

Prediction of Employee Attendance Factors Using C4.5 Algorithm, Random Tree, Random Forest

RIZA FAHLAPI, HERMANTO, ANTONIUS YADI KUNTORO,
LASMAN EFFENDI, RIDATU OCA NITRA

ABSTRACT

Research on the performance of workers based on the determination of standard working hours for absences conducted by workers in a certain period. In disciplinary supervision, workers are expected to be able to provide the best performance in the implementation of work in accordance with predetermined working hours. The measurement of the level of discipline of admission hours for placement workers is carried out every working day, continuously and continuously. Attendance monitoring already uses online attendance by using data downloaded from the online attendance provider as the main data. In addition, data collection is done by filtering employee absentee data and supporting information on the categories that cause mismatches in meeting work schedules. Mobilization of workers according to location and working hours has been regulated in company regulations allowing the placement of workers in accordance with the residence so as not to affect the desired work results the company is still within reasonable limits and can be increased. The assessment of this study as a progression factor inhibiting the company in achieving company targets. From the results of the author's analysis of the prediction of employee delay factors using three algorithms, namely the C.45 algorithm accuracy = 79.37% and AUC value = 0.646, Random Forest Algorithm accuracy = 78.58% and AUC value = 0.807 while for the Random Tree algorithm accuracy = 76.26% and the AUC value = 0.610.

Keywords: Employee, performance, accuracy, Random Three, Random Forest.

INTRODUCTION

The development of labor mobilization in the current technological era, allows increasingly narrow gaps for workers who have poor performance in their jobs. In this case, it needs to be followed up related to staffing absenteeism which is a basic concern, because the beginning of work discipline can be assessed from the compliance of employee working hours. Human Resources (HR) is the main aspect for the company or institution, because it is the main driver. For this reason, the company pays attention to this matter, especially in matters relating to motivation, job satisfaction, work involvement or workforce placement to commit to an organization, so that human resource management can run in accordance with what is expected.

To improve the image, work, and performance of the company towards the best direction and

support the sustainability of the company, it is necessary to unite the direction and outlook for workers that can be used as a guide or reference in carrying out both managerial and operational tasks in all fields of tasks and organizational units. For this reason, whether research in disciplining employees in terms of their presence to improve the image, work, and performance of employees can be proven through the algorithmic method that will be examined in this paper.

So far according to the company there are still many workers who are undisciplined and not monitored by employees who commit violations. To overcome this, then the initiative to drive back using an online attendance machine. Data taken from the online attendance machine in the company is used as a reference data in this study.

To prove that by implementing an online attendance machine can improve employee discipline and can show the maximum service

performance of workers to the company. From this background, researchers want to know the size of employee discipline by using data from an online attendance machine on employee performance satisfaction, especially in companies that are the focus of research.

Data Mining Process

Process with Data Mining must be in accordance with the procedures of the process itself, the process is Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is an effort to standardize process data mining. CRISP-DM process provides the life cycle of a data mining project. The process contains the project phase, each task, and the relationship between tasks.

There are six interrelated phases to describe CRISP-DM process in data mining, namely: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Rokach & Maimon, 2010), as in the following Figure 1.

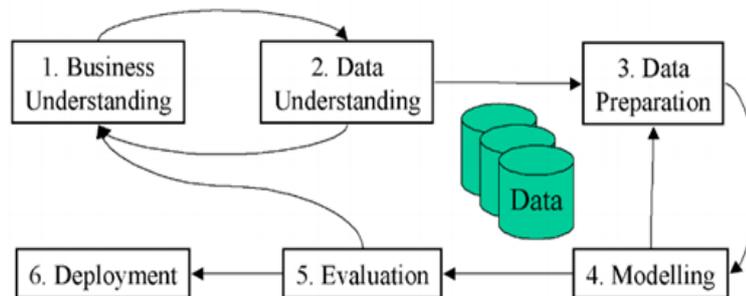


FIGURE 1. CRISP-DM Process
Source: (Rokach & Maimon, 2010).

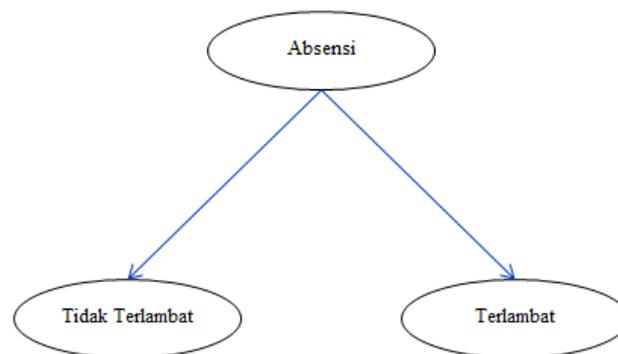


FIGURE 2. Binary feature testing requirements (Prasetyo, 2014)

Algorithm C4.5

Improved version of ID3 was introduced C4.5 algorithm. In ID3, induction of decision trees can only be done on the features of the type of category (nominal and ordinal), while the numeric type (interval or ratio) cannot be used. What distinguishes the improvement of the C4.5 algorithm from ID3 is that it can handle numeric type features, pruning decision tree, and deriving rule set. The gain criterion in determining the C4.5 algorithm is also used as a feature to break the nodes in the induced tree (Prasetyo, 2014).

