**NEWTON: Networking and Information Technology**
Vol. 2 No. 2 October 2022, Page. 72 - 78
E-ISSN: 2797-0728

# Classification of Prospective Students with Rapid Miner

**Agus Sifaunajah[1], Riesca Dewi Wahyuningtyas[2]***
[1,2] Information Systems, KH. Abdul Wahab Hasbullah University
*Email: dewiriesca@gmail.com

## ABSTRACT
*The application of information systems in various daily life processes results in data accumulation. Classification is a process in data mining to perform data analysis to generate models to describe the classes contained in the data. In this study, the ID3 Algorithm was used to classify prospective student data. This research employed a qualitative research model using the rapidminer tool application to carry out the classification process. Processed data is data obtained from surveys through google form media. The results of this study showed a classification model using the Interactive Dichotomizer Algorithm (id3), which can be developed into an information system for the admission of prospective students. It also revealed that data classification of students "cannot" or "can" at the foundation of the boarding school Bahrul Ulum Jombang in the form of a decision tree. Based on the decision tree, it was found that 50.72% (35 people) were classified as poor/incapable, and 49.27% (34 people) were classified as capable from a total of 69 data on prospective students.*

***Keywords***— *Information systems, classification, qualitative, Algorithm id3.*

## INTRODUCTION
There is a buildup of data when the application of information systems search for student data. An information system is a set of interrelated (integrated) elements that collect, store, process, and disseminate information to support decision-making and other purposes, both people and organizations that use classification in data mining (Yaqin *et al.*, 2021).

There are several functionalities of data mining, including analysis of associations between data, data classification, data clustering, and others. In this study, the functionality was data classification. Classification is a data analysis process that produces models to describe the classes contained in the data (Han, Kamber, & Pei, 2012). These models are called classifiers. This classifier is used to compile the types collected in the data.

The decisions that can be taken are greatly influenced by some time. One of the supporters needed is classifying student data within the Bahrul Ulum Islamic boarding school foundation Jombang. The ID3 Algorithm is one of the data mining algorithms that will be implemented to classify students to get the information needed using qualitative research methods with the Rapidminer application to process data and check student data at the boarding school foundation.

## METHOD
- ***Qualitative Research Method***

  **Qualitative research** is descriptive research that tends to use analysis. Process and meaning (subject perspective) are more highlighted in qualitative research. The theoretical basis is used as a guide so that the research focus follows the facts on the ground. Qualitative research has been known since the 1960s and is often called the alternative method. This method does not use detailed questions but starts with general ones that are tapered and detailed. Qualitative methods treat participants as subjects, not objects, so participants consider themselves valuable because their information is beneficial.
- ***Data Retrieval***

  With the accumulation of data in this study, researchers applied a survey system that looked for data by utilizing Google Forms to make it easier to search for data that would be applied using the RapidMiner tool application.

- ***Algorithm Id3 Workflow***

  Testing using the Id3 Algorithm is an algorithm to build a decision tree. This Algorithm was discovered by J. Ross Quinlan (1979) by utilizing Shanon's Information Theory. Id3 itself stands for Iterative Dichotomiser 3. The decision tree uses a hierarchical structure for supervised learning. The decision tree process recursively starts from the root node to the leaf node. Each branch states a condition that must be met, and at each end of the tree says the class of data. The decision tree process is changing the data's shape (table) into a tree model and then changing the tree model into a rule or the resulting rule.

  - **Entropy**

    Entropy is a formula for calculating the homogeneity of a sample

    Entropy(S) $\sum_{i=1}^{c} -pilog2p1$

    $Pi = \frac{Zi}{N} li$

    $Zi = positive\ example + Negative\ example$

    N = Amount of data

    **Figure 1.** *Entropy Formula*

    Information :

    a. Entropy(S) = 0 if all examples of S are in the same class.
    b. Entropy(S) = 1 if the number of positive and negative samples in S is the same.
    *c.* 0 < Entropy(S) < 1, if the number of positive models and the number of negative examples in S is not the same.

