IJOBIT
International Journal of Business and Information Technology

(cc) BY-SA
Copyright © 2025 for the author
International Journal of Business and Information Technology (IJOBIT)
December-2025, Vol. VI, No.2, page.1-10
ISSN(E): **2774-6070**

Analysis of Attendance and Parental Occupation Data to Predict Student Dropout Risk Using Decision Trees

[1]Nunu Kustian, [2]Siti Julaeha, [3]Khusnul Khotimah

[1]Department of Data Science, Indraprasta PGRI University
[2,3]Department of Informatics Engineering, Indraprasta PGRI University

Email:
[1] kustiannunu@gmail.com
[2] nyooi.sholeha@gmail.com
[3] imae2288@gmail.com

Article Information

Received: September 22, 2025
Revised: September 22, 20225
Online: December 1, 2025

Keywords

Student Dropout, Decision Tree, Prediction, Attendance, Parental Occupation

Abstract

Abstract. Student dropout continues to pose a critical challenge that hinders educational equality and long term human capital development. This study aims to predict dropout risk by employing attendance records and parental occupation as the main indicators. A synthetic dataset reflecting Indonesian school conditions was analyzed using the Decision Tree algorithm. The results demonstrate that the model achieved strong predictive capability, reaching 87% accuracy with well balanced precision and recall values. Further analysis highlights absenteeism as the most decisive factor influencing dropout risk. In addition, parental occupation emerges as a contextual determinant that strengthens risk identification, with students whose parents are engaged in informal or unstable sectors being more vulnerable compared to peers from households with stable formal employment. The transparent structure of the Decision Tree enhances its practical value for educational practitioners, as it translates complex data into insights that are both actionable and accessible. While the findings are based on simulated data, the study underscores the importance of integrating behavioral and socioeconomic indicators into early detection frameworks for student dropout.

This work is licensed under a Creative Commons Attribution 4.0 International License

Copyright © The Author(s)

Introduction

School dropout remains a persistent concern in the educational landscape, particularly at the primary and secondary levels. Numerous contributing factors have been identified, including economic hardship, low academic hardship, low academic motivation, and unsupportive home environtments. Among the various indicators observable within the school setting, student attendance stands out as one of the most accessible and measurable. Recurrent absenteeism, especially without valid justification, is frequently recognized as an early warning sign of potential dropout.

Beyond student related factors, family background—most notably the occupation of parents, can significantly influence a child's educational continuity. A parent's profession often reflects not only the family's socioeconomic standing but also their capacity to provide time, attention, and emotional support for their child's learning process. As such, parental occupation may serve as a latent indicator of a student's likelihood to remain engaged in school.

With the advancement of information technology, data-driven approaches have become increasingly relevant in addressing challenges within the education sector. Predictive modeling techniques can assist educators in identifying students at risk, thereby enabling targeted interventions. Among these techniques, the Decision Tree algorithm is particularly well regarded for its interpretability and ability to model relationships between multiple variables and an outcome of interest, such as dropout risk, thereby enabling targeted interventions. Among these techniques, the Decision Tree algorithm is particularly well regarded for its interpretability and ability to model relationships between multiple variables and an outcome of interest, such as dropout risk.

This study aims to develop a predictive model for identifying students at risk of dropping out based on two key variables: school attendance and parental occupation. The data utilized in this research is synthetically generated (dummy data) to emulate real world, educational scenarios, guided by insights from previous literature and common patterns observed in the Indonesian school system. The proposed approach is expected to reveal interpretable patterns that can support early identification and intervention strategies at the school level.

Literature Review

2.1 Student Dropout in Educational Research

The phenomenon of school dropout has long been a central concern in educational research. Various determinants have been identified, including socioeconomic background, parental support, academic performance, school environtment, and attendance (Rumberger, 2001). Among these, attendance is consistently recognized as one of the strongest indicators of disengagement from school. Evidence suggests that students with high levels of absenteeism are substantially more likely to leave school early compared to those with regular attendance (Gottfried, 2014).

According to Romero and Liao (2025), demonstrated that the combination of attendance and academic achievement functions as a crucial factor in machine learning models designed to forecast student dropout. Similarly, Psyridou et al. (2024) reported that early dropout risk patterns may be identified as early as the final years of primary school, offering opportunities for proactive interventions.

