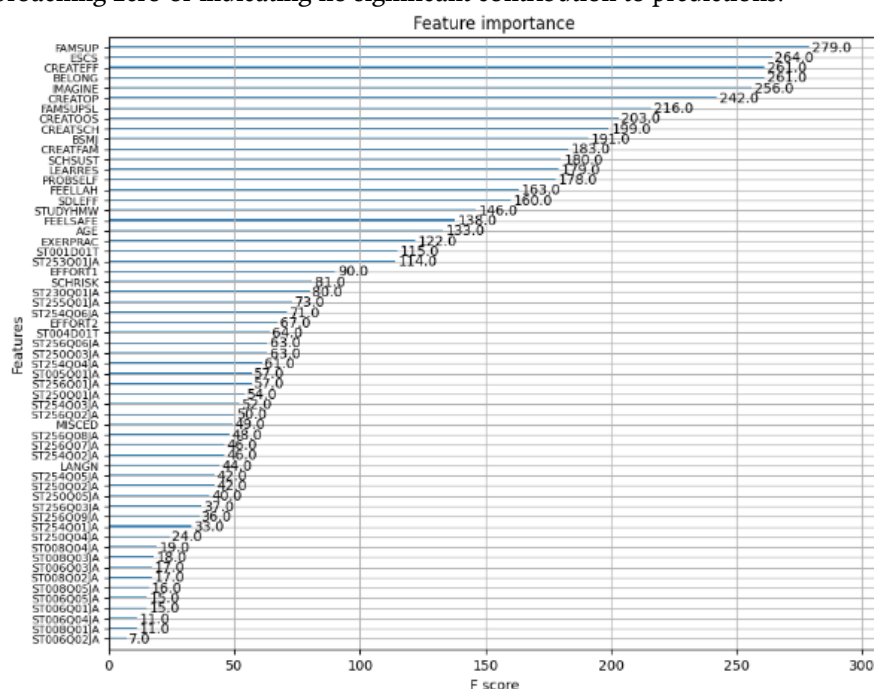


Figure 5. Features importance plot of XGBoost

Figure 5 shows the variables that contribute the most to making predictions in the XGBoost model. FAMSUP (Family Support) emerges as the most influential variable, followed by ESCS (economic, social, and cultural status index). The third variable is BELONG (Sense of Belonging), which measures students' sense of belonging and attachment to the school, evaluating how accepted, supported, and comfortable students feel in their school environment. The fourth variable is CREATEFF (Creative Self-Efficacy), which refers to students' confidence in their ability to think creatively and generate new ideas. The fifth variable is IMAGINE, which assesses students' ability in creative imagination or imaginative thinking. This variable measures how often and to what extent students use their imagination in creative thinking, both in school learning and everyday life. The sixth variable is CREATOP (Creative Problem-Solving Orientation), which refers to students' tendency to approach problems creatively.

CONCLUSION

Compared to GBDT, the XGBoost model performs better in terms of prediction, with an R^2 value of 0.5247, which is higher than the R^2 of GBDT, which reaches 0.5106. Additionally, XGBoost exhibits lower MSE (2788.1808), RMSE (52.8032), and MAE (41.5819) compared to GBDT, indicating that XGBoost is more accurate and effective in predicting students' reading literacy in PISA 2022.

This study also successfully identified significant variables that contribute to reading literacy through both models. The analysis results from GBDT indicate that variables such as ST253Q01JA (student access to technology at home), CREATOOS (creativity outside the school context), ESCS (economic, social, and cultural status index), SCHSUST (school sustainability efforts), PROBSELF (student self-efficacy in problem-solving), ST254Q04JA (students' perception of their literacy or learning environment), SCHRISK (perceived risks at school), ST001D01T (educational level), ST254Q03JA (learning motivation), and FAMSUP (family support in learning) play an important role in supporting student reading literacy performance, both directly and through their contribution to the learning environment and character development.

Meanwhile, in the XGBoost model, variables such as FAMSUP (family support), ESCS (socioeconomic and cultural status), BELONG (sense of belonging at school), CREATEFF (creative thinking self-efficacy), IMAGINE (creative imagination), and CREATOP (creative problem-solving orientation) play key roles in supporting students' reading literacy abilities. Additionally, other variables such as CREATOOS (creativity outside of school), CREATSCH (creative climate at school), BSMJ (school membership and happiness), and SCHSUST (school sustainability) indicate that the social and psychological environment of the school has a significant impact on fostering students' reading skills.

Overall, both models show that besides family support and socioeconomic status, psychosocial aspects such as the creative climate in schools, sense of attachment, and emotional support in the learning environment play a significant role in predicting students' reading literacy. These results provide practical implications for policymakers and educators to prioritize learning environments that support creativity, emotional support, and family involvement in the learning process. Further research needs to focus on tuning the parameters of the XGBoost and GBDT models to improve accuracy, such as optimizing the learning rate, max depth, and number of estimators. Additionally, the addition of new variables related to technology use and social interactions at school could enrich the model and improve reading literacy predictions. With further experimentation on these parameters and additional variables, it is expected that the model's performance can be enhanced and provide deeper insights into the significant factors contributing to predicting students' reading literacy.