  - **Gain**

    Gain(S, A) is the Information Gain of an attribute A in example S :

$$Gain(A) = Entropy(S) - \sum_{i=1}^{k} \frac{|Si|}{|S|} \times Entropy(Si)$$

**Figure 2.** *Gain Formula*

  - **Decision Tree**

    The decision tree is a tree that is used as a reasoning procedure to get answers to the problems entered. The tree that is formed is not always a binary tree. If all the features in the data set to use two kinds of categorical values, then the obtained tree form is in the form of a binary tree. If the quality contains more than two kinds of absolute values or uses a numeric type, the obtained tree form is usually not a binary tree.

    Flexibility makes this method attractive, in particular, because it provides the visualization of suggestions (in the form of a decision tree) that makes the prediction procedure observable (Gorunescu, 2011). An example of a decision tree is shown in the following figure.

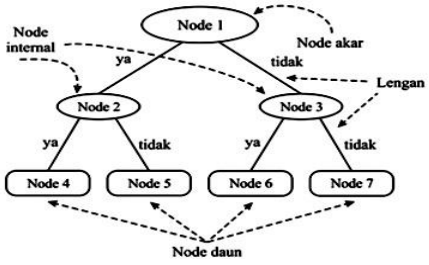


**Figure 3.** Sample Of Decision Tree

The characteristics of the decision tree, as shown in Figure 3, are formed by several elements as follows (Tan et al., 2006): The root node has no input arm and has zero or more output arms. An internal node is any non-leaf (non-terminal) node with precisely one input arm

and two or more output arms. This node represents tests based on feature values. Arm, each branch represents the value of the test results in the non-leaf nodes. Leaf node (terminal) is a node that has exactly one input arm and no output arm. This node represents the class label (decision).

- **Data Characteristic**

  The data characteristics are data types, accuracy, and precision, age, period, completeness, level of conciseness, relevance, and ease of access to data sources.

  - Class Target
    The target class is an attribute often used as a target class or called a label (target attribute). As for the desired target class, the target class is taken through Bps from 2 data, namely:
    - Capable
    - Incapable/Poor
  - The requirements of the class target are :
    - The capable family has the following categories:
      - Father's salary above Rp. 1,000,000 - 5,000,000 per month
      - Mother's salary above Rp. 500,000 - Rp.300.000 per month
      - Mother and Father's Job
      - Does not have a sibling or has a sibling but only 1
      - Have a vehicle with cash prices
      - The sibling has worked or currently pursuing higher education
    - Incapable/Poor Family has the following categories:
      - Father's salary above Rp. 500,000 - Rp. 1,000,000
      - Mother's salary above Rp.0 - Rp. 500,000
      - Mother and Father's Job
      - Have more than three siblings
      - Have a motorbike only by making payments on credit
      - Some are still working or are still studying
  - Attributes of Variables
    - Variables from work consist of father's job category, mother's job
    - The variable of salary consists of the father's salary and the mother's salary
    - Variable number of siblings consists of how many siblings, such as one, two, three, or four siblings
  - Data Identification
    To carry out identification, an inspection regarding the quality of checking the completeness of the data is carried out. Identification includes missing or blank values in the data.

- **Data Preparation**

  The data at this stage includes all the data to be entered into the modeling tool, and the data is processed from the beginning. The data is processed in ms. Excel by setting (parameters first) the target class or variable attribute and studying the data so that it is hoped that the data will have more knowledge of the processed data. Then it can help in selecting the data for the mining process. In carrying out data preparation, several stages need to be done. The stages include:

  - Data Sampling
    Data sampling is data that will be used for the following data mining process. Managed data in ms. excel is.
    - Father's job, Mother's job
    - Father's salary, mother's salary
    - Number of siblings
    - Eligibility = (Capable, Incapable/Poor)

- Data
  The raw data from students' data is then selected for the parameters to be analyzed. The taken parameters are variables or attributes previously made in the target class "capable" or "Incapable/poor." These attributes will become parameters explaining that the sample is the ability (capable or poor) of students at the boarding school foundation bahrul ulum jombang.
- Data Transformation
  After the data is selected and defined, then transform the information on specific parameters. To facilitate the process of changing the Entropy and Gain, calculations were done to make a decision tree.
- Cleaning Data
  This cleaning stage is carried out for the process of cleaning the data so that the data is ready for the modeling stage. This activity cleans up the missing data.