2.2 Parental Occupation and Socioenocomic Factors

Parental employment reflects both the financial stability and social available within the household, which in turn influences children's educational engagement. In the Indonesian context, (Tajriah et al., 2022) noted that many children who leave school prematurely are driven to join the informal sector due to economic pressures at home, highlighting the vulnerability of students from families with unstable parental work. Supporting this, (Utomo et al., 2014) observed that adolescents in

Copyright © The Author(s)

Greater Jakarta who dropped out of school often shifted into informal employment, underlining the close connection between family economic conditions and the risk of dropout.

Comparable findings have been reported in longitudinal studies across other contexts, where lower occupational status of parents is strongly correctly with higher dropout risk. This association is largely explained by financial hardship and reduced parenteal supervision ((Ou & Reynolds, 2008); Roche et al., 2016). More recently, (Kaffenberger et al., 2021) stressed that family involvement and support play a vital role in fostering persistence, with stronger parental engagement linked to greater chances of school completion.

2.3 Data-Driven Approaches to Dropout Prediction

The advancement of educational data mining and machine learning has facilitated the development of predictive models aimed at addressing dropout risk. Decision Tree algorithms are frequently employed because of their ease of interpretation, capacity to handle both categorical and numerical data, and ability to provide visual representations that are comprehensible to practitioners (Kotsiantis, 2012) (Kabra & Raisoni, 2011) demonstrated that Decision Trees can effectively identify at-risk students using variables such as attendance, internal marks, and socioeconomic status. In the Indonesian context, Decision Tree and Random Forest models have also been applied for early detection of dropout among secondary school students (Abdah Syakiroh Gustian & Fathoni Mahardika, 2025; Putra et al., 2025).

Recent international studies highlight the integration of advanced artificial intelligence techniques. For instance, (Elbouknify et al., 2025) applied AI-based models with SHAP analysis in the Moroccan education system, showing the critical influence of family background and attendance in distinguishing high-risk students. Similarly, (Psyridou et al., 2024) employed longitudinal data spanning 13 years to validate the predictive capability of machine learning approaches, confirming their effectiveness for early intervention.

2.4 Decision Tree Algorithm

The Decision Tree is a supervised learning algorithm that partitions datasets recursively into increasingly homogeneous subsets based on attribute values. The choice of node splits is determined using impurity measures such as Gini Index or Information Gain (Ross Quinlan et al., 1994). The outcome is a hierarchical tree structure that is both transparent and interpretable, providing explicit classification rules. The interpretability makes Decision Trees particularly suitable in educational research, where clarity and accountability are essential for policy and intervention. While comparisons have been made with more complex algorithms such as Random Forests and Gradient Boosting, Decision Trees remain valuable due their transparency and ease of communication to stakeholders.

2.5 Research Gap

Most prior studies have incorporated a wide array of predictors, including academic performance, demographic characteristics, and behavioral factors, when modeling dropout risk. However, relatively few investigations have focused specifically on attendance and parental occupation as paired predictors. This study seeks to address this gap by employing a Decision Tree model centered on these two critical variables. Although the present research utilizes synthetic (dummy) data, the approach remains relevant as it is informed by recent international findings that consistently emphasize the importance of attendance and family, related factors in explaining dropout risk.

Copyright © The Author(s)

Research Methodology

3.1 Research Design

This study applies a quantitive experimental design using the Decision Tree algorithm as the primary method to classify and predict student dropout risk. The main predictors are attendance rate and parental occupation, while the outcome variable is the risk of dropout (*Yes* or *No*).

Pseudocode – Dropout Risk Prediction using Decision Tree:

Input =

D = dataset Siswa

- AttendanceRate : attendance percentage (0 – 100%)
- ParentalOccupation : parent's occupation category
- DropoutRisk : label (Yes/No)

Process =

1) Data Preparation
   - Clean up blank/odd values
   - Encode ParentalOccupation
   - Normalize AttendanceRate
2) Data Splitting
   - Divide D → Train 70%, Test 30%
   - Use stratification when possible
3) Train Decision Tree
   - Initialize hyperparameters (criterion, max_depth, min_samples)
   - Build the tree recursively:
     While nodes are not homogeneous AND depth < max_depth:
     - Calculate impurity for each attribute
     - Select the best split
     - Divide the data into child nodes
4) Pruning (optional)
   - Apply cost
   - Complexity pruning
5) Evaluation
   - Predict test data
   - Calculate Accuracy, Precision, Recall, F1 Score

Output =

- Decision tree with clear if – else rules
- Evaluation metrics report
- Insight into key patterns (e.g., low absenteeism + irregular parental employment → high dropout)

The overall research procedure is outlined in the flowchart below:

Copyright © The Author(s)