The important thing in induction of a decision tree is how to state the test conditions on the node. There are 3 important groups in terms of node testing:

a. Binary Features

Having two different values is called a binary feature. The testing requirements when this feature becomes a node (both root and internal) have only two branch options. Examples of solutions are presented in Figure 2

b. Categorical type features

Features whose values are categorical (nominal and ordinal) can have several different values. One example is the 'weather' feature has 3 different values, and this can have many combinations of test solving conditions. In general there are 2, namely binary splitting (binary splitting) and (multi splitting).

c. The numeric type feature

For numeric features, tests in nodes (both internal and root) are expressed by comparison testing ($A < v$) or ($A \geq v$) with binary results, or for multi with results in the form of a range of values in the form $v_i \leq A < v_i + 1$, for $i = 1, 2, \dots, k$. In the case of binary solving, the algorithm checks all possible solving positions pemecahan and selects the best position v . For multi methods, the algorithm must check all possible continuous values.

Random Tree

Supervised learning algorithm developed by Breiman is called Random Tree (Kalmegh, 2015). Random Tree is a combination of 2 ideas of the classification algorithm, namely model (Decision Tree) of the single tree and Random Forest (Pfahring, 2011). Combining the concepts of both algorithms is expected to acquire a better decision tree algorithm. The success and efficiency of the Random Tree rely heavily on some particular technique features. Determine the best separation point for valuable attributes.

Random Forest

Random Forest was first introduced by Leo Breiman through his laboratory in 2001. One combination of tree prediction is called Random Forest, so relies Random vector value relies each of tree. For all trees in the forest takes separately and with the same distribution (Kalmegh, 2015). Random Forest grows a lot of classification trees. To classify the new object, input vector into each tree in the forest. there gives a classification called a "vote" for the class. The forest selects the classification that has the most votes (exceeding all trees in the forest) (Breiman & Cutler, 2007).

RapidMiner

RapidMiner is open-source software and solution for analysis of data mining, text mining, and prediction analysis. This software uses a variety of desktop techniques and predictions and providing insight to the user to make the most informed decisions. Approximately 500 data mining operators, including operators for input, output, data preprocessing, and visualization. It is a standalone software for data analysis and as a data mining machine that can be integrated into its products. Using Java language and can work in all systems (Aprilla Dennis, 2013).

Confusion Matrix

In supervised learning Confusion Matrix is a visualization tool which is commonly used (Gorunescu, 2011). Each column in the matrix is an example in the prediction class, each row represents an event in the actual class. One advantage of the Confusion Matrix is that it's easy to know if the data exists between two classes (mislabeling). Contains information about the actual and predicted in the classification system The performance is usually evaluated using data on the matrix.

Evaluation of the classification model is to estimate the right and wrong objects, the test sequence is contained in the confusion matrix. Predicted class displayed at the top of the matrix and the observed class on the left side. To predicted the observed class, Each cell contains a number that shows how many actual cases (Hastuti, 2012).

ROC Curve

The another way to test the classification performance with a ROC Curve, chart is a plot with an incorrect positive rate (FP) on the X-axis and the correct positive (TP) level on the Y-axis. Point (0.1) is the perfect classification which classifies all positive cases and negative cases correctly, because the wrong positive rate (FP) is 0 (no), and the correct positive rate (TP) is 1. The point (0.0) is a classification that predicts each case to be negative, while the point (1.1) corresponds to a classification that predicts each case to be positive. Point (1.0) is an incorrect classification for all classifications (Gorunescu, 2011).

K-Fold Cross Validation

The validation technique that divides the data into K sections and then each section will be done classification process using this technique will be conducted a trial of K. Each experiment will perform using a single data testing and the K-1 section will be training data, then the data testing will be exchanged for one piece of data training so that for each experiment will be obtained different testing data (Witten, Frank, & Hall, 2011). Data training will be used in conducting learning while data testing that has never been used as a learning and will serve of truth testing or accuracy. In Case The value of K used amounted to 10 or 10-Fold Cross-Validation.

T-Test

The T-Test is a hypothesis testing method using one individual (object research) using two different treatments. Although using the same object but the sample remains divided into two namely the data with the first treatment and the data with the second treatment. Performance can be known by comparing the condition of the first research object and the condition of the object in the second study (Hastuti, 2012).

Classification

In detail as a work that performs training/learning for the f target function that maps each vector (feature set) χ into one of many available Y-class labels. The training workshop will result in a model that is then stored as memory defined. The input and give the answer as the output in the classification Model has the same meaning as the black box, where a model that receives the input model can be performed (Prasetyo, 2014).

RESEARCH METHOD

CRISP-DM or Cross-Industry is a standard methodology for Data Mining Process. The result of collaboration from several companies, such as Daimler-Benz, Ohra, NCR Corp, and SPSS Inc. Which began to be initiated since 1999 (North, 2012).

Figure 3 CRISP-DM has six stages (North, 2012) (a) Business Understanding, first step to be obtained in the form of common questions, should be defined knowledge to managing profits, preventive action for anticipate faults defects of products, etc, (b) Data Understanding, We have to collect, identify, and understand the data assets. Every single data should be verifiable in correctness and reliable, (c) Data Preparation, these activities to cleaning, reformatting, reducing amount of noise data, and set up data consistent with the required format, (d) Modeling, the representation of the observations that result from the search, identification, and patterns contained, (e) Evaluation, the usability value can be determine to the model that we have made in the previous step, (f) Deployment, the results of all process.

1. CRISP-DM Stages

The research used experimental methods to compare and evaluate the workers' discipline at the opening hours or the working hours specified in the employment agreement. In designing research method of this experiment in data mining is Cross-Industry Standart Process for Data Mining (CRISP-DM) that consists of 6 phases steps (Larose, 2005).

