Analysis and Visualization of High School Student Achievement Data Using Decision Tree and Cross-Validation in Rapidminer

Radhitya Yunandri Hartanta[1]*, Asroni[2], Slamet Riyadi[3], Jeckson[4]

[1,2,3]*Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia*

[4]*Universitas Muhammadiyah Lampung, Pagar Alam, Labuhan, Labuhan Ratu, Kec. Kedaton, Kota Bandar Lampung, Lampung 35132*

**Corresponding author: radhitya.y.ft15@mail.umy.ac.id*

Abstract

The vision, mission, and work indicators of high school education are all geared toward producing graduates of excellent quality. Students' academic performance in the first and second grades indicates their eventual excellence as graduates. However, it is not feasible to ensure that students in the third year have good results only by comparing their academic achievements during the first and second grades. School leaders and policymakers can benefit from data analysis and visualization by gaining insight into recurring patterns and emerging trends in their stored data. Leaders and management at schools can benefit significantly from Rapidminer's decision tree and cross-validation data analysis methods when trying to figure out what to do about students performing well or poorly academically.

Keywords: *High School Student Achievement, Data Visualization, and Decision Tree*

1. Introduction

Data are essential to the daily operations of any organization, agency, or institution, from the routine transactional data stored in relational databases to the telemetry data on services provided to the data obtained from fields, such as social media and stored in a database, data warehouse, or big data. These data are precious when they can be transformed into diagnostic, predictive, or prescriptive information. Having information is insufficient. In order to make improvements to their operations, high schools need the ability to take data-driven action. Whether redistributing funds to meet pressing demands, recognizing failing strategies, and pivoting to a new course of action, agility is essential. It explains why it is crucial to utilize data to generate compelling stories. Using data for diagnosis and predictions is also an issue in high schools. Despite the importance of students' final grades to their academic and professional futures, the school has not identified, characterized, or predicted the factors contributing to students' success and graduation.

2. Method

The initial step of this study required students' demographic data and test scores from their first through last years of schooling. The subsequent step was to analyze the information in the data source to make predictions about the graduation rate of students. The mining process in this study was carried out using a decision tree algorithm. It is envisaged that more realistic predictions could be made using the cross-validation method. For students whose G1 and G2 grades were included in the prediction, the results were expressed as condition scores or G3 (designated third-year grades). The vast quantity of data currently available is a defining feature of the modern era. Tools with predefined functionality exist for processing and visualizing massive data sets using a variety of graphical representations. Since the human brain analyzes images or visuals faster than words, data visualization is a considerably more efficient method of obtaining information than analysis performed on numeric data [1], [2]. Data visualization is a method solution for conveying complex information using the human visual system to facilitate comprehension [3]. This visualization yields an efficient and visually appealing graph well-suited to the data and is compatible with mobile platforms[4], [5].

A high level of expertise is required to conduct the required predictive analysis, data mining, text analytics, and data preparation [6]–[8]. In reality, a data scientist would be regarded as a super coder with a Ph.D. in statistics and an in-depth familiarity with every conceivable business challenge. Experts in this field were naturally hard to come by. Thus, Rapidminer was developed to allow laypeople to draw the same conclusions as data scientists [9]. It helps make it simpler for people to implement the solutions and discover new ones. To help businesses move more quickly, Rapidminer equips business analysts and data scientists with the tools to unearth previously unseen patterns and advantages. As a result, there is a chance for a ton of profit in the market [10], [11]. Rapidminer was employed in the following way to analyze and visualize prediction data for high school students.

1. Obtaining the indicators required for analysis regarding the grades of high school students
2. Collecting supporting data
3. Analyzing the data obtained using Rapidminer with the decision tree and cross-validation methods to analyze the factors influencing students' final grades
4. Acquiring a visualization of the decision tree method for use by writers and those who need data and information to facilitate observation of the factors influencing students' final grades

3. Results and Discussion

The performance of students of Gabriel Pereira and Mousinho da Silveira High School [12] was analyzed using data derived from https://www.kaggle.com/datasets/larsen0966/student-performance-data-set. The investigation involved 395 data with the following columns.

Table 1 Columns in Student Data

Column Name	Description
Error	Not attending class
Roll call	Attending class
G1	Period 1 score
G2	Period 2 score
Condition	Predicted score

The final value or G3 in Table 1 was given the condition label by changing the G3 value to the following condition.

1. 0-15 : Failed
2. 16-30 : Passed

Subsequently, the data were analyzed using Rapidminer's decision tree method to identify the factors that most impacted the condition.

