Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Improving the Classification Model of Smart Indonesia Program Recipients in Koorwilbidikcam Sumber Using the C4.5 Algorithm

Sigit Saputra[1*], Nana Suarna[2], Irfan Ali[3], Dodi Solihudin[4]

[1,2,4] Teknik Informatika, STMIK IKMI CIREBON
[3]Rekayasa Perangkat Lunak, STMIK IKMI CIREBON
sigitsaputra725@gmail.com[1]*

Abstract

This study aims to implement the C4.5 classification algorithm in determining the recipients of the Smart Indonesia Program (PIP) in the Koorwilbidikcam Sumber area. The main problem faced is the inaccuracy in identifying recipients of assistance which causes suboptimal distribution. This study uses a quantitative approach with data mining techniques, analyzing students' social and economic data. The dataset used consists of 550 student data with variables such as parental occupation, income, means of transportation, and ownership of KIP or SKTM. The classification process is carried out using the C4.5 algorithm with the Knowledge Discovery in Databases (KDD) stages which include data selection, preprocessing, data transformation, and data mining using RapidMiner. The results of the study show that the C4.5 algorithm is able to identify significant patterns in the data and produce decision trees that can be used to support decision making. The implementation of this algorithm improves the accuracy and efficiency of the PIP recipient selection process, as well as reducing subjective bias in the determination. Thus, this study contributes to the development of a fairer and more targeted educational assistance distribution system. The use of data mining-based technology such as the C4.5 algorithm also opens up opportunities for technology integration in decision making in the education sector, so that it can be a model for other areas with similar problems. In this study, it produced an accuracy rate of 85.45% from 550 data and from 110 testing data that had been tested.

Keywords: *C4.5 Algorithm, Smart Indonesia Program (PIP), Classification*

1. Introduction

Education plays a very important role in the development of a country. The Smart Indonesia Program (PIP) is one of the government's initiatives to improve access and quality of education, especially at the elementary school level (SD). The Smart Indonesia Program (PIP) is government support in the form of direct payments to students according to predetermined criteria. There are several types of scholarships, one of which is a scholarship provided by the Indonesian government, namely the Smart Indonesia Program Scholarship (PIP) for elementary, junior high and high school levels, where according to (Ratna et al., 2022) the Smart Indonesia Program Scholarship is carried out to minimize the dropout rate as stated in the instruction of the President of the Republic of Indonesia number 7 of 2014 concerning the implementation of the productive family program through the Smart Indonesia Program (PIP) which aims to prevent students from the possibility of dropping out of school.

This study aims to implement the C4.5 classification algorithm in the context of determining PIP recipients in the Koorwilbidikcam Sumber. Using data mining technology, this study will explore patterns in students' social and economic data to improve the accuracy and precision of determining PIP beneficiaries. The research method uses the C4.5 algorithm, which is a widely known decision tree program used in practice. This algorithm uses information entropy and normalized information gain to build a decision tree for classification purposes. The results of this study are not only in the form of technology implementation but also an important step in improving the quality and sustainability of the PIP program at the elementary level. The allocation of PIP funds that are right on target can support the participation of students in need and strengthen the educational base of Koorwilbidikcam Sumber as a whole. The C4.5 algorithm was chosen because of its ability to identify meaningful patterns in data and produce interpretable decision rules.

The C4.5 algorithm is one of the algorithms used to classify data by forming a decision tree. The process in the decision tree is to change the form of data (table) into a tree model, change the tree model into a rule, and simplify the rule.

2. Research Methodology

This research adopts a quantitative approach to analyze the social and economic data of elementary school students in the Koorwilbidikcam Sumber region. The primary objective is to improve the classification accuracy of recipients of the Smart Indonesia Program (Program

Indonesia Pintar, PIP) using the C4.5 algorithm. The study involves several sequential steps: problem analysis, observation, data collection, data analysis, and drawing conclusions.

2.1. Research Design

The research flow begins with identifying problems related to the inaccuracy of PIP recipient selection. Observations were conducted at several elementary schools in the area to better understand the socioeconomic diversity of the student population. Data collection was then carried out using Dapodik (Basic Education Data) consisting of 550 student records across 8 attributes: Name, Address, Type of Residence, Parent's Occupation, Parent's Income, Mode of Transportation, SKTM Ownership (Certificate of Poverty), and KIP Ownership (Smart Indonesia Card).