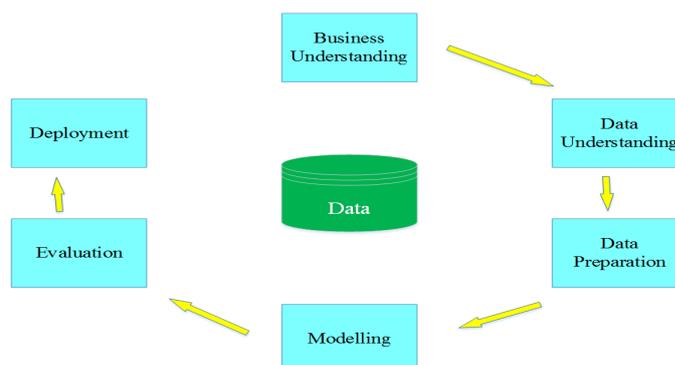


Figure 3. CRISP-DM (North, 2012)

2. Framework

The author makes a framework of the thinking of Irian which is useful as a guideline in the research \so that this research can be done consistently. This research consists of several stages as seen in Figure 3.2 problem (problem) In this research is whether the algorithm can be applied to the determination of the measure of labor discipline in the working hours and which algorithm will provide the best model for classifications of performance determinant of workers.

For that created model using algorithms C 4.5, random Tree and Random Forest to solve the problem then done testing to both performance of the model. After testing against both models formed, it will be tested with 10 fold cross-validation. Accuracy models will be formed measured using the confusion matrix Under Curve Area (AUC) and Receiver Operating Characteristic (ROC) Curve. To develop applications (deployment) based on the model that has been made then the tools used Rapid Miner 8.0. The following depictions of thought frameworks have been performed as follows figure 4.

RESULT AND DISSCUSSION

1. Implementation Of The Methodology

Based on the research methodology that has been described in chapter III, the implementation of the methodology is done in this study.

2. Research Methods

In this research, the author conducts research using the CRISP-Standard research method (Cross-Standart Industry for Data Mining). The stages of Crisp-Dm consist of Business Understanding, Data understanding.

3. Business Understanding

Understanding the data of workers in PT Permata Karya Jasa that in reviewing personnel performance in the level of discipline is still common, therefore researchers here propose a new attribute of testing existing data and develop it into a new predictive value to be used as an evaluation material in determining the effectiveness of workers in carrying out their duties. Later, this above will be tested by comparison of two algorithms namely C 4.5, random Tree, and Random Forest.

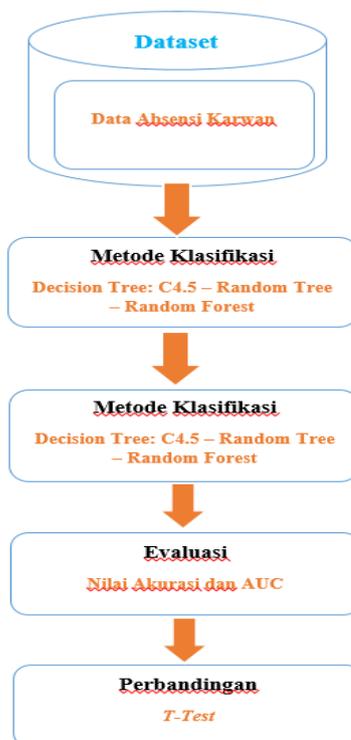


Figure 4. Framework

4. Data Understanding

At this stage the author examines the data Absent officer over the last few months, the data taken is the hour data in the work of employees in the period of January to April 2019, and for now, also added data from the domicile of each worker. For the data that we present in the implementation, among others: date, list of worker names, hours of attendance, entry/exit workers, attendance category, and also domicile. The data

downloaded from an online attendance worker in a project of 215 employees. Where the number of workers consists of 3538 worker attendance records. The number of attendance is influenced by how active the worker in the arrival attendance. Here is the initial data before the number of values in the Convert see table 1 below. Table 2 above is the number of employee absentee days from all workers for several periods that have been downloaded and the data will be processed.

TABLE 1. List of Employee Attendan

Date	In Time	Employee Name	Pin	Age	Status
1/4/2019	08.08.45	EDI POERNOMO	2119920064	26	Present
1/4/2019	14.54.41	ADRIANSYAH	2119910031	27	Present
1/4/2019	07.34.02	DIAN HANAFI S	2119820074	36	Present
1/4/2019	07.06.28	RASWADI	2118910140	27	Present
1/4/2019	07.19.25	R. DWI ADE ISWAHYUDI	2119930048	25	Present
1/4/2019	07.07.44	RONY PRASTYO RAHARJO	2119870049	32	Present
1/4/2019	05.59.28	Heru Widodo	2119900050	29	Present
1/4/2019	07.23.30	MARIA ELISAWATI	2119940057	25	Present

TABLE 2. Table of Number of Attendance Days

Row Labels	Numbers of Employees
ADMIN	1325
ICT PROJECT	986
JASA KEPROYEKAN	257
PJOP	970
Grand Total	3538

TABLE 3. Example table of attendance results

Gender	Attendance Status	Working		Machine
		Status	Status	
F	PRESENT	Active	Valid	JZOE#392#AMANTAASSI
M	PRESENT	Active	Valid	DDXN#372#DIAN
M	PRESENT	Active	Valid	PPCU#271#KUSWANTO
M	PRESENT	Active	Valid	UAGO#281#MOHREZAFAT
M	PRESENT	Active	Valid	TUBI#304#AFRIANGGAP
F	PRESENT	Active	Valid	UQCE#169#KARISMANIN
M	PRESENT	Active	Valid	MEQS#160#SUTARNO
F	PRESENT	Active	Valid	Winda Megasari
F	PRESENT	Active	Valid	LICD#15#GIWANGKATO
M	PRESENT	Active	Valid	TEPA#5#HABIBURAHM

5. Data Preparation

This step adjusts the attributes of the data table that will be in the process using the Rapid Miner Framework 8.0. In the previous year, it has been described several tables that will be processed in this study, this time the author created a new table that combines the tables into one table. Data in table 4 below is data from tables that will be inserted into the process of data mining that will be used in Rapid Miner 8.0 Framework. Some sample calculations using several methods in this project is possible to produce different results, so the accuracy in analyzing the results of the prediction is very certain of the results that will be produced by the researchers.

