Employees' Attendance Patterns Prediction using Classification Algorithm Case Study: A Private Company in Indonesia

Nunung N. Qomariyah and Yudho G. Sucahyo

***Abstract*—**Employee's attendance status is a criterion that has been used for performance appraisal in a company. By utilizing data mining technology, the relationship between employee's characteristics and level of employee absences can be learned from the employee's attendance data pattern.

This research used case study as research method. Objective of this research was to find the special characteristics of groups of employees which showed frequent absence pattern in the office. Decision tree as part of classification algorithms in the data mining has been chose to observe special pattern of absenteeism within two year period. Software Orange Data Mining was used as a tool in processing the data. At the end of the data analysis, it was found five rules for classifying characteristics of employees related to absenteeism.

***Keywords*—**data mining, employee's absence, decision tree, classification.

I. Introduction

CLASSIFICATION is a process to finds the same properties on a set of objects in database, and classifies them into different classes according to the specified classification models. The algorithm given some input, analyze them and produces a prediction [1]. The prediction accuracy defines how "good" the algorithm is. To establish classification model, a sample database 'E' is treated as training set, where each tuple consists of set of attributes that same as set of attributes of tuples in a large database 'W'. Each tuple is identified by a label or class identity. Purpose of this classification is to analyze a set of training data and specified accurate description or model for each class based on some features captured from the data. Description of each class then used to classify the tested data in the database 'W', or to constructs a better description for each class in the database. A common example of the use of the classification model is prediction of the bank lending risk [2]. The data to be examined consists of the people who have received some credits for the bank. Most creditors pay the obligation well and some of them do not pay it well. Data mining classification will be able to define what the most influential attributes are. In this research, classification was chose as an appropriate method to find out the characteristics of high frequent absence employees.

One of the popular classification models that have been widely used nowadays is *decision tree*. This is a predictive model using hierarchical tree structure or structures [3]. Decision Tree is very popular because it is easy to interpret. It has flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents the results of the test, and leaf nodes represent classes or class distributions.

Nunung N. Qomariyah is a lecturer of Universitas Pembangunan Jaya (UPJ) in Indonesia. She is specialized in Data mining and Artificial Intelligence. (e-mail: nunung.nq@upj.ac.id).

Yudho G. Sucahyo is a lecturer of University of Indonesia (UI) in Indonesia. He is specialized in Data Mining and E-Government System. (e-mail: yudho@cs.ui.ac.id).

II. Case Study

A company that became an object of study of this research was a private company operated in Jakarta, Indonesia, which has about 30 of total employees. The company operated in manufacturing business industry especially in Consumer Goods Sanitary Napkin since 1997. Human Resource Department (HRD) in the company maintained records for the employee's attendance in office from 2006. However in this research, only two years period of data, from 2009 until 2011, that was analyzed. This is a period when the policies of un-attendance have been applied to all employees. The total number of record processed in this research was 14,400. For privacy and security of company data, some attributes such as employee's name, were hidden.

III. Tools For Data Mining

In this research, we used Orange as a tool for data mining. Orange is open source software for data visualization and analysis through visual programming or Python scripting [4]. There are some features available in Orange: preprocessing data, scoring, filtering, modeling, model evaluating, and exploration techniques. All this features are implemented in C++ language for its speed and Python language for its flexibility. Orange is free software under the terms of the GNU (General Public License). It is maintained by Faculty of Computer and Information Science, University of Ljubljana, Slovenia, together with open source community [4].

IV. Research Design

There were 14.400 data of employee attendance records classified into three categories based on the frequency of absence every month. HRD rules was used to determining the target class for the data, they were:

1. If there were more than 10% of absence calculating in period of three months, it would be considered as "frequent absent employee" (we say "*sering bolos*" in Indonesian)
2. If there were 5 - 10% of absence calculating in period of three months, it would be considered as "rare absent employee" (we say "*jarang bolos*" in Indonesian)
3. If there were below 5% of absence calculating in period of three months, it would be considered as "frequent present employee" (we say "*rajin masuk*" in Indonesian)

When a set of rules has been decided in the decision tree model, the next important thing that should be carried out was Pruning Decision Tree, which would eliminate unnecessary rules [5]. The Pruning Decision Tree steps are:

1. *Eliminate unnecessary antecedents by:*

a. Build a contingency table for each rule that has multiple antecedents.
b. Simplify rules by eliminating antecedent rule which does not affect the conclusion by using the following independence tests:

- Chi-Square Test if the expected values frequency is above 10.
- Yates Correction for Continuity if the expected values frequency is between 5 and 10.
- Fisher's Exact Test if the expected values frequency is below 5.

