Using C4.5 algorithm to predict students monthly payment on islamic boarding school

S R Istiana, I Waspada
Department of Informatic, Faculty of Science and Mathematics, Diponegoro University Semarang, Indonesia
E-mail: selviratnaistiana@if.undip.ac.id

Abstract. Assalafi AlFithrah is islamic boarding school located in Meteseh Semarang. In 2012 until 2016, it received 262 students from various regions in Indonesia. Fulfillment of life necessities of students is one of the things that should be considered by boarding school management. The basic necessities of life is food, clothing, place, education and health care. The monthly payment needed to cover the necessities of students in Assalafi AlFithrah. There are number of students whos payment were delayed. From these problems, it was necessary to predict the delay of monthly payment to estimate whether students will be late to pay or not. Data mining as solution to predict the delay of the monthly payment on students. One of a robust classification technique of data mining on categorical data is C4.5. Experiment conducted based on Knowledge Discovery Database (KDD) with C4.5 algorithm resulted average accuracy 81.15%, with average value of precision 77.62% and average value of recall 91.90%. This Application can give recommendations for scholarship to poor families. With this application can improve service and trust of parents towards the survival of students in schools.

1. Introduction
Assalafi AlFithrah is islamic boarding school located in Meteseh Semarang. In 2012 until 2016, it received 262 students from various regions in Indonesia. Fulfillment of life necessities of students is one of the things that should be considered by boarding school management. The basic necessities of life are food, clothing, place, education and health care [1]. The monthly payment needed to cover the necessities of students in Assalafi AlFithrah. There are number of students whos payment were delayed. From these problems, it was necessary to predict the delay of monthly payment to estimate whether students will be late to pay or not, so that the boarding school can give special attention to students who are late in paying monthly payment.

In technological development there are many studies that use predictions to overcome problems that will occur in the future. Classification of data is one of the important tasks in data mining [2]. Decision tree is one of the most widely used technique [3], and is very popular among researchers [3] because of their simplicity [4], intelligibility [5], ease of implementation [5], decision tree construction faster and produce better accuracy [6] then other classification algorithms. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class [6].

One of a robust classification technique of data mining on categorical data is C4.5. There have been several previous studies using the C4.5 algorithm. Sri Wahyuni et al (2017) to know the factors of
students who experienced dropout with C4.5 algorithm [7]. Then Mashael Al-Barak dan Muna Al-Razgan implement C4.5 algorithm to predict students' final GPA based on their grades in previous courses and collected transcripts data of female students who graduated from Computer Sciences in King Saud University in the year 2012 [8]. Kalpesh Adhatrao et al (2013) implement C4.5 and ID3 algorithm to predict the student performance and collected first year student data of Fr. C.R.I.T., Navi Mumbai [9]. The advantage of C4.5 algorithm is that the performance of C4.5 algorithm is better when compared to the algorithm ID3, C5.0 and CART [10]. Then, C4.5 algorithm is the fastest performing algorithm and has the highest accuracy when applied to loan data [11].

Based on the description described above, then in this research the writer build the Student's Monthly Payment Prediction Application by Using C4.5 Algorithm. It is to facilitate Assalafi Al Fithrah Meteseh Semarang to know the late monthly payment of student early.

2. Methodology
The methodology used in this research is as follows:

2.1. Domain Understanding and KDD Goals
Assalafi AlFithrah Meteseh Semarang received 262 students from 2012 to 2016. From interviews conducted with the boarding school treasurer stated that delays in monthly payments occur if students do not pay for 3 months or more in a row, where at this time the boarding school can give warnings to students who are late in making monthly payments. Delay in monthly payments will affect the welfare of students in the boarding school, so this can interfere with the learning process of students. So from that it is necessary to predict the delay in monthly payment of students, so that students who are expected to make delays can be given early treatment.

2.2. Selection and Addition
The data that will be used for prediction is the monthly payment data of students at the Assalafi Islamic Boarding School Al Fithrah Meteseh Semarang from 2012 to 2016. The data totaled 262 data. Monthly payment data at the boarding school is recorded in the monthly payment notebook. Data collection of student's monthly payments has not been saved on the computer, and only a portion of the student biodata has been stored on the computer. From this data, it is necessary to transfer data from notebook data into computer. Monthly payment data consists of payment date, name, address, monthly payment details. Biodata recorded on the computer include the name, address, gender, place and date of birth, the name of the student guardian, class (if the students are also at the school), students 'rooms, students' information (active / alumni). From the data obtained, there are data that have not been filled completely including data on the names of student guardians, students 'rooms and students' information (active / alumni). The completed data will be used such as data the name of the student, gender, city, parents' occupation, age, and monthly payment information data. From 262 student data there are 141 data with monthly "Rutin" and 121 data payment information with monthly payment information "Terlambat".