In the work location table above, the researchers reprocessed the data by converting redundant values or values that were too diverse

into smaller groups to facilitate the formation of the model. For example, the value with worker domicile is categorized by distance. Table 5 is a table of categorization of attributes.

6. Modeling

Dataset form previous stage is used as input for the classification algorithm, which is made as a training dataset. In the insert dataset in the rapid miner, This research will be used three types of algorithms to classify C 4.5, random Tree, and Random Forest. The process design used following the caption Figure 1.

7. Evaluation

The usability value can be determine with 10-fold cross-validation model. the process design used following the caption Figure 2.

TABLE 4. Work Location Tables

Project	Work Location	Distance	Mode of Transport
ICT PROJECT	Jakarta Ketapang	23	Public Transport
ADMIN	JAKARTA	32	Public Transport
PJOP	Bitung	29	Private Vehicle
PJOP	SURABAYA	10	Private Vehicle
PJOP	MEDAN	11	Private Vehicle
ICT PROJECT	Jakarta Ketapang	8	Public Transport
ICT PROJECT	Jakarta Ketapang	18	Private Vehicle
ICT PROJECT	Jakarta Ketapang	27	Private Vehicle
ICT PROJECT	Jakarta Manhattan	20	Public Transport
PJOP	JAKARTA	51	Public Transport

TABLE 5. Label Category

Time In	Status
< 08.00	Punctual
> 08.00	Late

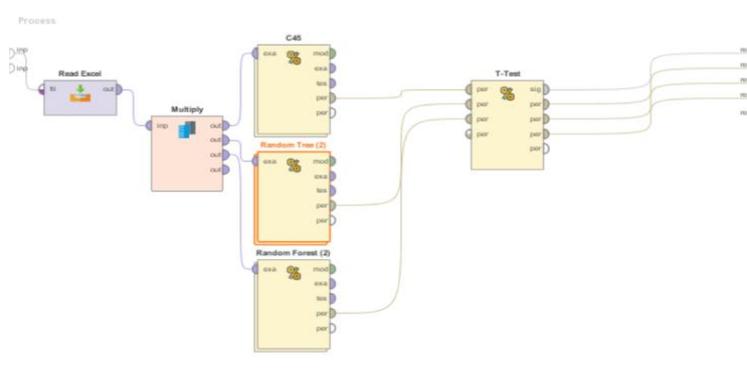


FIGURE 1. Design process comparison of 3 algorithms (C4.5, Random Tree and Random Forest)

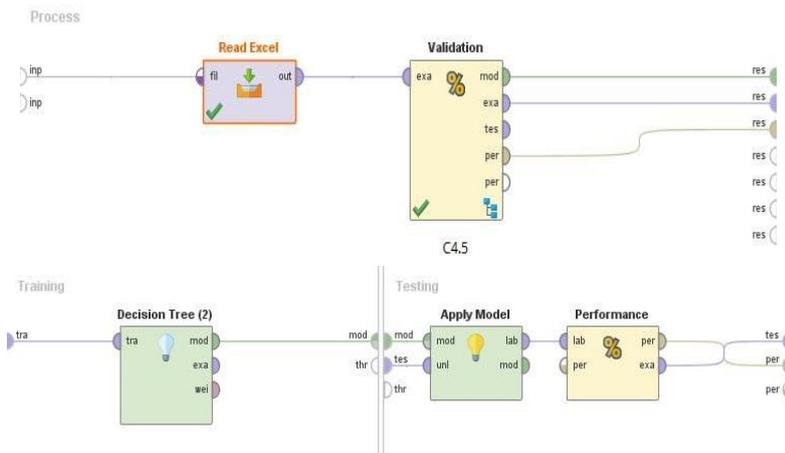


FIGURE 2. Testing of the C4.5 Algorithm Model

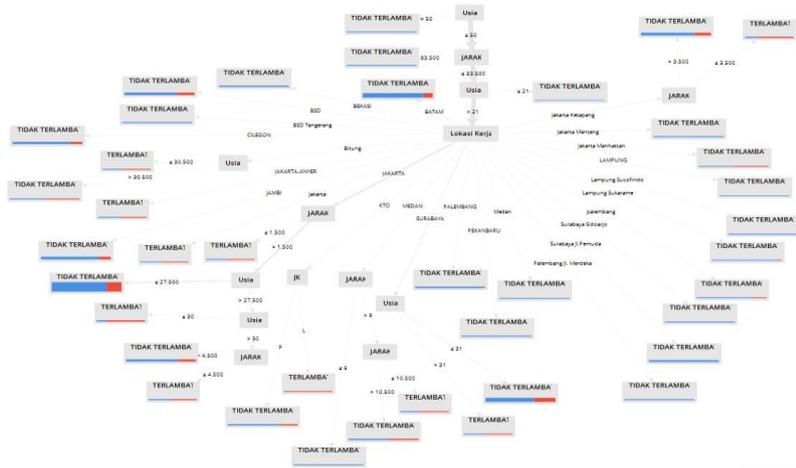


FIGURE 3. C4.5 Algorithm Decision Tree Model

After testing with the model above, the results formed will look like in Figure 3.