- **Modeling**
  At this stage, the data mining processing process using the id3 algorithm flow is carried out concerning the desired data form from the application of data mining, namely knowing the decision tree on the students' data of the Bahrul Ulum Jombang Islamic Boarding School Foundation with the id3 algorithmic flow. It is intended to measure RESULT AND accuracy, which uses the rapidminer application. The following is an overview of the research process.

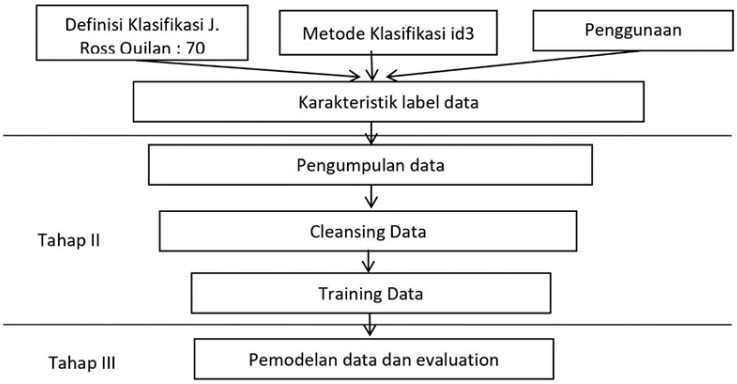


**Figure 4.** Research instrument flow chart

- **Evaluation Stage**
  The evaluation stage is the stage for conducting interpretation trials of the data mining results that have been generated in the previous step.

## DISCUSSION

- ***Data Set ( Primary Data )***
  The raw data received from the students of the Bahrul Ulum Islamic Boarding School Foundation, Jombang, were then selected for the parameters to be analyzed. The taken parameter is the target class previously created with the target class "capable" and "poor." The attribute will be a parameter or input variable. The data on the Bahrul Ulum Islamic Boarding School Foundation Jombang explained that the characteristics of "capable" and "poor" meant that the students of this Islamic Boarding School foundation were able to experience education with the help of the Bahrul Ulum Islamic Boarding School Jombang Foundation. The following figure is the processed data on rapidminer.