2.2. Data Sources

The data were obtained through direct interviews with officials from Koorwilbidikcam Sumber. These interviews provided insights into the recipient eligibility criteria, which include economic status, educational background, parental occupation, and household dependents. The discussion also revealed several challenges in recipient selection, such as incomplete data, changing economic conditions, and limited verification resources. Officials expressed optimism that implementing a classification algorithm could improve the fairness and efficiency of the selection process.

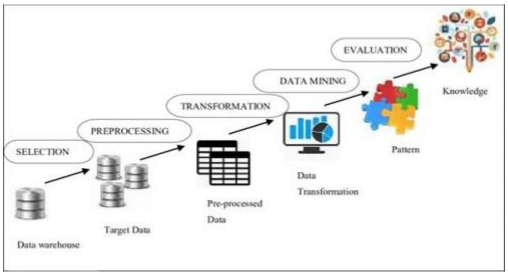
2.3. Population and Sample

The sample consists of 550 student data entries sourced from Dapodik through authorized access. These data points serve as the main dataset for classification modeling. All variables are processed to reflect the eligibility status for PIP support.

2.4. Data Collection Techniques

Data collection included both interviews and extraction of Dapodik datasets. After collection, the data were cleaned, organized based on selected attributes, and processed using the RapidMiner software tool. The aim was to build a decision-support model using the C4.5 classification algorithm.

2.5. Data Analysis Technique



The data analysis process follows the Knowledge Discovery in Databases (KDD) methodology, which involves six primary steps:

1. Data Selection
   Relevant attributes are selected to ensure the analysis only involves scientifically justified variables.
2. Data Preprocessing
   The data are cleaned by removing unnecessary, missing, or inconsistent entries to enhance data quality before classification.
3. **Data Transformation**
   Data are converted into Excel format and imported into RapidMiner for further processing with the C4.5 algorithm.
4. **Data Mining**

This study applied the C4.5 decision tree algorithm to classify students eligible for the Program Indonesia Pintar (PIP). From 550 records, the data was split into 440 training and 110 testing entries. The C4.5 algorithm builds the tree by selecting the most informative attributes based on the Gain Ratio, which is a normalized measure of information gain. The tree construction process includes calculating entropy to measure uncertainty, computing gain ratio for each attribute, selecting the best attribute as the root, and recursively repeating the process for remaining branches. This continues until no significant gain is obtained. The final decision tree consists of interpretable rules to support eligibility classification.

The entropy and information gain formulas used are:

$$Entropy\ (S) = -\sum_{i=1}^{n} pi * log_2 pi$$

$$Gain\ (S, A) = Entropy\ (S) \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy}\ (Si)$$

Explanation:
S: Set of cases

n: Number of partitions *S*
pi: Proportion of *Si* to *S*
A: Attribute
|*Si*|: Number of cases in partition i
|*S*|: Total number of cases in *S*

5. Evaluation

The performance of the model is evaluated using accuracy, recall, and precision metrics. The final model achieved an accuracy rate of 88.18%, confirming the algorithm's effectiveness in identifying eligible PIP recipients.

3. Results and Discussion

This study uses a dataset obtained from direct interviews with one of the officials at Koorwilbidikcam Sumber, resulting in 550 data entries**.** The dataset is presented in Table 3.1 below:

Table 1: Dataset

No	Name	Address	Type of Residence	Parent's Occupation	Parent's Income	Transportation	SKTM Holder	KIP Recipient
1	Amar Hidayat	Pasalakan	With Parents	Entrepreneur	Rp. 2,500,000 – Rp. 3,000,000	Walking	No	No
2	Nadin Ismayanti	Kaliwadas	With Parents	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
3	Lilik Nadiroh	Pasalakan	With Parents	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
4	Indra Saputra	Perbutulan	With Parents	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
5	Dea Almaretha	Sendang	With Parents	Farmer	Rp. 2,000,000 – Rp. 3,000,000	Walking	No	Yes
6	Sandi Permadi	Sendang	With Parents	Farmer	Rp. 2,000,000 – Rp. 3,000,000	Bicycle	No	Yes
7	Hanun Permata	Perbutulan	With Parents	Entrepreneur	Rp. 2,500,000 – Rp. 3,000,000	Walking	Yes	No
8	Indy Sulistiawati	Perbutulan	With Parents	Unemployed	0 – Rp. 300,000	Bicycle	Yes	Yes
9	Asna	Pasalakan	With Parents	Porter	0 – Rp. 500,000	Bicycle	Yes	Yes
10	Hisyam Syamsuri	Pasalakan	With Parents	Vendor	Rp. 1,000,000 – Rp. 2,000,000	Bicycle	Yes	Yes
…	…	…	…	…	…	…	…	…
545	Sinta Nirmala	Sumber	With Parents	Laborer	0 – Rp. 500,000	Bicycle	Yes	Yes
546	Endang Siliarti	Sumber	With Parents	Laborer	0 – Rp. 500,000	Bicycle	Yes	Yes
547	Entin	Sumber	With Parents	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
548	Ratna Asih	Sumber	With Parents	Porter	0 – Rp. 500,000	Walking	Yes	Yes
549	Temu Syarifah	Pasalakan	With Parents	Porter	0 – Rp. 500,000	Walking	Yes	Yes
550	Andi Setiadi	Kaliwadas	With Parents	Police Officer	Rp. 3,500,000 – Rp. 4,500,000	Walking	No	No