There is a performance vector from the C4.5 algorithm :

Performance Vector:

accuracy: 79.37% +/- 1.88% (micro average: 79.37%)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2501 559

LATE: 171 307

precision: 63.92% +/- 5.95%

(micro average: 64.23%) (positive class: LATE)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2501 559

LATE: 171 307

recall: 35.46% +/- 7.32%

(micro average: 35.45%) (positive class: LATE)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2501 559

LATE: 171 307

AUC (optimistic): 0.783 +/- 0.029

(micro average: 0.783) (positive class: LATE)

AUC: 0.772 +/- 0.030

(micro average: 0.772) (positive class: LATE)

AUC (pessimistic): 0.760 +/- 0.032

(micro average: 0.760) (positive class: LATE)

Based on the results of the test using Rapid Miner then obtained decision tree and rule-rule model as shown above with attribute of age value, entry, JK, calculated work, Project, Job Site, domicile, distance, transportation, label. However, in the decision tree above not all attributes appear because the attribute has a small gain value.

From the results of the test model that has been done is to get the accurate results and the Area Under Curve (AUC). The result of the ROC chart with the value of Under Curve Area (AUC) of 0,646 with the performance accuracy is Fair Classification, as seen in Figure 4. From the results of testing with the C4.5 Algorithm model, the accuracy value is 79.37%, as shown in Table 6.

The True Positive Number (TP) is 2501 records classified as LATE and False Negative (FN) as many as 171 records classified as LATE. 559 records for Positive False are classified as

LATE selected, and 307 records for True Negative are classified as NOT LATE selected. Similar to the evaluation process in the C4.5 algorithm above, the Random Forest algorithm is also evaluated as below Figure 5.

The From the results of the model testing, the next process is to get the results of accuracy and Under Curve Area (AUC) and the results of the ROC graph with the value of Area Under Curve (AUC) of 0.807 with accuracy performance, which is Good Classification, as seen in Figure 6.

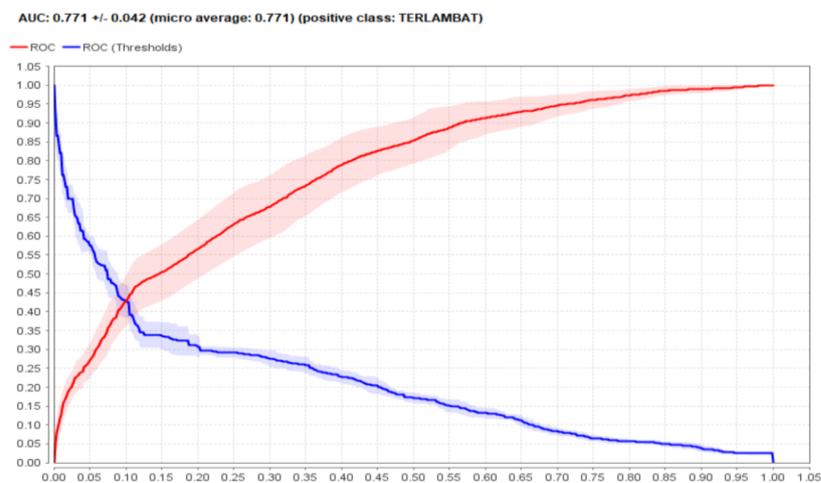


FIGURE 4 AUC value in C4.5 Algorithm

TABLE 6. Model for C4.5 Algorithm
Accuracy: 79.37% +/- 1.88% (micro average: 79.37%)

	true NOT LATE	true LATE	Class Presicion
NOT LATE	2501	559	81.73%
LATE	171	307	64.23%
Class Recall	93.60%	35.45%	