The basis for selecting the attributes of age, city, parents occupation, and gender are as follows:
• The age of the student is used as one of the attributes in the prediction of monthly payments. Age can be used to identify the level of delinquency in children. Someone has gone beyond childhood, but is still not mature enough to be said to be an adult, is in a period of transition and self- discovery, which is why they often do acts known as juvenile delinquency [12].
• The origin of the region can be used as an excuse for delayed monthly payments. Student who come from outside the area will usually feel uncomfortable at being in a boarding school [13]. From the inconvenience, the student were likely to violate the rules in the Islamic boarding school.
• Economic problems can cause delays in the monthly payment of students [14]. Family economy is about the occupation of parents. So, the occupation of parents can be the reason for the delay in monthly payments to students, especially for students who are less fortunate because the boarding school has non provided scholarships for underprivileged students [15].
Male adolescents have higher juvenile delinquency when compared to female adolescents [16]. Delinquency is the improper use of money given by parents to make payments. From this, the authors take gender as one of the attributes in predicting the delay in monthly payments.

2.3. Preprocessing and Data Cleaning
The data used are student biodata data and student monthly payment data. The student biodata data is partly stored in the form of spreadsheets and some are stored in the form of printed documents while monthly payment data is stored in the form of a payment book. The data obtained is then combined and which data will be used. The data of student biodata consists of name, address, gender, place and date of birth, name of student guardian, class (if students are also at the school), student's room, student's information (active / alumni). Monthly payment data consists of number, date of payment, name, address, monthly payment details. Biodata data and monthly payment data are combined and entered into spreadsheets. Details of monthly payments are filtered whether students have been late in paying monthly fees or not. Then data on names, gender, origin, parents' occupation, and age and monthly payment information, because the data is completed. The entry form for student data has been handled in such a way that data on names, gender, origin, parents' occupation, and age is not empty and consistent. Preprocessing is done to improve the quality of data that will be processed in data mining. In this stage imbalancing data is done, because the "Terlambat" class and the "Rutin" class is not balanced.

2.4. Data Transformation
Data from the region is a wide range of data. The area of origin in the student data contains the name of the district or home town of each student. The origin of the region will be divided into two groups, namely “Luar Semarang” and “Semarang”. “Luar Semarang” and “Semarang” were chosen because the majority of student came from Semarang and the majority of student came from Central Java.

At the “Usia” of 13-18 years children enter the transition period and search for identity that is why they often do juvenile delinquency [12]. Then based on the theory of juvenile delinquency, age grouping will be divided into 2 categories > = 13 years and <13 years. Then based on the theory of juvenile delinquency, age grouping will be divided into 2 categories > = 13 years and <13 years. Occupation grouping will be divided into four categories, “Buruh”, “Karyawan Swasta”, “PNS” and “Lainnya”. Data transformation can be seen in Table 1.

| No | Attributes       | Value of Attributes | Information                        |
|----|------------------|---------------------|------------------------------------|
| 1  | Usia             | >=13                | Category for age more than same as 13 years |
|    |                  | <13                 | Category for age less than 13 years |
| 2  | Jenis Kelamin    | L                   | Category for male sex              |
|    | (Jenkel)         |                     |                                    |
| 3  | Pekerjaan        | Buruh               | Category for parent occupation labor (Buruh) |
|    |                  | Karyawan Swasta     | Category for parent occupation     |
|    |                  |                     | Private Employee (Karyawan Swasta) |
|    |                  | PNS                 | Category for parent occupation     |
|    |                  |                     | civil servant “PNS”                |
|    |                  | Lainnya             | Other category for parent occupation |
| 4  | Kota             | Semarang            | Semarang category for Semarang city / regency |
|    |                  | Luar Semarang       | Luar Semarang for cities / regencies outside Semarang |
2.5. Data Mining
This stage will extract information on monthly payment data. To achieve the KDD goal, the decision tree method is used as a data mining method and the C4.5 algorithm. Before making the decision tree using the C4.5 algorithm, several processes are carried out, namely data transformation, data partitioning, and imbalancing data process. Flowcharts from the decision tree formation stage can be seen in Figure 1.