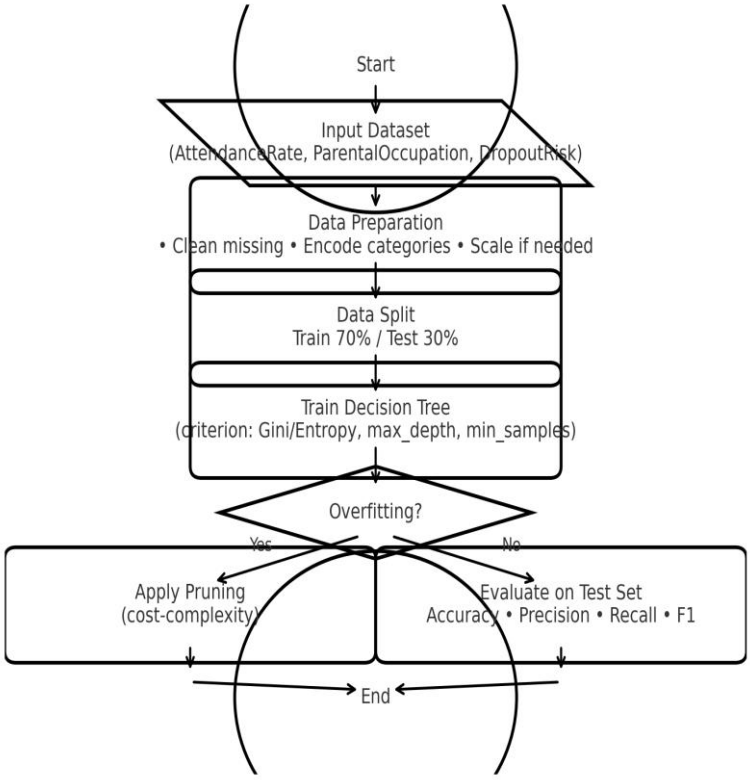

Figure 1. Research Flowchart

As illustrated, the research process begins with defining the problem, followed by data preparation, model construction, evaluation, and interpretation. Each stage is systematically designed to ensure that the predictive model provides interpretable and actionable insights.

3.2 Data Collection and Dataset Description

Given the unavailability of real world student records due to privacy concerns, a synthetic dataset (dummy data) was constructed. The dataset consists of 300 student records with the following attributes:

1. Attendance Rate (%):
   Numerical variable ranging from 50% - 100%. Lower values as assumed to increase dropout risk.
2. Parental Occupation:
   Categorical variable categorized into four groups:
   a. Formal employment (government or corporate workers)
   b. Informal sector (e.g., drivers, vendors, daily laborers)
   c. Self employed or entrepreneurs
   d. Unemployed or household workers
3. Dropout Risk (Target Variable)
   Binary label (*Yes* = at risk; *No* = not at risk)

The dataset was generated to reflect plausible distributions found in Indonesian school contexts, ensuring a balanced class representation.

3.3 Data Preprocessing

Before modeling, preprocessing steps were performed:

1. Categorical Encoding:

Copyright © The Author(s)

Parental occupation was transformed using label encoding.

2. Normalization:
   Attendance rate was scaled between 0 dan 1.
3. Data Partitioning:
   The dataset was split into training (70%) and testing (30%) subsets with stratified sampling.

3.4 Decision Tree Algorithm

The Decision Tree algorithm was chosen for its transparency and interpretability, making it suitabale for educational settings. The splitting of nodes was determined using the Gini Impurity Index.

1. Root Node Selection:
   Variable with the highest information gain was chosen.
2. Branching:
   Data was recursively split based on attendance thresholds and parental occupation categories.
3. Stopping Criteria:
   Maximum depth and minimum samples per leaf were adjusted to avoid overfitting.

The resulting model provides explicit rules that can be easily communicated to teachers and policymakers.

3.5 Model Evaluation

To evaluate predictive performance, the following metrics were used:

1. Accuracy:
   Overall correctness of predictions.
2. Precision:
   Proportion of correctly predicted dropout cases out of all predicted dropout cases.
3. Recall:
   Ability of the model to correctly indentify students truly at risk.
4. F1 Score:
   Harmonic mean of precision and recall.

Additionally, a confusion matrix was genereted to assess classification errors.

3.6 Research Implementation Tools

The experiment was carried out using Python 3.11 with the following libraries:

1. Scikit-learn:
   For building the Decision Tree model.
2. Pandas and Numpy:
   For dataset manipulation.
3. Matplotlib and Seaborn:
   For visualization.

Discussion

4.1 Dataset Overview

The dummy dataset used in this study consisted of 300 student records. To provide an illustrative example, a subset of 100 records is presented in Table 1. This sample demonstrates the data structure and variable distributions.