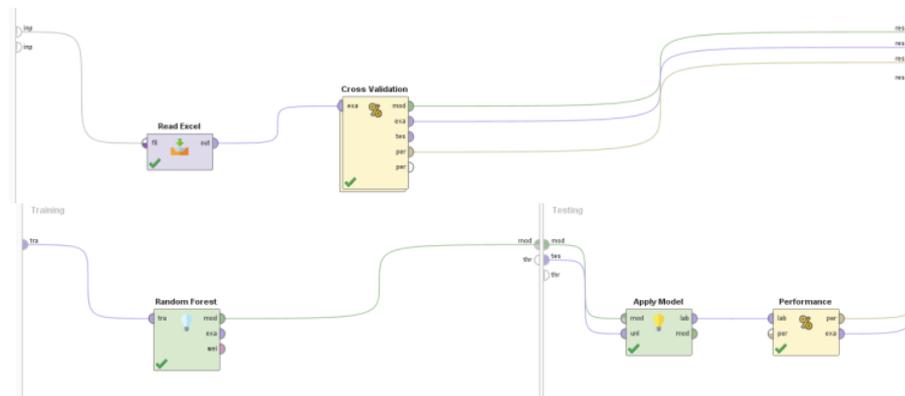


Figure .5 Testing the Random Forest Model

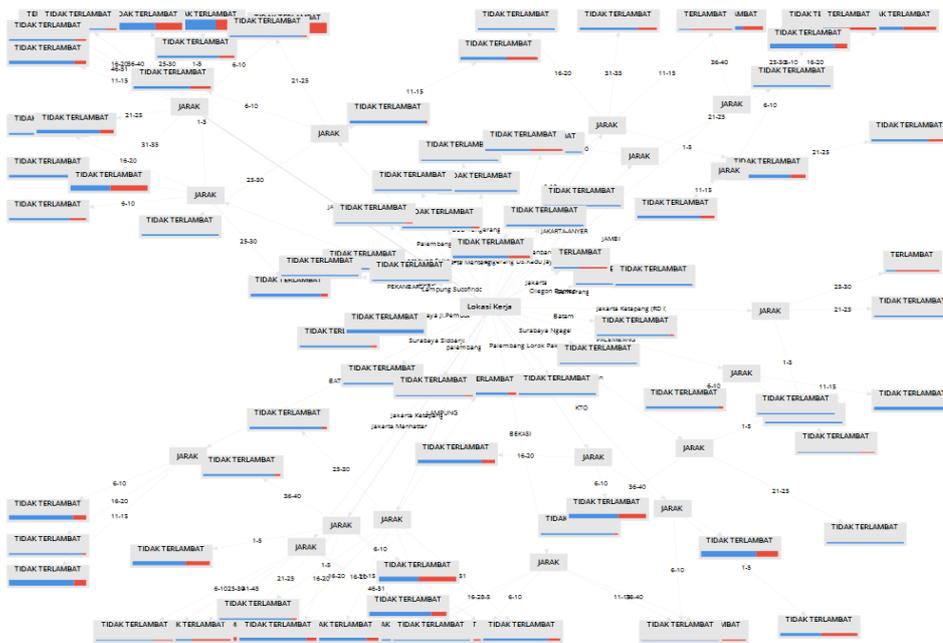


FIGURE 6 Algorithm Random Forest Model

PerformanceVector:

accuracy: 79.11% +/- 1.49% (micro average: 79.11%)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2537 604

LATE: 135 262

precision: 65.94% +/- 6.36% (micro average: 65.99%) (positive class: LATE)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2537 604

LATE: 135 262

recall: 30.26% +/- 4.24% (micro average: 30.25%) (positive class: LATE)

ConfusionMatrix:

True: NOT LATE TOO LATE

NOT TOO LATE: 2537 604

LATE: 135 262

AUC (optimistic): 0.806 +/- 0.033 (micro average: 0.806) (positive class: LATE)

AUC: 0.804 +/- 0.033 (micro average: 0.804)

(positive class: LATE)

AUC (pessimistic): 0.802 +/- 0.033 (micro average: 0.802) (positive class: LATE)

Based on the results of the test using Rapid Miner then obtained decision tree and rule-rule model as shown above with attribute of age value, entry, JK, calculated work, Project, Job Site, domicile, distance, transportation, label. However, in the decision tree above not all attributes appear because the attribute has a small gain value.

From the results of the test model that has been done is to get the accurate results and the Area Under Curve (AUC). The result of the ROC chart with the value of Under Curve Area (AUC) of 0,804 with the performance accuracy is Fair Classification, as seen in Figure 7. From the results of testing done by the Random Forest Algorithm model, the accuracy value is 78.58%, as shown in Table 7.

The True Positive Number (TP) is 2523 records classified as NOT LATE and False Negative (FN) as many as 149 records classified as LATE, while 581 records for Positive False are classified as LATE selected, and 285 records for True Negative are classified as NOT LATE selected.

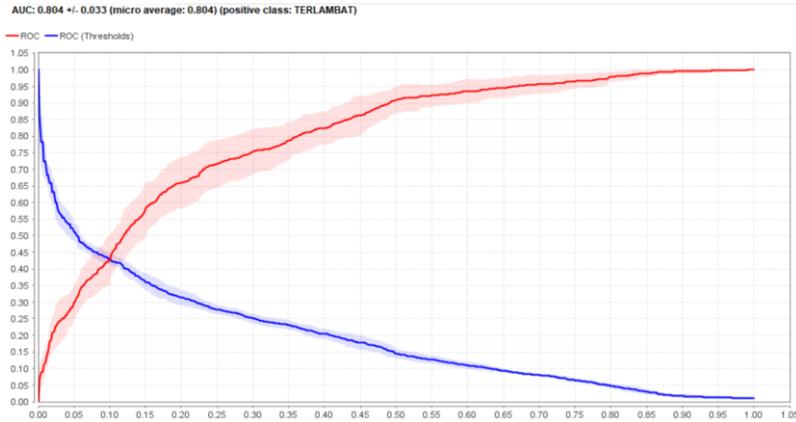


FIGURE 7. AUC value in the Random Forest Algorithm

TABLE 7. Models for Random Forest Algorithms
Accuracy: 78.58% +/- 2.03% (micro average: 78.58%)

	true NOT LATE	true LATE	Class Precision
NOT LATE	2523	581	81.28%
LATE	149	285	65.67%
Class Recall	94.42%	32.91%	