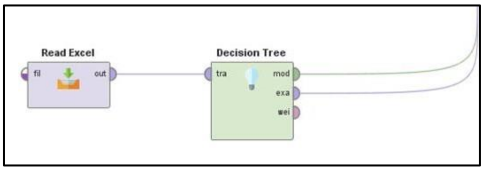


Figure 1. Decision Tree Process in Rapidminer

Figure 2 depicts the results obtained from the decision tree.

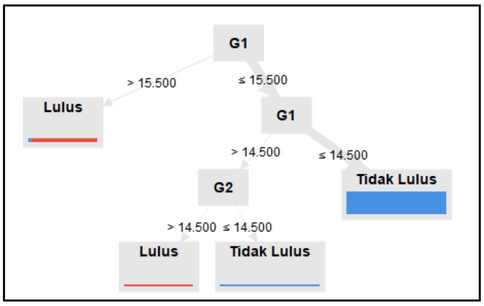
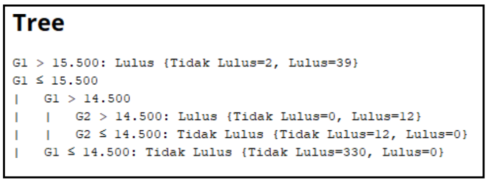


Figure 2. Decision Tree Results

Students predicted to pass with a G1 score of higher than 15.5 had 39 correct predictions and two incorrect ones considered failed. Figure 3 describes the results.

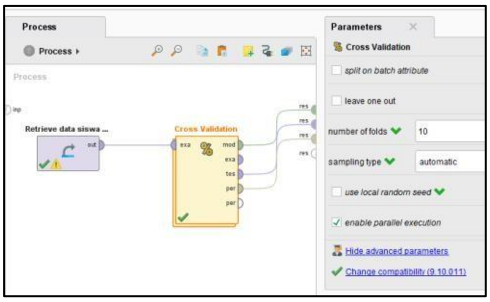


Tree

```
G1 > 15.500: Lulus {Tidak Lulus=2, Lulus=39}
G1 ≤ 15.500
|   G1 > 14.500
|   |   G2 > 14.500: Lulus {Tidak Lulus=0, Lulus=12}
|   |   G2 ≤ 14.500: Tidak Lulus {Tidak Lulus=12, Lulus=0}
|   G1 ≤ 14.500: Tidak Lulus {Tidak Lulus=330, Lulus=0}
```

Figure 3. Description of Decision Tree results in Rapidminer

Cross-validation was performed ten times to provide more accurate validation results, forming a process in Figure 4.



□ split on batch attribute

□ leave one out

number of folds ✓ 10

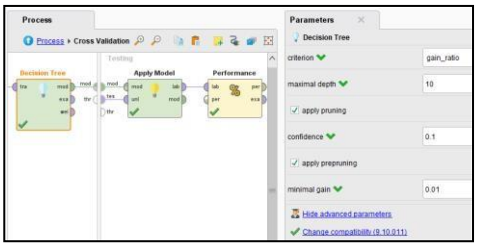
sampling type ✓ automatic

□ use local random seed ✓

☑ enable parallel execution

Figure 4. Decision Tree and Cross-Validation

Figure 5 exhibits the details of the cross-validation process.



criterion gain_ratio
maximal depth 10
☑ apply pruning
confidence 0.1
☑ apply prepruning
minimal gain 0.01
Hide advanced parameters
Change compatibility (9.10.011)

Figure 5. Cross Validation

The experiment continued using the decision tree and cross-validation methods. Figure 6 illustrates the experimental results.

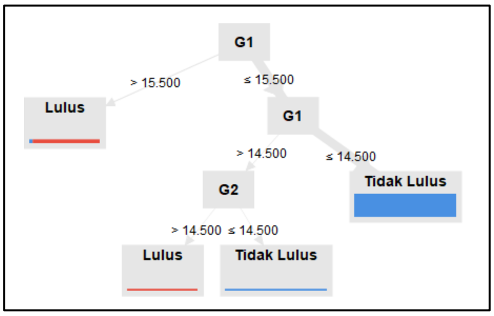
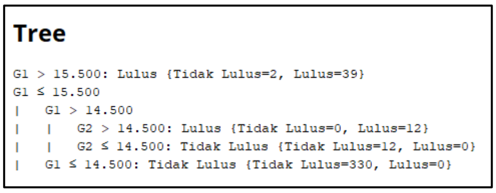


Figure 6. Decision Tree and Cross-Validation

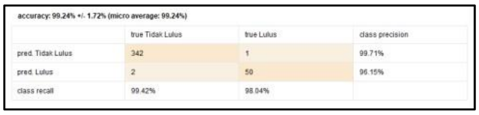
It turned out that both the decision tree method with and without cross-validation produced identical results when tested with 395 data and the same factors. These results are described in Figure 7.