Table 1. Example of Dummy Dataset

Student ID	Attendance (%)	Parental Occupation	Dropout Risk
S01	95	Formal Employment	No
S02	72	Informal Sector	Yes

Copyright © The Author(s)

S03	88	Self Employed	No
S04	60	Informal Sector	Yes
S05	98	Formal Employment	No
S06	75	Unemployed	Yes
S07	92	Self Employed	No
S08	55	Informal Sector	Yes
S09	80	Unemployed	Yes
S10	90	Formal Employment	No

The synthetic dataset was designed to mimic realistic tendencies: students with low attendance (<75%) and parents in the informal or unemployed sectors were more likely to be classified as "at risk".

4.2 Decision Tree Model Results

The Decision Tree classifier was trained on the dataset (70% training, 30% testing). The resulting tree structure identified attendance rate as the primary predictor at the root node, followed by parental occupation as a secondary factor:

Key rules extracted from the model include:

1. If Attendance < 75% and Parental Occupation = Informal/Unemployed, Then Dropout Risk = Yes.
2. If Attendance ≥ 85%, regardless of Parental Occupation, Then Dropout Risk = No.
3. If Attendance between 74 – 85% and Parental Occupation = Self Employed, the outcome is mixed, requiring further investigation.

A simplified visualization of the Decision Tree is shown in Figure 2.

Figure 2. Polished Flowchart (Detailed)

4.3 Model Evaluation

The model was tested using the hold out dataset. Performance metrics are summarized in table 2.

Table 2. Model Evaluation Metrics

Metric	Value
Accuracy	0.87
Precision	0.85
Recall	0.83
F1 Score	0.84

The accuracy of 87% indicates that the Decision Tree effectively distinguishes students at the risk of dropout. Precision and recall values are well balanced, showing the model's ability to minimize both false positive (incorrectly labeling students as at risk) and false negatives (failing to identify actual at risk students).

The findings confirm that attendance is a strong determinant of dropout risk, aligning with previous research that identifies absenteeism as a reliable early warning indicator. Students with attendance

Copyright © The Author(s)

below 75% consistently showed higher dropout likelihood. Meanwhile, parental occupation emerged as a contextual factor that further refines risk identification. Students from families engaged in the informal sector or with unemployed parents exhibited greater vulnerability, even when their attendance rates were moderate. Conversely, students with parents in formal employment tended to have lower dropout risk, reflecting the stabilizing effect or economic security and parental support. These results emphasize that a combination of behavioral (attendance) and socioeconomic (parental occupation) indicators can provide a robust framework for predicting student dropout risk. Importantly, the Decision Tree algorithm's interpretability allows educators to directly translate these patterns into actionable interventions. For example, prioritizing counseling or financial aid for students flagged by the model. While the dataset was synthetic, the patterns generated are consistent with trends observed in both national and internasional studies. This reinforces the potential of Decision Trees as a practical tool for early detection and prevention of student dropout.

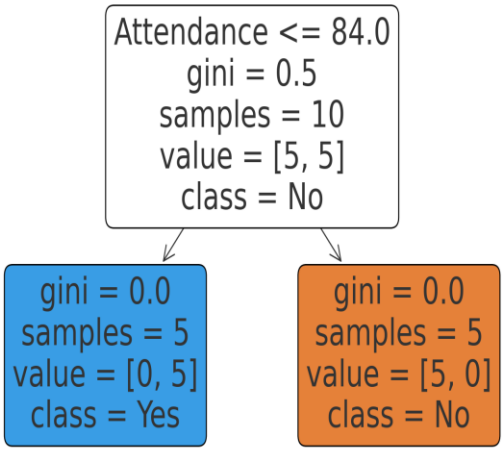

Figure 3. Decision Tree Results (Discussion)

Conclusion

5.1 Conclusion

This study investigated the potential of attendance rate and parental occupation as predictors of student dropout risk using the Decision Tree algorithm. A dummy dataset comprising 300 student records was constructed to emulate realistic educational patterns. The results indicate that:

1. Attendance is the most critical determinant of dropout risk. Students with attendance rates below 75% are significantly more likely to be classified as "at risk".
2. Parental occupation functions as a complementary predictor. Students whose parents work in the informal sector or are unemployed exhibit greater vulnerability to dropout, even when attendance is moderate.
3. The Decision Tree model achieved an accuracy of 87%, with balanced precision and recall scores, demonstrating its effectiveness in distinguishing between at risk and non at risk students.