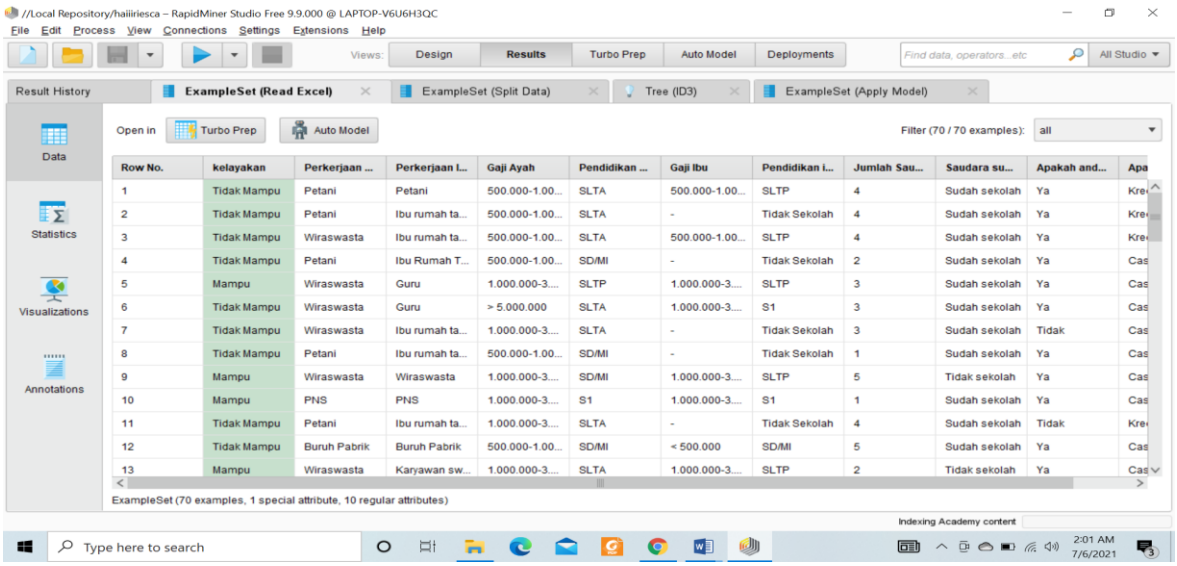


**Figure 5.** Processed data on RapidMiner

- **Processing Data**

In this process, the data cleaning software rapidminer performs data processing starting from reading the primary data feasibility of the students of the Bahrul Ulum Jombang Islamic Boarding School foundation to separate or filter empty data to make data processing at the next stage run well. Cleaning data is the correct data to be used in the following data.

**Table 1.** Processed Data in Excel

| Attribute | Category | Amount | Capable | Poor | Entropy | Gain |
|---|---|---|---|---|---|---|
| Total | | 345 | 170 | 175 | 0.999848483 | |
| | | | | | | |
| Father's Salary | | | | | | 0.222833527 |
| | < 500.000 | 25 | 0 | 25 | 0 | |
| | 500.000-1.000.000 | 105 | 25 | 80 | 0.791858353 | |
| | 1.000.000-3.000.000 | 130 | 95 | 35 | 0.840358672 | |
| | > 5.000.000 | 85 | 10 | 35 | 0.890334183 | |
| Mother's Salary | | | | | | 0.216815385 |
| | - | 37 | 13 | 24 | 0.93526914 | |
| | < 500.000 | 161 | 55 | 106 | 0.926355683 | |
| | 500.000-1.000.000 | 64 | 24 | 40 | 0.954434003 | |
| | 1.000.000-3.000.000 | 64 | 59 | 5 | 0.395537806 | |
| | > 5.000.000 | 19 | 19 | 0 | 0 | |
| Father's job | | | | | | 0.069777051 |
| | Entrepreneur | 145 | 75 | 70 | 0.999142104 | |
| | Farmer | 125 | 45 | 80 | 0.942683189 | |
| | Teacher | 15 | 5 | 10 | 0.918295834 | |
| | Civil Servant | 45 | 35 | 10 | 0.764204507 | |
| | Factory Workers | 10 | 5 | 5 | 1 | |
| Mother's Job | | | | | | 0.168671041 |
| | Entrepreneur | 50 | 35 | 15 | 0.881290899 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Farmer | 60 | 15 | 45 | 0.811278124 | |
| | Teacher | 30 | 15 | 15 | 1 | |
| | Civil Servant | 15 | 15 | 0 | 0 | |
| | Factory Workers | 5 | 0 | 5 | 0 | |
| | Housewife | 165 | 75 | 90 | 0.994030211 | |
| Father's Education | | | | | | 0.02215492 |
| | Elementary School | 100 | 40 | 60 | 0.970950594 | |
| | Junior High School | 60 | 30 | 30 | 1 | |
| | Senior High School | 150 | 75 | 75 | 1 | |
| | Undergraduate Degree | 35 | 25 | 10 | 0.863120569 | |
| Mother's Education | | | | | | 0.03997623 |
| | No school level | 12 | 4 | 8 | 0.918295834 | |
| | Elementary School | 86 | 40 | 46 | 0.99648599 | |
| | Junior High School | 70 | 30 | 40 | 0.985228136 | |
| | Senior High School | 122 | 56 | 66 | 0.995148096 | |
| | Undergraduate Degree | 50 | 35 | 15 | 0.881290899 | |
| Vehicle | | | | | | 0.002156212 |
| | Cash | 245 | 125 | 120 | 0.999699543 | |
| | Kredit | 100 | 45 | 55 | 0.992774454 | |
| Sibling | | | | | | 0.173041197 |
| | 1 | 65 | 50 | 15 | 0.779349837 | |
| | 2 | 70 | 45 | 25 | 0.940285959 | |
| | 3 | 100 | 50 | 50 | 1 | |
| | 4 | 55 | 5 | 50 | 0.439496987 | |
| | 5 | 45 | 20 | 25 | 0.99107606 | |
| | 6 | 10 | 0 | 10 | 0 | |

- **Modelling**

In this picture, the id3 algorithm data is the stage of making id3 algorithm training data by modeling through producing a decision tree about the students' data of the bahrul ulum jombang Islamic boarding school foundation. Based on the results of the id3 algorithm data, the result can be obtained.