2. *Eliminate unnecessary rule.*

This research was conducted in Bahasa Indonesia, so there were many variables that we wrote in Bahasa. We also included the translation of the variable in English next to the table or script.

V. Data Testing And Analysis

Data testing and analysis was carried out by using the Orange software. The whole process was shown in Orange canvas below:

1. *Data Preparation*

In the data preparation step, there were three important things should be conducted:

A. Data Integration

The raw data from HRD was available in the format of Microsoft Excel. In this step, we should combine the employee master data with employees' attendance data to gather the required information. Then the combined data was exported into a text format (tab delimited format) to be read by Orange software. Those data still contained some noises which were should be cleaned up in the next step.

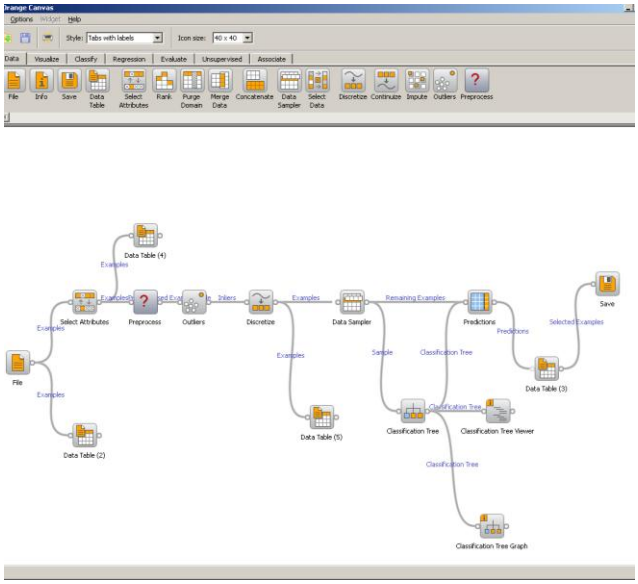

Fig. 1 Data Processing showed in Orange Canvas

The following schema merging the two separate data:

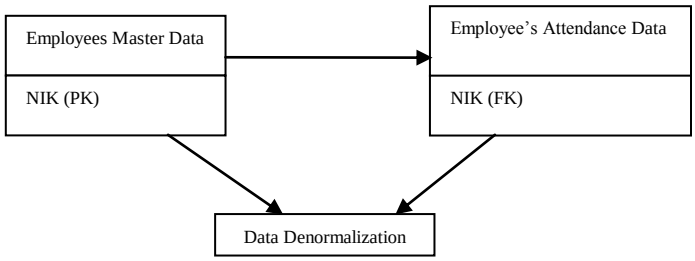

Fig. 2 Integration Process and Data Denormalization

The raw data contained 142 original attributes, which had been selected to be used in the data processing. The attributes included in the data processing was shown in the table below:

TABLE 1
THE CHOSEN ATTRIBUTES

No.	Attribute
1	NIK (Identity number of employee)
2	Present Level
3	Status
4	Position Name
5	Work period
6	Age as of This Year
7	Sex
8	Marital Status
9	Number of Children

B. Data Cleaning and Transformation

The raw data still contained invalid records, some of them were miss type such as a record "D" in the present level column, this column should be filled by number 1 to 7 or empty for contract-based employee. The invalid records were removed from the data and not included in the next process. We used Preprocess feature in Orange and activated the option "remove missing". Once the data was clean from noise then

the transformation process was carried out to standardize the data format. We used Ms Excel for this step. We counted work period, current age and number of children.