Tree

```
G1 > 15.500: Lulus {Tidak Lulus=2, Lulus=39}
G1 ≤ 15.500
|   G1 > 14.500
|   |   G2 > 14.500: Lulus {Tidak Lulus=0, Lulus=12}
|   |   G2 ≤ 14.500: Tidak Lulus {Tidak Lulus=12, Lulus=0}
|   G1 ≤ 14.500: Tidak Lulus {Tidak Lulus=330, Lulus=0}
```

Figure 7. Description of Decision Tree and Cross-Validation

The confusion matrix table additionally demonstrates the performance of cross-validation. It achieved an accuracy of 99.24% with a standard deviation of 1.72%. Figure 8 exhibits an accuracy rate of 99.71% for the 342 correct failed predictions and one incorrect prediction. Moreover, there were 50 correct passed predictions and two incorrect predictions, generating an accuracy rate of 96.15%, as displayed in Figure 8.


accuracy: 99.24% +/- 1.72% (micro average: 99.24%)

	true Tidak Lulus	true Lulus	class precision
pred. Tidak Lulus	342	1	99.71%
pred. Lulus	2	50	96.15%
class recall	99.42%	98.04%	

Figure 8. Performance Vector and Confusion Matrix

4. Conclusion

Research began with data collection, followed by processing the data using Rapidminer tools. This study employed the decision tree and cross-validation methods to process the data and obtain accurate prediction results. The goal was to determine factors affecting students' grades and how well these methods could predict their grades based on their current grades and demographic information.

References

[1] M. A. Borkin *et al.*, "What makes a visualization memorable?," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[2] Z. Bylinskii *et al.*, "Learning Visual Importance for Graphic Designs and Data Visualizations," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, Québec City QC Canada: ACM, Oct. 2017, pp. 57–69. doi: 10.1145/3126594.3126653.

[3] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, "The Science of Visual Data Communication: What Works," *Psychol Sci Public Interest*, vol. 22, no. 3, pp. 110–161, Dec. 2021, doi: 10.1177/15291006211051956.

[4] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, "TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility," in *ACM SIGGRAPH 2003 Papers*, San Diego California: ACM, Jul. 2003, pp. 453–462. doi: 10.1145/1201775.882291.

[5] L. Nair, S. Shetty, and S. Shetty, "Interactive visual analytics on Big Data: Tableau vs D3. js," *Journal of e-Learning and Knowledge Society*, vol. 12, no. 4, 2016, Accessed: Dec. 04, 2023. [Online]. Available: https://www.learntechlib.org/p/173675/

[6] D. Tao, P. Yang, and H. Feng, "Utilization of text mining as a big data analysis tool for food science and nutrition," *Comp Rev Food Sci Food Safe*, vol. 19, no. 2, pp. 875–894, Mar. 2020, doi: 10.1111/1541-4337.12540.

[7] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 1, 2020.

[8] F. Martínez-Plumed *et al.*, "CRISP-DM twenty years later: From data mining processes to data science trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, 2019.

[9] P. Ristoski, C. Bizer, and H. Paulheim, "Mining the web of linked data with rapidminer," *Journal of Web Semantics*, vol. 35, pp. 142–151, 2015.

[10] A. Mammadov, "DERIVATION OF PRESCRIPTIVE ACCIDENT PREVENTION MODEL FROM PREDICTIVE MODELS USING ML ALGORITHMS," Master's Thesis, Middle East Technical University, 2021. Accessed: Dec. 04, 2023. [Online]. Available: https://open.metu.edu.tr/handle/11511/95191

[11] H. Abusnina, "Combining engineering and data-driven approaches to model the risk of excavation damage to underground natural gas facilities," PhD Thesis, Rutgers University-School of Graduate Studies, 2019. Accessed: Dec. 04, 2023. [Online]. Available: https://rucore.libraries.rutgers.edu/rutgers-lib/61672/

[12] D. A. Petrusevich, "Clustering of secondary school students in Portugal," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 012140. Accessed: Dec. 04, 2023. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1691/1/012140/meta