3.1. Data Selection

In this stage, attribute selection was performed to identify which data features were relevant for the classification process. Initially, the dataset consisted of eight attributes: Name, Address, Type of Residence, Parent's Occupation, Parent's Income, Transportation, SKTM Holder, and KIP Recipient. After review, three attributes—Name, Address, and Type of Residence—were considered irrelevant to the prediction goal and thus excluded. The remaining five attributes—Parent's Occupation, Parent's Income, Transportation, SKTM Holder, and KIP Recipient—were selected for further processing and model training in RapidMiner.

3.2. Preprocessing

At this stage, data preprocessing is carried out to identify and correct potential errors or inconsistencies within the dataset. The preprocessing process includes handling irrelevant, inaccurate, and missing data. This step ensures that the data used for model training is clean, complete, and ready for further transformation. The final result of the selected and preprocessed data is shown in Table 3.2 below:

Table 2: Preprocessed Data Sample

No	Parent's Occupation	Parent's Income	Transportation	SKTM Holder	KIP Recipient
1	Entrepreneur	Rp. 2,500,000 – Rp. 3,000,000	Walking	No	No
2	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
3	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
4	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
5	Farmer	Rp. 2,000,000 – Rp. 3,000,000	Walking	No	Yes
6	Farmer	Rp. 2,000,000 – Rp. 3,000,000	Bicycle	No	Yes
7	Entrepreneur	Rp. 2,500,000 – Rp. 3,000,000	Walking	Yes	No
8	Unemployed	0 – Rp. 300,000	Bicycle	Yes	Yes
9	Porter	0 – Rp. 500,000	Bicycle	Yes	Yes
10	Vendor	Rp. 1,000,000 – Rp. 2,000,000	Bicycle	Yes	Yes

…	…	…	…	…	…
545	Laborer	0 – Rp. 500,000	Bicycle	Yes	Yes
546	Laborer	0 – Rp. 500,000	Bicycle	Yes	Yes
547	Laborer	0 – Rp. 500,000	Walking	Yes	Yes
548	Porter	0 – Rp. 500,000	Walking	Yes	Yes
549	Porter	0 – Rp. 500,000	Walking	Yes	Yes
550	Police Officer	Rp. 3,500,000 – Rp. 4,500,000	Walking	No	No

3.3. Data Transformation

Data transformation is conducted to convert and structure the dataset from Excel format into a form suitable for processing within the RapidMiner environment. This includes assigning roles to each attribute to ensure proper operation flow during training. As shown in Figure 4.1**,** the Set Role operator was used to assign the target label:

- Attribute: “KIP Recipient”
- Target Role: “Label”

3.4. Data Mining

The next step after the data transformation process is the implementation of the C4.5 Decision Tree algorithm using RapidMiner. This algorithm is used to build a decision tree model that can be utilized for classification or prediction based on the given data. The stages involved in the data mining process are as follows:

1. Importing Data
   The first step involves importing the dataset in Excel format into the RapidMiner environment using the Read Excel operator. The attributes used in this study include Parent’s Occupation, Parent’s Income, Means of Transportation, SKTM Holder, and KIP Recipient (as the label)
2. Setting the Target Attribute (Label)
   The Set Role operator is used to define how each attribute will be treated by other operators in the model. In this case, the attribute "KIP Recipient" is assigned the role of "label", which serves as the target for classification. This allows the algorithm to learn from the input attributes and predict whether a student is eligible for the Smart Indonesia Program.
3. Splitting the Dataset
   The dataset is then divided into two subsets using the Split Data operator:
   a. Training Data (80%)
   b. Testing Data (20%)
   This division is essential to train the model on a substantial portion of the data while reserving a separate subset for performance evaluation. The commonly used 80:20 ratio ensures that the model is trained effectively and validated on unseen data.
4. Training and Testing the Data
   The dataset consisting of 550 records was divided into 440 training data and 110 testing data using the Split Data operator. The training phase involved feeding five selected attributes—parent’s occupation, parent’s income, transportation, SKTM holder, and KIP recipient—into the C4.5 Decision Tree algorithm to build a classification model. The model parameters used in the Decision Tree operator included: Gain Ratio as the criterion, a maximum depth of 10, and pre- and post-pruning enabled, with minimum gain and leaf size thresholds to avoid overfitting.
   Once the model was trained, it was applied to the testing data using the Apply Model operator. Model performance was evaluated using the Performance (Classification) operator, with metrics such as accuracy, weighted mean recall, and weighted mean precision. This process ensured that the model could be validated against unseen data and confirmed its ability to generalize classification patterns accurately. In this phase, the training data is used to train the C4.5 algorithm. During this process, the model learns patterns from the provided features (parent’s occupation, income, transportation, SKTM status) in relation to the label (KIP recipient). The resulting model is then evaluated using the test data to assess its predictive performance.
5. Decision Tree Output
   After computing and evaluating the dataset using the C4.5 algorithm, a complete decision tree model was generated. This tree illustrates the classification paths derived from key attributes such as parent’s income, occupation, transportation, and SKTM ownership. The visual structure of the tree allows for a clear interpretation of how eligibility decisions are made based on specific attribute combinations.
6. Decision Tree Description
   The resulting tree structure provides a comprehensive overview of the classification logic. Each node and branch represents a conditional rule, making the model highly interpretable. The results also reinforce the effectiveness of the C4.5 algorithm in producing reliable and understandable decision-making models. The use of C4.5 is recommended in data mining tasks where explainability and structured output are critical.
7. Accuracy Rate
   The model was evaluated using the testing dataset within RapidMiner. The classification accuracy achieved was 85.45%, which indicates a high level of reliability. This means that the generated rules correctly predicted student eligibility in over 85% of test cases. The evaluation process used the Split Data technique to ensure that the testing data remained unseen during training, thereby validating the model’s ability to generalize to new inputs.

3.5. Discussion

The results of this study confirm that the C4.5 algorithm is highly effective for classifying potential recipients of the Smart Indonesia Program (PIP). The decision tree produced by the model provided clear and interpretable rules based on key socio-economic attributes, such as parental income, occupation, and SKTM ownership. The accuracy rate of 85.45%, achieved using RapidMiner, indicates that the model's decision rules are highly reliable and closely aligned with real-world eligibility outcomes. This finding is consistent with previous studies. For instance, Eko Budiarto et al. demonstrated that the C4.5 algorithm outperformed Naive Bayes in peer classification tasks. ADA Abdurrahman also highlighted that C4.5 improves upon ID3 by supporting numeric attributes, pruning capabilities, and generating rule sets. Similarly, Weni Ratna Sari Oktapia Ningse et al. achieved 98% accuracy using C4.5 for identifying PIP recipients at MIS Al-Koirot.

Compared to these prior studies, this research used a more updated and targeted dataset obtained from Dapodik 2024, focusing specifically on students within the Koorwilbidikcam Sumber region. While previous research involved students from Islamic junior high schools or pesantren, this study applied the algorithm to a broader demographic and used different sources and data formats. Additionally, this study exclusively focused on C4.5, while others also used comparative algorithms. The overall process included data cleaning, transformation, training, and testing within the RapidMiner environment. The C4.5 model proved robust in identifying eligible students, classifying them into "Yes" or "No" categories. Given its high accuracy and interpretability, this algorithm is highly recommended as a decision support tool for education stakeholders in selecting PIP beneficiaries effectively and transparently.

4. Conclusion

This study explores the application of the C4.5 decision tree algorithm in classifying students eligible for the Smart Indonesia Program (PIP) in the Koorwilbidikcam Sumber region. The findings indicate that C4.5 is highly effective in processing socio-economic data and generating accurate classification models. With an accuracy rate of 85.45%, the model successfully identified eligibility patterns using critical attributes such as parental income, occupation, transportation, and SKTM ownership.

The attribute selection process within the C4.5 framework helped highlight the most influential factors, offering a clear understanding of what contributes to student eligibility. Furthermore, the interpretability of the decision tree makes it a practical decision-support tool for educational institutions and government stakeholders. Overall, this research confirms that the C4.5 algorithm is both reliable and efficient in supporting fair and data-driven beneficiary selection processes for social assistance programs like PIP.