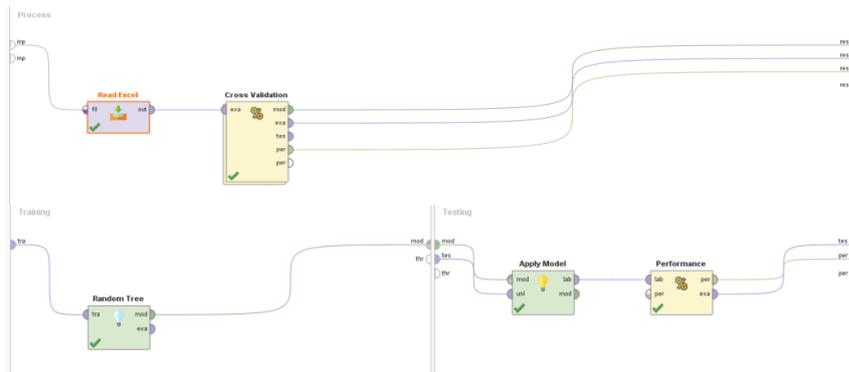


FIGURE 4.9 Random Tree Model Testing

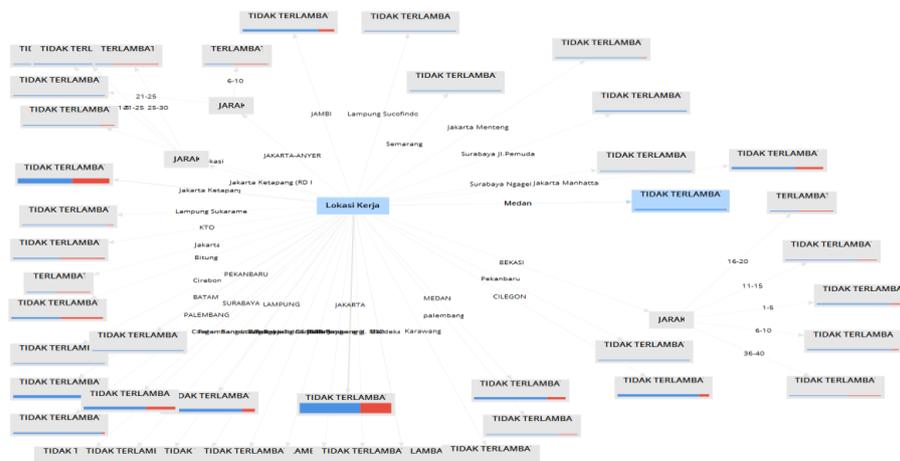


FIGURE 10. Decision Tree Model of Random Tree Algorithm

Performance Vector:

accuracy: 76.26% +/- 1.35% (micro average: 76.26%)

Confusion Matrix:

True: NOT LATE	TOO LATE
NOT LATE: 2642	810
TOO LATE: 30	56

precision: 67.59% (positive class: TOO LATE)

Confusion Matrix:

True: NOT LATE	TOO LATE
NOT LATE: 2642	810
TOO LATE: 30	56

recall: 11.33% +/- 15.74% (micro average: 11.32%) (positive class: TOO LATE)

Confusion Matrix:

True: NOT LATE	TOO LATE
NOT LATE: 2642	810
TOO LATE: 30	56

AUC (optimistic): 0.896 +/- 0.112 (micro average: 0.805) (positive class: TOO LATE)

AUC: 0.805 +/- 0.141 (micro average: 0.805) (positive class: TOO LATE)

AUC (pessimistic): 0.805 +/- 0.087 (micro average: 0.805) (positive class: TOO LATE)

Based on the results of testing using Rapid Miner, a decision tree and model rules are obtained with attributes of Age, Entry, Gender,

Calculated work, Project, Job Location, Domicile, Distance, Transportation, label. However, in the decision tree above, not all attributes appear because the attribute has a small gain value.

The model testing results lead to get the results of accuracy and Under Curve Area (AUC) to and to get the results of the ROC chart with the value of Area Under Curve (AUC) of 0.805 with accuracy performance, namely Fair Classification, as seen in Figure 11. Tests that have been done with the Random Tree Algorithm model got an accuracy value of 76.26%, as shown in Table 8.

The True Positive Number (TP) is 2642 records classified as NOT LATE and False Negative (FN) as many as 30 records are classified as LATE. 810 records for Positive False are classified as LATE selected, and 56 records for True Negative are classified as NOT LATE selected.

8. Deployment

This stage is the last stage in the standard modeling in data mining (CRISP-DM). In this stage, the report will be made in the form of writing the results of thesis and journal research from the introduction to conclusions, as well as creating a Graphical User Interface (GUI) so that users who use the results can interact and apply it easily. Based on the results of the analysis of each of the algorithm tests above, the results can be summarized as in Table 9.



Figure 11. AUC value in Random Tree Algorithm

TABLE 8 Models for Random Tree Algorithms
Accuracy: 76.26% +/- 2.41% (micro average: 76.26%)

	true NOT LATE	true LATE	Class presicion
NOT LATE	2642	810	76.54%
LATE	30	56	65.12%
Class Recall	98.88%	6.47%	

TABLE 9 Comparison of Performance Algorithms

	C.45	Random Forest	Random Tree
ACCURACY	79,37%	78.58%	76,26%
AUC	0.763	0.646	0.805

TABLE 10. Results of testing of T-Test C4.5, Random Forest and Tree

A	B	C	D
	0.796 +/- 0.017	0.793 +/- 0.019	0.763 +/- 0.024
0.796 +/- 0.017		0.706	0.002
0.793 +/- 0.019			0.006

The comparison result of the three performance algorithms above, the results of the tests for Random Forest and C.45 are higher than the Random Tree algorithm. The accuracy value for the Random Forest 78.58% and C.45 algorithm models is 79.37% and the accuracy value for the Random Tree algorithm is 76.26% with an accuracy difference of 3.11%. Seeing the results of calculations in Table 4.9 above and by applying the Area Under Curve (AUC) accuracy performance classification, for C4.5 algorithm with AUC of (0.763) and Good Classification for Random Tree algorithm with an AUC value of (0.805) while for the smallest classification for the Random Forest algorithm with an AUC value of (0.646). After testing with 10 Fold Cross Validation in this study, it was tested again using T-Test to test the truth and falseness of the models. In this test, a comparison between three algorithms to get the results of the T-Test statistics calculation algorithms, After testing Comparison will be made between three algorithms to get the results of the T-Test statistic calculations in Table 10.