- Work period = Year (Date Hired - To Present)
  We rounded to the annual work period, if the period of employment was less than 1 year it was considered 1 year of work period.
- Current Age = Year (Age as of this Year)
- Number of Children = Count (Name of Child 1, Child 2 Name, Child's Name 3)

Here is an example of the data that has been transformed:

TABLE II
TRANSFORMED DATA

NIK	Present Level	Status	Position Name	Working Period	Age	Sex	Marital Status	Number of Children
7000009	7	Active	Assistant HRD Manager	5	35	Male	Married	3
7000010	5	Active	Director of Sales	5	37	Male	Married	3
7000011	7	Active	Area Sales Supervisor	5	46	Male	Married	2
7000012	5	Active	Finance Supervisor	4	58	Male	Married	1
7000013	5	Contract	Sales Executive	5	49	Male	Married	1
7000014	5	Active	Sales Executive	4	37	Male	Married	1
7000015	7	Active	Area Sales Manager	4	37	Male	Married	2
7000016	5	Active	Area Sales Supervisor	4	35	Male	Single	0
7000017	5	Active	Area Sales Supervisor	4	49	Male	Single	0

The next step was classifying the data using HRD parameters. The transJJformation process was performed on the target class by counting daily attendance data for each row record. In Ms Excel we used a function COUNTIF().

Examples of some results of the target class transformation are shown as follows:

TABLE III
DAILY ATTENDANCE DATA

NIK	Day of Work in a Month (date)														
	1	4	5	6	7	8	11	12	13	14	15	18	19	20	21
7000009	1	1	1	1	1	1	1	C	C	C	C	1	1	1	1
7000010	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7000011	1	1	1	1	1	1	1	1	1	1	1	S	S	S	1
7000012	1	1	1	1	1	1	1	1	1	1	I	1	1	1	1
7000013	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7000014	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

*Translation:
C = Cuti (authorized leave); S = Sakit (sick leave); I = Ijin (unattended for other reason); A = Absent

TABLE IV
TARGET CLASS TRANSFORMATION

NIK	Recapitulation of attendance (I/S/C/A)			Target class
	Month 1	Month 2	Month 3	
7000009	4	1	3	Sering bolos
7000010	0	0	0	Rajin masuk
7000011	3	0	0	Jarang bolos
7000012	1	0	0	Rajin masuk
7000013	0	0	0	Rajin masuk
7000014	0	0	0	Rajin masuk

*Translation:
Sering bolos = frequent absent employee; Jarang bolos = rare absent employee; Rajin masuk = frequent present employee

2. *Data Exploration*

There were two steps in the data exploration as follows:

A. Data outlier identification

In this step we identified the outlier data using Z = 2,0. We found 12 data were outlier. These outlier data showed as below:

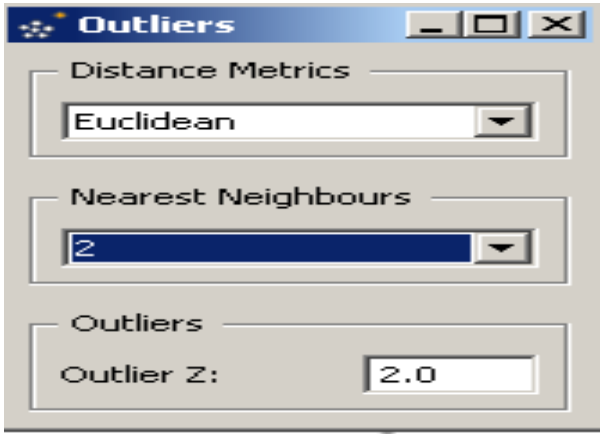

Fig. 3 Data Outlier Settings

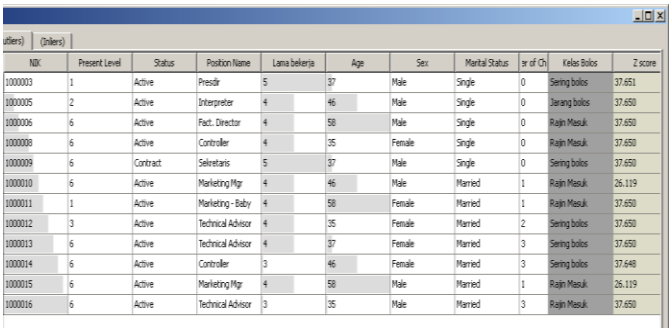

Fig. 4 Data Outlier Results

B. Discretization

Discretization step was performed to the continuous attributes: age and work period. We did this to simplify the data. We used *Equal-Width Discretization* with *number of interval* = 5 for age attribute, and *number of interval* = 6 for work period attribute. The results of discretization process shown as bellow:

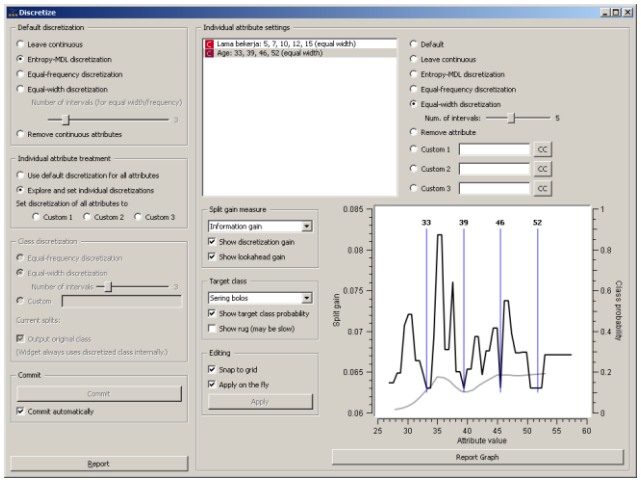

Fig. 5 Discretization Process

3. *Data Modeling*

In the modeling phase there were two important things we should perform. The first step was data treatment by utilizing the Sampler Data feature in Orange. We divided the data for training and testing, the ratio was 80% : 20% respectively. Then the second step was to determine the technique, we chose decision tree classification algorithm to examine the data.

A. Testing the training data

The parameter settings in the Classification Tree to form a decision tree based on attributes was shown as bellow:

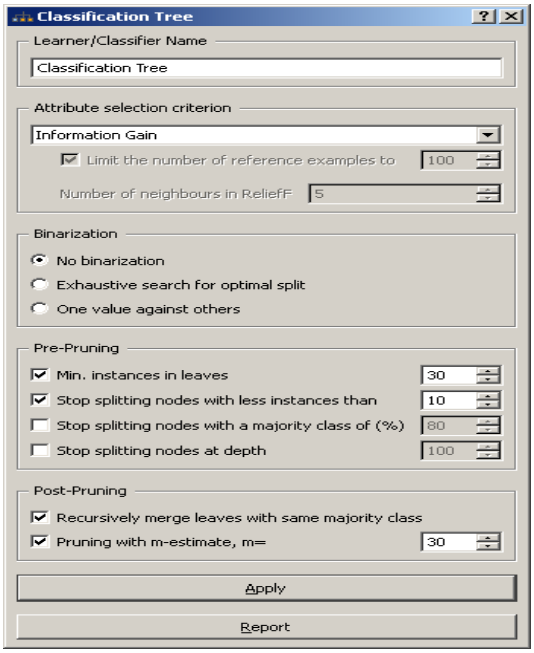

Fig. 6 Parameter Settings

Pruning value with smaller m-estimate will produce more rules that show high level of accuracy, but if the value of m-estimate is greater, it will lead to low level of accuracy. In this research, the chosen value of m = 30 that formed 36 nodes and 25 leaves.

The classification tree and rules that was generated from the specified algorithms and parameter settings was shown below:

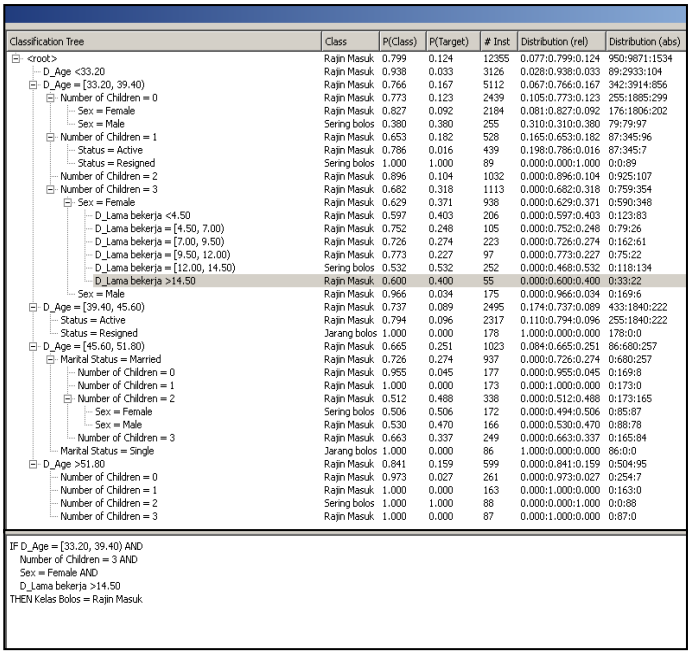

Fig. 7 Classification Tree and Rules

B. Examining the testing data

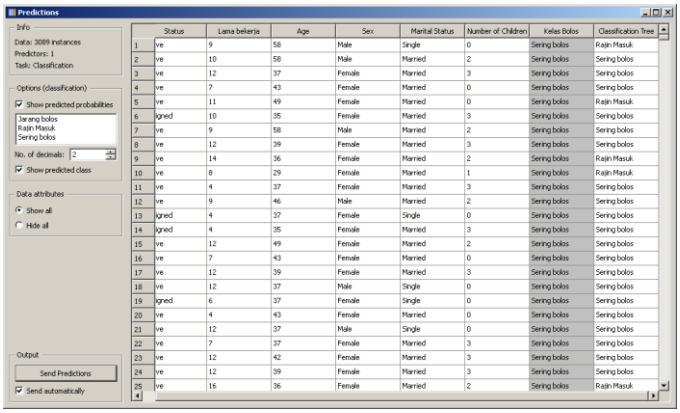

Fig. 8 Data Testing Examining Accuracy

This step was performed using Prediction feature in Orange. We have got the correct prediction as much as 2936 events, and the incorrect prediction as much as 153 events. From this result it could be seen that the test accuracy was 95.05%.

VI. CONCLUSION

From the results of the decision tree, there were 5 rules which led to end conclusion of "frequent absent employee", as follows:

We could translate the rules into English, it mean the employees with these characteristics:

1. Female employee
2. She had 3 children
3. She's aged was around 33-39 years old
4. She had working period around 12-14 years

Had tend to had more days of work absence than other characteristics.

TABLE V
RULES OF CLASSIFICATION

Rules	Class	P(class)	P(target)	Number of Instances	Distribution (relative)	Distribution (absolute)
IF D_Age >51.80 AND Number of Children = 2 THEN Kelas Bolos = Sering bolos	Sering bolos	1	1	88	0.000:0.000: 1.000	0:0:88
IF D_Age = [45.60, 51.80) AND Marital Status = Married AND Number of Children = 2 AND Sex = Female THEN Kelas Bolos = Sering bolos	Sering bolos	0.506	0.506	172	0.000:0.494: 0.506	0:85:87
IF D_Age = [33.20, 39.40) AND Number of Children = 3 AND Sex = Female AND D_Lama bekerja = [12.00, 14.50) THEN Kelas Bolos = Sering bolos	Sering bolos	0.532	0.532	252	0.000:0.468: 0.532	0:118:134
IF D_Age = [33.20, 39.40) AND Number of Children = 1 AND Status = Resigned THEN Kelas Bolos = Sering bolos	Sering bolos	1	1	89	0.000:0.000: 1.000	0:0:89
IF D_Age = [33.20, 39.40) AND Number of Children = 0 AND Sex = Male THEN Kelas Bolos = Sering bolos	Sering bolos	0.506	0.506	172	0.000:0.494: 0.506	0:85:87

*Translation:
Kelas Bolos = Target Class

The results of this study were quite interesting. However, with the limitations of this study, which only used limited data sample we cannot generalized the accuracy. It is expected that in future work, the sample size can be increased, so it will get the higher accuracy. This limitation was due to the limitations of both the resource and processor memory that was used. It was recommended using a large capacity of machine to process the larger data.

REFERENCES

[1] Hand, D. J. Discrimination and Classification. John Wiley. Chichester, 1981.
[2] Bhasin, M. L. Data Mining: A Competitive Tool in the Banking and Retail Industries. The Chartered Accountant, October 2006.
[3] Berry, M. J. and Linoff, G. S. Mastering data mining. John Wiley& Sons. New York, 2000.
[4] Orange Data Mining Documentation. http://www.ailab.si/orange/doc/
[5] Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques. Morgan Kauffman. MA, USA. 2012.