![Flowchart Decision Tree Formation](image1)

Data transformation will be made into discrete data (categories). Data whose scope is too broad and will affect the KDD process so that it needs to be grouped into several small groups. Then randomly divide the monthly payment data into training data and test data. The training data will be used in the decision-making process, while the test data will be used to measure the performance of the decision tree that has been made. The training data that will be used in the decision-making process must be balanced first to avoid any tendency towards the majority class in the tree being made. The process to balance training data is done using random oversampling (ROS) methods. The ROS method is done by calculating the difference between the majority class and minority class. Then randomly selected data from minority classes. The data is then added to the minority class. The addition of data will be repeated until the amount of data in the majority class and minority class is the same [17]. The ROS method is illustrated by a flowchart in Figure 2.

![Flowchart Random Oversampling](image2)
After the training data is balanced, the training data is ready to be used for decision tree formation with the C4.5 algorithm. The C4.5 algorithm is an algorithm derived from ID3 [18]. Some developments were carried out in the C4.5 algorithm. The performance principle of the C4.5 algorithm is based [18] and [19] are:

- Make a decision tree. The purpose of this decision tree construction algorithm is to create a model of a training data set that will be used to predict new data classes.
- Pruning decision tree. Decision tree construction can be large and not easily "read", C4.5 algorithm can be simplified by pruning decision trees based on confidence values. Pruning also aims to reduce the prediction error rate on new data.
- Make a decision tree that has been built. Rule in the form of if-then comes from the decision tree by tracing from the root node to leaf node.

The basic algorithm used by algorithm C4.5 in the decision tree is the greedy algorithm that builds a top-down decision tree recursively doing divide and conquer [20]. Following are the steps of decision tree formation using C4.5 algorithm [18]:

1. Calculate the Entropy of case \( S \)

\[
Entropy(S) = -\sum_{j=1}^{k} \frac{freq(eq_j)}{|S|} \log_2 \left( \frac{freq(eq_j)}{|S|} \right) \tag{1}
\]

Information:
- \( Entropy(S) \) = entropy from case \( S \)
- \( k \) = number of classes in the case \( S \)
- \( j \) = iteration for the class in the case \( S \)

If the total case only has one class (routine or late), then make the node as a leaf with the majority class.

2. Calculates information gain for each attribute. Information gain equation can be seen in the equation (2).

\[
gain(X) = Entropy(S) - Entropyx(S) \tag{2}
\]

Information:
- \( Gain(X) \) = Gain sought from attributes \( X \)
- \( Entropy(S) \) = Entropy from case \( S \)
- \( Entropyx(S) \) = Entropy for attribute \( X \) in the case \( S \)

3. Calculate split info for each attribute. The split info equation can be seen in the equation (3).

\[
SplitInfo(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \left( \frac{|S_i|}{|S|} \right) \tag{3}
\]

Information:
- \( SplitInfo(X) \) = split info from attributes \( X \)
- \( n \) = Number of case partitions \( S \)
- \( |S_i| \) = amount of data on the partition of case \( S \)
- \( |S| \) = Total number of data case \( S \)
- \( Entropy(S) \) = entropy on the partition \( i \)

4. Use of information gain in the ID3 algorithm It has a strong bias for testing with many results. The gain ratio is added to the C4.5 algorithm to overcome the lack of information gain. The formula for calculating gain ratio can be seen in the equation (5).

\[
gainratio(X) = \frac{gain(X)}{\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \left( \frac{|S_i|}{|S|} \right)} \tag{5}
\]

Information:
- \( gainratio(X) \) = gain ratio sought from attribute \( X \)
gain (X) = information gain from attribute X
splitinfo(X) = split info from attribute X
5. Attributes with the biggest gain ratio are selected as nodes.
6. Share data based on attribute values of selected attributes. Then use it to do the next step.
7. Repeat steps 1 through 6 until all attributes are used or fulfill a stop condition.