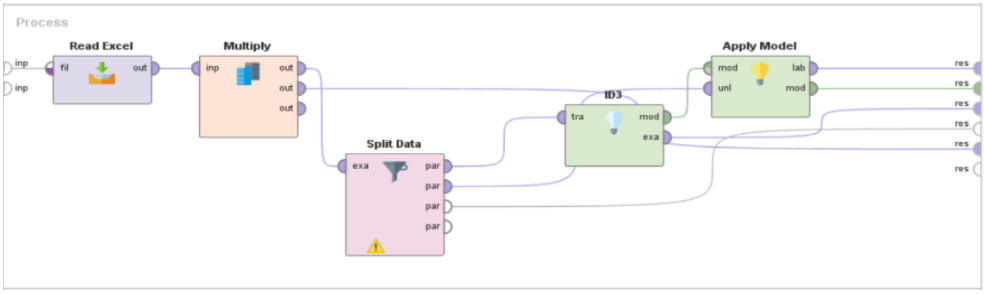


**Figure 6.** The Result of Algoritma id3 Data

- **Decision Tree**

In this step, we have calculated the gain value. We choose the maximum gain value between each attribute from the gain value. The maximum gain value is Gain (Father's salary).

Figure 7 showed the formation of a decision tree, namely the father's salary attribute. The formed branch contained categorical data on the father's attributes, namely father's salary <500,000, father's salary of 500,000 - 10000000, father's salary of 1,000,000 - 3,000,000, father's salary> 5,000,000.

Figure 8 explained the formation of a decision tree, namely the mother's salary attribute. The formed branch was category data on the salary attribute, namely the mother's salary <500,000, the mother's salary 500,000 - 1000.000, the mother's salary 1,000,000 - 3,000,000, and the mother's salary> 5,000,000.

Figure 9 revealed the formation of a decision tree, namely the sister attribute. The formed branch contained categorical data on siblings' characteristics: brothers/Sisters 1,2,3,4,5, and 6.