ACKNOWLEDGEMENTS

This research would not have been possible without the invaluable support of my advisor and the institution. I would like to express my deepest gratitude to my supervisor, Prof. Ema Utami, S.Si., M.Kom, for her continuous guidance, insightful feedback, and encouragement

throughout the course of this research. Her expertise and direction have been instrumental in shaping the outcomes of this study.

I would also like to extend my sincere thanks to the faculty members of Universitas Amikom Yogyakarta for the knowledge and academic guidance they provided during my studies, which greatly contributed to the success of this research.

REFERENCES

- Ahmad, G. N., Fatima, H., Shafiullah, Salah Saidi, A., & Imdadullah. (2022). Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access*, 10(August), 80151–80173. <https://doi.org/10.1109/ACCESS.2022.3165792>
- Breiman. (2019). Classification and Regression Trees. In *Sustainability (Switzerland)* (Vol. 11, Issue 1). http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI
- Bu, Y., & Chen, F. (2023). What key contextual factors contribute to students' reading literacy among top-performing countries and economies? Statistical and machine learning analyses. *International Journal of Educational Research*, 122(July), 102267. <https://doi.org/10.1016/j.ijer.2023.102267>
- Dilekçi, A. (2022). Evaluation of Turkey's PISA Reading Literacy Scores. *International Journal of Education and Literacy Studies*, 10(1), 138. <https://doi.org/10.7575/aiac.ijels.v.10n.1p.138>
- Farhangi, F., Sadeghi-Niaraki, A., Razavi-Termeh, S. V., & Choi, S. M. (2021). Evaluation of tree-based machine learning algorithms for accident risk mapping caused by driver lack of alertness at a national scale. *Sustainability (Switzerland)*, 13(18). <https://doi.org/10.3390/su131810239>
- Guoqing Chen, T. Z. (2023). Combined Prediction Model of Gasoline Octane Loss Value Based on XGboost-GBDT Combined Prediction. *Sustainability (Switzerland)*, 11(1), 1–14. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI
- Su, Q., Chen, L., & Qian, L. (2024). Optimization of big data analysis resources supported by XGBoost algorithm: Comprehensive analysis of industry 5.0 and ESG performance. *Measurement: Sensors*, 36(October), 101310. <https://doi.org/10.1016/j.measen.2024.101310>
- Katagiri, K., & Fujii, T. (2022). Partitioned Path Loss Models Based on Coefficient of Determination. *International Conference on Information Networking, 2022-Janua*, 198–203. <https://doi.org/10.1109/ICOIN53446.2022.9687163>
- Kuan, S. I., Kim, J., Kwon, O. H., & Song, H. J. (2022). Canopy $\frac{1}{2}$ K-means Combined Collaborative Filtering Using RMSE-minimization. *Proceedings - 2022 IEEE International Conference on Big Data and Smart Computing, BigComp 2022*, 31–34. <https://doi.org/10.1109/BigComp54360.2022.00016>
- Liu, H., Chen, X., & Liu, X. (2022). Factors influencing secondary school students' reading literacy: An analysis based on XGBoost and SHAP methods. *Frontiers in Psychology*, 13(September). <https://doi.org/10.3389/fpsyg.2022.948612>
- Mahdavi, M. (2024). Efficient Hardware Acceleration of Mean Squared Error Calculation

- Through In-Memory Computing. *2024 IEEE International Conference on Omni-Layer Intelligent Systems, COINS 2024, Imc*, 1–6. <https://doi.org/10.1109/COINS61597.2024.10622442>
- Moitra, A. (2024). *MSBoost : Using Model Selection with Multiple Base Estimators for Gradient Boosting*.
- Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLoS ONE*, 18(2 February), 1–8. <https://doi.org/10.1371/journal.pone.0279774>
- Shen, E. Z. (2022). Short-time cab speed prediction model based on XGBoost. *Research Square*, 1, 0–13.
- Smiti, S., Soui, M., & Ghedira, K. (2024). Tri-XGBoost model improved by BLSmote-ENN: an interpretable semi-supervised approach for addressing bankruptcy prediction. *Knowledge and Information Systems*, 66(7), 3883–3920. <https://doi.org/10.1007/s10115-024-02067-w>
- State, T. (2022). *PISA 2022 Result The State of Learning and Equity in Education: Vol. I* (Issue 2).
- Tağa, T. (2023). *What does PISA Assess in Reading Literacy ? Misconceptions and Misuses. c*.
- Wang, S., & Ma, J. (2023). A novel GBDT-BiLSTM hybrid model on improving day-ahead photovoltaic prediction. *Scientific Reports*, 13(1), 1–13. <https://doi.org/10.1038/s41598-023-42153-7>
- Wang, Y., King, R., Haw, J., & Leung, S. on. (2023). What explains Macau students' achievement? An integrative perspective using a machine learning approach (¿Cuál es la explicación del rendimiento de los estudiantes macaenses? Una perspectiva integradora mediante la adopción del enfoque del aprendizaje aut. *Infancia y Aprendizaje*, 46(1), 71–108. <https://doi.org/10.1080/02103702.2022.2149120>
- Wang, Y., Yan, Z., & Xing, L. (2021). A Movie Score Prediction Model Based on XGBoost Algorithm. *Proceedings - 2021 International Conference on Culture-Oriented Science and Technology, ICCST 2021*, 486–491. <https://doi.org/10.1109/ICCST53801.2021.00108>