After the tree is formed then the tree is run down to overcome the problem of overfitting data. Trees are run by using the error-based pruning method. Error Based Pruning is one way to do a popular pessimistic error evaluation used in C4.5 algorithm. In Error Based Pruning confidence intervals are calculated to show the probability of misclassification and the upper limit of the error rate to be compared [21]. The error-based pruning method is as follows:
1. Calculate the total number of cases, the number of cases with routine decisions, and cases with late decisions from the sub-tree to be counted.
2. Calculate the p value for each node. Value is obtained by dividing the number of cases with the decision of the minority class with the total number of cases.
3. Calculate the error estimate from each node. Error estimates are calculated using equation (6).

\[ p = \frac{\text{error data training}}{\text{confidence limit}}, \text{where } z = z_1 - (a/2) \text{ to } a \text{ confidence level} \ n = \text{number of data in leaf} \]

4. Calculate the average estimate error for the child node according to the ratio. The ratio for the child node is calculated by dividing the number of cases of child nodes with the number of cases of parent nodes.
5. Comparing error estimates for children and parents. If the error estimates that the parent is smaller than the child, the tree will be cut. Then the parent node will be converted into leaves with the majority class value. Conversely, if the child node is smaller than the parent, the tree is not cut.
6. Repeat steps 1 through 5 until all sub-trees are checked.

After the process is passed, decision tree modeling can be used for prediction.

3. Experimental result
3.1. Interpretation decision tree
The pattern generated from the KDD process. The use of the Data Mining Algorithm can be displayed in the form of a decision tree. The results of the decision tree formed from the C4.5 Algorithm process from the data of late monthly payments can be seen in Figure 3.

From Fig. 8 rules can be taken as follows:
• IF pekerjaan “LAINNYA” AND jenkel “L” AND kota “LUAR SEMARANG” THEN “TERLAMBAT”
• IF pekerjaan “LAINNYA” AND jenkel “L” AND kota “SEMARANG” THEN “RUTIN”
• IF pekerjaan “LAINNYA” AND jenkel “P” THEN “RUTIN”
• IF pekerjaan “KARYAWAN SWASTA” THEN “RUTIN”
• IF pekerjaan “BURUH” THEN “TERLAMBAT”
• IF pekerjaan “PNS” THEN “RUTIN”

3.2. Evaluation decision tree
Decision tree performance measurement is done by measuring the performance of decision trees that have been built. Performance measurement is done by random subsampling and confusion matrix measurements. Confusion matrix is obtained by comparing the results of the prediction of the application and the actual results. To determine the performance of the decision tree, precision, recall, and accuracy are used.

The performance measurement process was carried out 15 times with each test using 52 different test data. The results of the calculation of precision, recall and accuracy can be seen in Table 2. Accuracy values show the closeness of results to the actual values. Precision values show how close the results of an examination are repeated with the same sample. The smaller precision value will more accurate and the greater the precision value, less accurate the method. Recall is the quality of how complete relevant results are displayed.

| Measurement | Precision | Recall | Accuracy |
|-------------|-----------|--------|----------|
| 1           | 78.79%    | 92.86% | 82.69%   |
| 2           | 78.13%    | 89.29% | 80.77%   |
| 3           | 75%       | 96.43% | 80.77%   |
| 4           | 81.25%    | 92.86% | 84.62%   |
| 5           | 75%       | 85.71% | 76.92%   |
| 6           | 78.79%    | 92.86% | 82.69%   |
| 7           | 83.33%    | 89.29% | 84.62%   |
| 8           | 80%       | 85.71% | 80.77%   |
| 9           | 74.29%    | 92.86% | 78.85%   |
| 10          | 67.57%    | 89.29% | 71.15%   |
| 11          | 78.79%    | 92.86% | 82.69%   |
| 12          | 86.21%    | 89.29% | 86.54%   |
| 13          | 72.97%    | 96.43% | 78.85%   |
| 14          | 74.29%    | 92.86% | 78.85%   |
| 15          | 80%       | 100%   | 86.54%   |
| Average     | 77.62%    | 91.90% | 81.15%   |
| ± Standard Deviation | 0.05 | 0.04 | 0.04 |

The highest accuracy of 15 tests is 86.54%. The parameters used in decision tree formation are as follows:
• The minimum number of data (objects) on a branch is 0
• The use of error-based pruning as a pruning method with a confidence value is 0.25
• Using random oversampling as an imbalanced data method

The average of testing was 15 times with each test using 52 different test data, obtained an average of 77.62% precision, an average recall of 91.90% and an average accuracy of 81.15%. The standard deviation for precision is 0.05, the standard deviation of the recall is 0.04, and the standard deviation accuracy is 0.04. The standard deviation values less than 1. It can be interpreted that the result of model is quite good. It shows the stable results and does not cause bias.