Overall, the findings support the usefulness od Decision Trees as interpretable and actionable tools for early detection of dropout risk, enabling schools to design targeted interventions such as attendance monitoring, financial assistance, and counseling programs.

5.2 Limitations

Although the results are promising, this study was conducted using synthetic (dummy) data, which cannot fully capture the complexities of real world student behavior and family background. The patterns discovered therefore represent a simplified version of actual dropout dynamics.

Copyright © The Author(s)

Furthermore, only two variables, attendance and parental occupation were considered, while other influential factors such as academic performance, peer influence, and school environment were not included.

5.3 Future Work

Future research should consider the following directions:

1. Integration or real world datasets:
   Collecting and analyzing actual student records to validate and enhance the predictive model.
2. Inclusion of additional variables:
   Incorporating academic grades, school facilities, teacher quality, and psychosocial factors to build more comprehensive models.
3. Comparative modeling:
   Evaluating the performance of Decision Trees against more advanced algorithms such as Random Forest, Gradient Boosting, or Neural Networks to determine trade offs between accuracy and interpretability.
4. Practical deployment:
   Developing decision support dashboards for teachers and school administrators that can visualize at risk students in real time and recommend intervention strategies.

By extending the current framework, future studies can contribute to more effective policies and practices for dropout prevention, ultimately supporting educational equity and student success.

References

Abdah Syakiroh Gustian, & Fathoni Mahardika. (2025). Analisis Klasifikasi Risiko Dropout Mahasiswa Menggunakan Algoritma Decision Tree dan Random Forest. *Jupiter: Publikasi Ilmu Keteknikan Industri, Teknik Elektro Dan Informatika*, *3*(4), 182–189. https://doi.org/10.61132/jupiter.v3i4.980

Elbouknify, I., Berrada, I., Mekouar, L., Iraqi, Y., Bergou, E. H., Belhabib, H., Nail, Y., & Wardi, S. (2025). *AI-based identification and support of at-risk students: A case study of the Moroccan education system*. https://doi.org/https://doi.org/10.48550/arXiv.2504.07160

Gottfried, M. A. (2014). Chronic Absenteeism and Its Effects on Students' Academic and Socioemotional Outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *19*(2), 53–75. https://doi.org/10.1080/10824669.2014.962696

Kabra, R. R., & Raisoni, G. H. (2011). Performance Prediction of Engineering Students using Decision Trees. In *International Journal of Computer Applications* (Vol. 36, Issue 11). https://doi.org/10.5120/4532-6414

Kaffenberger, M., Sobol, D., & Spindelman, D. (2021). *The Role of Low Learning in Driving Dropout: A Longitudinal Mixed Methods Study in Four Countries*. https://doi.org/10.35489/BSG-RISE-WP_2021/070

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Ou, S. R., & Reynolds, A. J. (2008). Predictors of Educational Attainment in the Chicago Longitudinal Study. *School Psychology Quarterly*, *23*(2), 199–229. https://doi.org/10.1037/1045-3830.23.2.199

Psyridou, M., Prezja, F., Torppa, M., Lerkkanen, M. K., Poikkeus, A. M., & Vasalampi, K. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-63629-0

Putra, L. G. R., Prasetya, D. D., & Mayadi, M. (2025). Student Dropout Prediction Using Random Forest and XGBoost Method. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, *9*(1), 147–157. https://doi.org/10.29407/intensif.v9i1.21191

Copyright © The Author(s)

Romero, S., & Liao, X. (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. *PLOS ONE*, *20*(6 June). https://doi.org/10.1371/journal.pone.0325047

Ross Quinlan, by J., Kaufmann Publishers, M., & Salzberg, S. L. (1994). *Programs for Machine Learning* (Vol. 16).

Rumberger, R. W. (2001). *Why Students Drop Out of School and What Can be Done*. https://escholarship.org/uc/item/58p2c3wp

Tajriah, S., Goretty Djandon, M., & Sulaiman, H. (2022). *Motif Anak Putus Sekolah Yang Bekerja Pada Sektor Informal (Studi Kasus) di Kelurahan Ekasapta Kecamatan Larantuka Kabupaten Flores Timur*. 58–74. https://doi.org/https://doi.org/10.37478/sajaratun.v7i2

Utomo, A., Reimondos, A., Utomo, I., McDonald, P., & Hull, T. H. (2014). What happens after you drop out? Transition to adulthood among early school-leavers in urban Indonesia. *Demographic Research*, *30*(1), 1189–1218. https://doi.org/10.4054/DemRes.2014.30.41