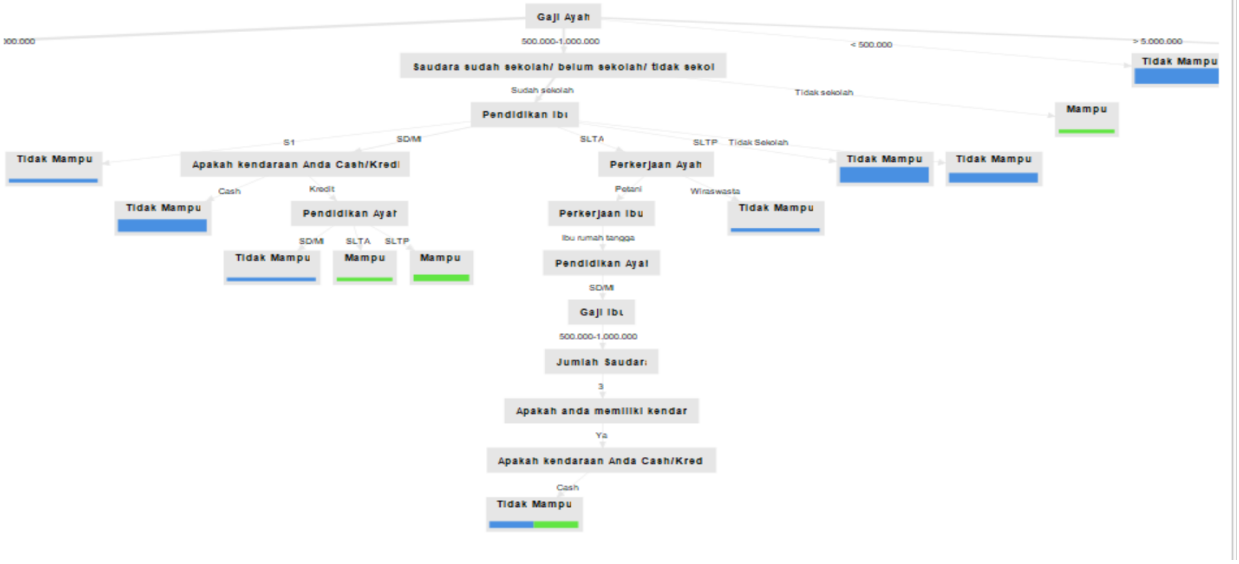


**Figure 7.** Decision tree

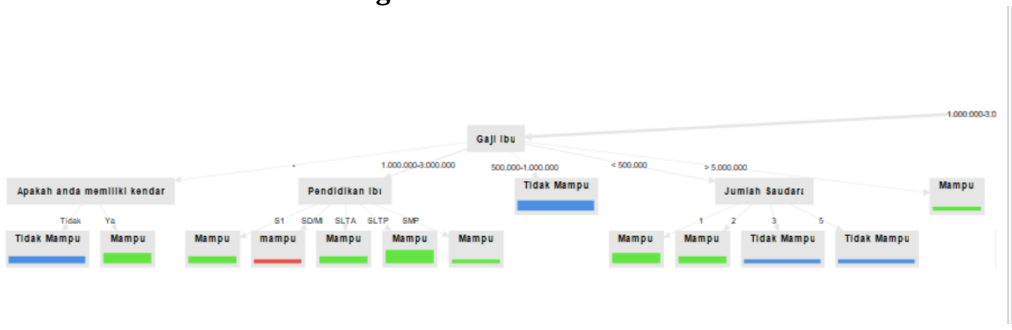


**Figure 8.** Decision tree


**Figure 9.** Decision tree of Sibling

- **Rule**

  Association Rules or Association Analysis is a methodology for finding memorable/interesting relationships (associations) hidden in large data sets (or data sets). One application of the Association rules method is the Market Basket Analysis.