Based on Table 4.10 above, it can be analyzed that the C4.5, Random Forest and Random Tree algorithms have insignificant differences in values, and have a probability of 0.002 - 0.006 which is 0.706.

CONCLUSION

The research can be concluded : from the results of the tests that have been conducted to produce, the value of accuracy for the factor of delay of workers at PT. Permata Karya Jasa uses the data mining classification algorithm, it can be proved by the results of accuracy and AUC values of each algorithm, for C.45 accuracy = 79.37% and AUC = 0.646, Random Forest accuracy = 78.58% and AUC = 0.807 while for the Random Tree algorithm accuracy = 76.26% and AUC = 0.610. As for the advice given by the author to refine the results of this study is the application of the algorithm C.45 and the Random Forest is expected to be able to provide solutions for workers placement PT. Permata Karya Services in order to be more disciplined in carrying out their duties as workers, especially in carrying out attendance at work in accordance with company regulations and work agreements. The following suggestion is to use another algorithms such as Particle Swarm Optimization (PSO) or another algorithm likes Genetic Algorithm (GA) as well as other algorithms to increase the level of accuracy especially in providing further analysis for

more specific results in order to foster employee attendance.

REFERENCES

- Oded Maimon, Lior Rokach, (2010), 2nd Edition Data Mining And Knowledge Discovery Handbook.
- Prasetyo, Eko. (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Yogyakarta: Andi Offset.
- Kalmegh, S.R. (2015). Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data. *International Journal of Emerging Technology and Advanced Engineering*, 5(1), 507–517.
- Pfahring, B. (2011). Semi-random model tree ensembles: An effective and scalable regression method. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-25832-9_24
- Breiman, L., & Cutler, A. (2007). Random forests — Classification description: Random forests.
- Aprilla Dennis. (2013). Belajar Data Mining dengan RapidMiner. *Innovation and Knowledge Management in Business Globalization: Theory & Practice, Vols 1 and 2*. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Gorunescu, F. (2011). Data mining: Concepts, models and techniques. *Intelligent Systems Reference Library*. <https://doi.org/10.1007/978-3-642-19721-5>.
- Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan*.
- Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). In *Complementary literature None*.
- North, M. (2012). Data Mining for the Masses. In *Computer*.
- Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. In *Discovering Knowledge in Data: An Introduction to Data Mining*. <https://doi.org/10.1002/0471687545>.
- Hermanto, B., & SN, A. (2017). Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 11(1),43.<https://doi.org/10.22146/ijccs.15946>
- Anggarwal, C.C. (2015). Data Mining: The Textbook. Switzerland: Springer.
- Bahar, (2011). Penentuan Jurusan Sekolah Atas Dengan Algoritma Fuzzy C-Means.
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). San Francisco: Morgan Kaufmann.
- Nguyen, H. K. and Chew, M. T. (2017). RFID-based attendance management system. *IEEE. 2017 2nd Workshop on Recent Trends in Telecommunications Research (RTTR)*, pp. 1-6.
- Nowakowski, S., Ognjanović, I., Grandbastien, M., Jovanovic, J., and Šendelj, R. (2014). Two Recommending Strategies to Enhance Online Presence in Personal Learning Environments. *Springer. Recommender Systems for Technology Enhanced Learning*, pp 227-249.
- Shipway, N. J., Huthwaite, P., Lowe, M. J. S., and Barden, T. J.. (2019). Performance Based Modifications of Random Forest to Perform Automated Defect Detection for Fluorescent Penetrant Inspection. *Springer. Journal of Nondestructive Evaluation, Vol. 38, No. 37*, pp. 1-11.
- Wang, H. (2017). Design for Attendance System with the Direction Identification Based on RFID. *International Conference on Machine Learning and Intelligent Communications. Springer. MLICOM 2017: Machine Learning and Intelligent Communications*, pp 282-290.
- Witten, Ian H., (2011). Data Mining Practical Machine Learning Tools and Techniques.

Younis, M. I., Al-Tameemi, J. F. A., Ismail, W., and Zamli, K. Z.. (2013). Design and Implementation of a Scalable RFID-Based Attendance System with an Intelligent Scheduling Technique. Springer. Wireless Personal Communications, Vol. 71, No. 3, pp. 2161-2179.

AUTHORS :

Riza Fahlapi

Computer Science, Faculty Information and Technology, STMIK Nusa Mandiri, Jakarta, Jalan Kramat Raya No. 18, Senen, Jakarta Pusat.

Email: riza.fahlapi@gmail.com

Hermanto

Computer Science, Faculty Information and Technology, STMIK Nusa Mandiri, Jakarta, Jalan Kramat Raya No. 18, Senen, Jakarta Pusat.

Email: hermanto.hmt@bsi.ac.id

Antonius Yadi Kuntoro

Computer Science, Faculty Information and Technology, STMIK Nusa Mandiri, Jakarta, Jalan Kramat Raya No. 18, Senen, Jakarta Pusat.

Email: antonius.aio@nusamandiri.ac.id

Lasman Effendi

Computer Science, Faculty Information and Technology, STMIK Nusa Mandiri, Jalan Kramat Raya No. 18, Senen, Jakarta Pusat.

Email: lasman.lef@bsi.ac.id

Ridatu Oca Nitra

Computer Science, Faculty Information and Technology, STMIK Nusa Mandiri, Jalan Kramat Raya No. 18, Senen, Jakarta Pusat.

Email: ridatu.rdo@bsi.ac.id