To improve future implementations, it is recommended to combine C4.5 with other machine learning algorithms to enhance classification performance. Additionally, incorporating more detailed student data—such as number of family dependents, academic performance, extracurricular participation, and achievement records—could improve prediction accuracy. Future studies may also focus on hyperparameter optimization to further refine model outcomes and ensure better generalization across different educational contexts.

References

[1] A. Abdurrahman, "Komparasi Algoritma Naïve Bayes dan C4.5 dalam Menentukan Kelayakan Bantuan KIP (Kartu Indonesia Pintar) (Studi Kasus: DKI Jakarta)," 2020. [Online]. Available: https://repository.mercubuana.ac.id/id/eprint/68157

[2] P. Beras, R. Oleh, J. P. Gultom, and A. Rikki, "Implementasi Data Mining menggunakan Algoritma C4.5 pada Data Masyarakat Kecamatan Garoga untuk Menentukan Pola Penerima Beras Raskin," ARTICL, vol. 2, no. 1, pp. 11–19, 2020.

[3] E. Budiarto, Rino, S. Hariyanto, and D. Susilawati, "Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SD Maria Mediatrix Menggunakan Algoritma C4.5," Algor, vol. 3, no. 2, 2022. doi: 10.31253/algor.v3i2.1019

[4] P. P. Haryoto, H. Okprana, and ..., "Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru," TIN: Terapan Informatika, 2021. [Online]. Available: http://ejurnal.seminar-id.com/index.php/tin/article/view/919

[5] N. Hidayah, "Sistem Klasifikasi Penerima Beras Miskin Menggunakan Algoritma Decision Tree C4.5 (Studi Kasus: Kelurahan Tambakmerang, Girimarto)," eprints.uty.ac.id, 2019. [Online]. Available: http://eprints.uty.ac.id/4136/

[6] I. Zamjani, Herlinawati, N. S. Perdana, F. Widiaputera, and S. N. Azizah, "Biaya Satuan & Lini Masa Pengelolaan Program Indonesia Pintar," 2020.

[7] W. R. S. Oktapia Ningse, S. Sumarno, and Z. M. Nasution, "Klasifikasi Algoritma C4.5 untuk Penentuan Penerima Program Indonesia Pintar pada MIS Al-Khoirot," [Online]. Available: http://download.garuda.kemdikbud.go.id/article.php?article=2998597&val=27036

[8] N. Aprilyani, I. Zulfa, and H. Sulaiman, "Penerapan Algoritma Decision Tree C4.5 Untuk Model Penentuan Penerima Beasiswa Program Indonesia Pintar (PIP): Studi Kasus SMA Negeri 3 Timang Gajah," J. Tek. Inform. Dan Elektro, 2023.

[9] N. Nurahman, "Klasifikasi Penerima Bantuan Sosial di Desa Batuah Menggunakan Metode Algoritma C4.5," J. Tekinkom, vol. 1, no. 1, 2022. [Online]. Available: http://jurnal.murnisadar.ac.id/index.php/Tekinkom/article/view/516

[10] E. Prasetyaningrum and P. Susanti, "Analisa Tingkat Kepuasan Pelanggan Pada Percetakan CV. Mega Media Menggunakan Algoritma C4.5," Sisfotenika, vol. 13, no. 1, pp. 65–75, 2023.

[11] N. W. O. Pratiwi, N. W. Utami, and I. Putra, "Klasifikasi Penentuan Penerima Bantuan Sosial Tunai (BST) Menggunakan Algoritma C4.5 di Desa Keramas, Gianyar, Bali," J. Inform. Teknol., 2022. [Online]. Available: http://www.jurnal.uts.ac.id/index.php/JINTEKS/article/view/1667

[12] R. A. Saputra, S. Wasiyanti, and D. Pribadi, "Information Gain Pada Algoritma C4.5 Untuk Klasifikasi Penerimaan Bantuan Pangan Non Tunai (BPNT)," Indones. J. Business Intell. (IJUBI), vol. 4, no. 1, pp. 25–33, 2021. doi: 10.21927/ijubi.v4i1.1757

[13] Secretary General of the Ministry of Education, Culture, Research, and Technology, "Peraturan Sekretaris Jenderal Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 14 Tahun 2022 tentang Petunjuk Pelaksanaan Program Indonesia Pintar Pendidikan Dasar dan Menengah," JDIHN, 2022. [Online]. Available: https://jdih.kemdikbud.go.id/

[14] W. Susanto and A. Mulyani, "Analisa Algoritma C4.5 Terhadap Penentuan Rekomendasi," OKTAL: J. Ilmu Komputer dan Sains, vol. 1, no. 10, pp. 1607–1616, 2022.