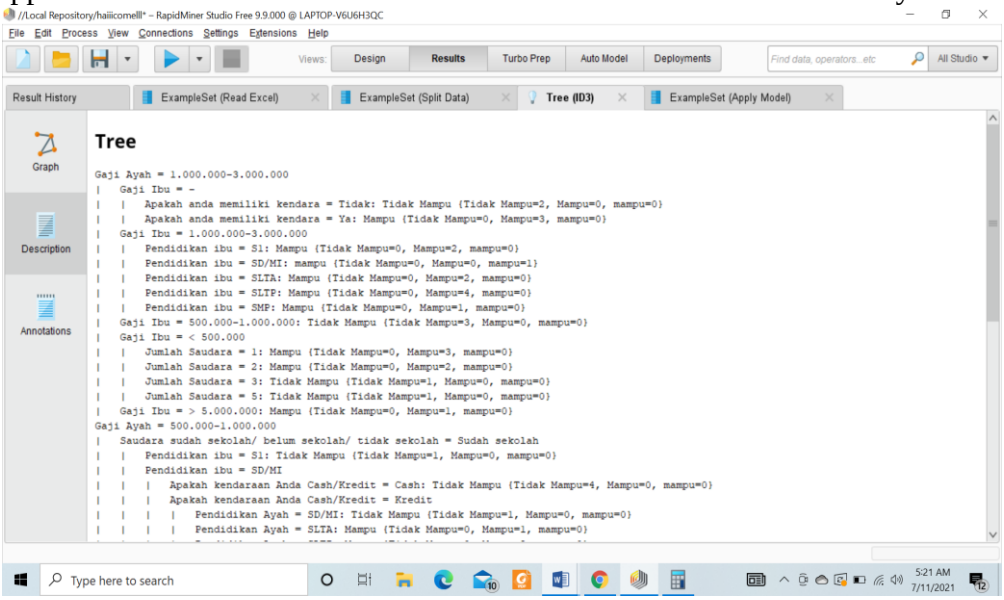


**Figure 10.** Processed Data on RapidMiner

## DISCUSION

The result of this discussion is the data mining process will be faced with the process of collecting data very quickly and large. Data mining is a specific data collection process to identify valuable data. Data mining is used to make decisions based on huge data sets. The current process will be faced with collecting high-speed and large data using the id3 method in the repidminer tool. The source of this huge data flow varies. It can come from data obtained through the Google Form process, distributed to students of the Islamic Boarding School Foundation to obtain valid data, or directly from the Bahrul Ulum Jombang Islamic Boarding School foundation.

It's not easy to store large amounts of data that there are massive amounts. The rapidminer tool application can make it easier to do calculations. This system is responsible for keeping and calculating big data quickly. Data mining is beneficial for users both at the corporate and individual levels. Data mining is looking for patterns that cannot be found using simple analytical techniques. It is necessary to use complex mathematical algorithms to study the data and then evaluate possible future events based on the data.

## CONCLUSION

The results of this study were conducted to build a model using the Interactive Dichotomizer(id3) Algorithm. Data on poor students at the Bahrul Ulum Jombang Islamic Boarding School foundation could be collected by producing a decision tree.

A mother's salary below > 500,000 and a father's salary of 1000.000-300.000 produce more than three children, then it is declared incapable, while a mother's salary is above 1000.000-3000000 and a father's salary is > 500.000 having 1, 2 or 3 children is still considered capable.

The data processed using rapidminer with a percentage of 50.72% stated that 35 people were incapable, and a percentage of 49.27% said that 34 people were capable from a total of 69 student data.

## REFERENCES

Munthe, I. R., & Sihombing, V. (2018). Klasifikasi Algoritma Iterative Dichotomizer (ID3) untuk Tingkat kepuasan pada Sarana Laboratorium Komputer. *Jurnal Teknologi Dan Ilmu Komputer Prima (JUTIKOMP)*, *1*(2), 27–34. https://doi.org/10.34012/jutikomp.v1i2.237

Safii, M. (2019). Implementasi Data Mining Dengan Metode Pohon Keputusan Algoritma Id3 Untuk Menentukan Status Mahasiswa. *Jurnal Mantik Penusa*, *2*(1), 82–86.

Agustina, D. melina, & Wijanarto. (2016). Analisis Perbandingan Algoritma ID3 Dan C4 . 5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum pada PDAM Kabupaten Kendal. *Journal of Applied Intelligent System*, *1*(3), 234–244.

Srimenganti, I., Taufik, I., & Mulyana, E. (2018). Implementasi Algoritma Decision Tree ( ID3 ) Untuk Penyakit Campak. *Seminar Nasional Teknik Elektro*, 235–242.

Amalia, A. E., & Naf'an, M. Z. (2017). Implementasi Algoritma ID3 Untuk Klasifikasi Performansi Mahasiswa (Studi Kasus ST3 Telkom Purwokerto). *Seminar Nasional Teknologi Informasi Dan Multimedia 2017*, 115–120.

Widiyati, D. K., Wati, M., & Pakpahan, H. S. (2018). Penerapan Algoritma ID3 Decision Tree Pada Penentuan Penerima Program Bantuan Pemerintah Daerah di Kabupaten Kutai Kartanegara. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, *2*(2), 125. https://doi.org/10.30872/jurti.v2i2.1864

Sitepu, A. (2012). Karakteristik Keluarga Menurut Peringkat Kemiskinan: Studi Pendahuluan untuk Perumusan Kriteria Fakir Miskin. *Informasi*, *17*(01), 48–63. https://ejournal.kemsos.go.id/index.php/Sosioinforma/article/viewFile/930/490

Yaqin, N., Hariono, T., & Rohman, R. U. (2021, December). Automatic Water Level Control Tem On Hydroponic Plants Based On Arduino. In *Multidiscipline International Conference* (Vol. 1, No. 1, pp. 612-617).