3.3 Discovered knowledge
The results of the KDD process are used to predict the delay in monthly student payments whether a student will make a late or routine monthly payment. These results can be used to overcome the delay in monthly payments earlier, so as to guarantee the fulfillment of student's basic needs. For students who are predicted to delay payments, early treatment can be carried out, such as appealing to
“santri”, recommendations for scholarship to poor families. With this application can improve service and trust of parents towards the survival of students in schools.

4. Conclusion
The conclusions that can be drawn from the results of this study are the use of C4.5 algorithm for predictions of student monthly payment delays in Assalafi Alfiithrah Meteseh Semarang produces an average accuracy of 81.15%, with an average value of precision is 77.62% and the average recall value is 91.90%. The most influential attribute in predicting student monthly payments is the occupation of student’s parent the attribute that does not affect the late monthly payment is the origin of student.

Reference
[1] Shiman D 1999 ECONOMIC AND SOCIAL JUSTICE A Human Rights Perspective (Washington: University of Minnesota).
[2] Chandra B, Mazumdar S, Arena V and Parimi N 2002 WISE 2002 - Proc. 3rd Int. Conf. Web Inf. Syst. Eng. Work. 160.
[3] Berzal F, Cubero J, Cuenca F, and Martín-Bautista M 2003 Data Knowl. Eng. 44 31
[4] Su J and Zhang H 2006 “A Fast Decision Tree Learning Algorithm,” in Proceedings of the 21st national conference on Artificial intelligence.
[5] Anyanwu M and Shiva S 2009 Int. J. Comput. Sci. Secur. 3 230
[6] Lavanya D and Rani K 2011 Int. J. Comput. Appl. 261.
[7] Wahiyuni S, Saputra K, and Perangin-perangin I 2017 “the Implementation of Decision Tree Algorithm C4. 5 Using Rapidminer in the Implementation of Decision Tree Algorithm C4. 5 Using Rapidminer in Analyzing Dropout Students,” Conference: International Conference on Technical and Vocation Education and Training, At Padang Indonesia.
[8] Al-Barrak M and Al-Razgan M 2016 Int. J. Inf. Educ. Technol. 6 528
[9] Adhutrao K, Gaykar A, Dhawan A, Jha R, and Honrao V 2015 Int. J. Data Min. Knowl. Manag. Process 3 39
[10] Hssina B, Merbouha A, Ezzikouri H, and Erritali M 2014 Int. J. Adv. Comput. Sci. Appl. 4 13.
[11] Mohankumar M, Amuthakkani S, and Jeyamala G 2016 Int. J. Adv. Res. Biol. Eng. Sci. Technol. 2 360.
[12] Samara D, Humaedic S, and Santoso M 2017 Jurnal Penelitian dan PPM 4 129
[13] Kosasih A, Komariah S, and Setiawan Y 2014 Jurnal pendidikan Sosiologi 5 1
[14] Iskandar M 2018 “Metode Musyrif Dalam Mengatasi Kenakalan Santri Di Pondok Pesantren Attaqwa Putera Kecamatan Babelan Kabupaten Kabu,” (Undergraduate Thesis-UMS Alaudin Makasar)
[15] Suaranbt, “Telat Bayar SPP Karena Kendala Ekonomi.” 2017. [Online]. Available: http://www.suarantb.com/news/2017/03/22/27542/mahasiswa.iain.mengaku.telat.bayar.spp.karena.kendala.ekonomi. [Accessed: 09-Apr-2018].
[16] Riskinayasari G 2015 “Kenakalan Remaja Ditinjau dari Kosep Diri dan Jenis Kelamin,” (Undergraduate Thesis-UMS Surakarta)
[17] Saifudin A and Wahono R 2015 J. Softw. Eng. 1 76.
[18] Quinlan J 1993 C4.5 programs for machine learning (Amerika: The Morgan Kaufmann).
[19] Han J and Kamber M 2006 Data Mining: Concepts and Techniques (United States: Elsavier Science and Tehnology)
[20] Ananda D and Wibisono A, 2014 J. Inf. Syst. 10 29
[21] Grabczewski K 2014 Meta-Learning in Decision Tree Induction (Switzerland: Springer).