Implementation Of Machine Learning To Determine The Best Employees Using Random Forest Method

Putri Taqwa Prasetyaningrum[1]
Department of Information System
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
putri@mercubuana-yogya.ac.id

Irfan Pratama[2]
Faculty of Information Technology
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
irfanp@mercubuana-yogya.ac.id

Albert Yakobus Chandra[3]
Faculty of Information Technology
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
albertch@mercubuana-yogya.ac.id

***Abstract*— In the world of work the presence of the best employees becomes a benchmark of progress of the company itself. In the determination usually by looking at the performance of the employee e.g. from craft, discipline and also other achievements. The goal is to optimize in decision making to the best employees. Models obtained for employee predictions tested on real data sets provided by IBM analytics, which includes 29 features and about 22005 samples. In this paper we try to build system that predicts employee attribution based on A collection of employee data from kaggle website. We have used four different machines learning algorithms such as KNN (Neighbor K-Nearest), Naïve Bayes, Decision Tree, Random Forest plus two ensemble technique namely stacking and bagging. Results are expressed in terms of classic metrics and algorithms that produce the best result for the available data sets is the Random Forest classifier. It reveals the best withdrawals (0,88) as good as the stacking and bagging method with the same value.**

Keywords—random forest; machine learning; best employees; key performance index

I. INTRODUCTION

The development of digital technology has been increasingly advanced, all forms of information have switched from conventional (physical) form to digital form. It allows us to process these data with certain mechanisms to then generate knowledge that can be used for strategic purposes. in recent years there is a field of science that is very trending in the world of information technology related to other fields including business namely Data Science. Data Science is a field of science that specifically studies data, especially quantitative or numerical data both structured and unstructured so that the data can provide an understanding of existing problems or facts[1].

The many data if analyzed traditionally will not be effective to obtain information or patterns contained in it, a method that is quite capable to perform data analysis is machine learning, which is a method of data extraction that is a combination of artificial intelligence and computer science or computer science[2]. Process prediction approach has been proposed, implementing different data processing schemes and prediction algorithms[3]. Specifically, the processing of data conceptually can be done using the Data Mining process. One of the processes of Data Mining that can provide knowledge about the interrelationship between data variables is Random Forest. Random Forest Regression Algorithms are used to match the data and human resource from the test data[4]. Today Big Data is becoming a relevant issue for the world, interest in Data Science and Machine Learning is growing. There are typical types of Data Analytics ranging from Descriptive Analytics that evolves into something further developed, as Predictive Analytics. Predictive analysis allows investigators to work on authentic and up-to-date data to help predict future environmental possibilities. These predictive insights promote much better decision making and better results. The use of predictive analytics is vast, empowering organizations to improve quite a lot of aspects of their business fortifying their decision-making power, one of which is human resources[5].

In recent years, the company has increasingly paid attention to human resources (HR), such as the quality of and skills represent real growth factors and competitive advantages for the company[6]. This encourages that HR data contains a lot of noise and errors. Therefore building accurate analytics models is challenging for HR[7]. If the data in HR available more, the extreme gradient enhancement is recommended to be used as the most reliable algorithm. this requires minimal data preprocessing, having predictive power, and rank important features automatically and reliably[8]. A common problem with data is that it's missing data. Most real data sets have been lost Value. Inhibiting lost values makes analysis easier by creating a complete data set because it eliminates complex pattern handling issues of missing security[9]. Missing data pose several problems for the data analysis[10]. In previous studies the approach of handling lost data so that minimize the damage, underlying assumptions and possible costs and its benefits[11].

K-NN Imputation is a method to estimate a missing values that occur on a dataset which usually a classification purposed dataset based on the neighboring pattern, and said that KNN Impute provide more robust and sensitive method for missing values imputation[12]. The results of previous studies, with Random forest model to predict employees with these 10 features through random forests produces more accurate and precise, with and 89% accuracy and 72% precision[13]. From the results of previous research analysis of predictions employee delay factor using three algorithms, namely the accuracy of the C.45 = 79.37% and AUC value = 0.646, Random Forest Algorithm Accuracy = 78.58% and AUC value

= 0.807 while for random tree algorithm accuracy = 76.26%and AUC value = 0.610[14]. By using Random forest can identify whether run broiler breeders lay eggs or not on a certain day during the egg-laying period with an accuracy of about 85%[15]. On ensemble method based architecture using random forest research importance to predict employee's turn over has achieved the highest accuracy of 99.4%[16].

With enough data, Machine learning can be used to predict the best employees. This research shows that the machine learning, random forests, can improve accuracy and precision of predictions, and points to variables and behavior indicators that have been found to have indicators correlation with employees. Based on the problems previously raised, researchers are interested in conducting research related to employee data in Bank Rakyat Indonesia. The use of random forest algorithms is included in the part of machine learning which is a computer learning process from the data will be processed in such a way that it can be recognized the characteristics or pattern models of existing data in the hope of providing knowledge to be able to improve the effectiveness of the best employee assessment

II. METHOD

A. *Data*

The data used in this study is a public dataset retrieved from Kaggle data repository with "people analytics" query. The dataset is about employees track record on a company which used to determine whether the employee is on their best performance or not. The data consist of 29 data attributes and 1 class label with 22005 instances. The information is various, from the demographic information such as, age, gender, and marital status into more professional characteristics such as, job duration on current level, the job level, employee type, person level, annual and sick leaves, and the achievement meter which shows each employee personal achievement on the job. The sample dataset can be seen in Table 1.

TABLE I. DATASET SAMPLES

Job_level	*Person_level*	*Gender*	*...*	*Best Performance*
JG04	1.17	Male	…	0
JG04	1.83	Male	…	1
JG03	0.75	Male	…	0
JG03	0	Male		0
JG04	1.17	Male		0
…	…	…	…	…
JG04	1.5	Male	…	0
JG04	1.75	Female	…	1
JG04	1.42	Female	…	0
JG04	1.5	Male	…	0

From Table I can be inferred that the data characteristics is mixed between categorical(nominal) data and numerical data. In the following step, the data will be undergoing a preprocessing step to ensure the readiness of the data to be processed further.

B. *Research Design*

The following stages is the research design of this study.

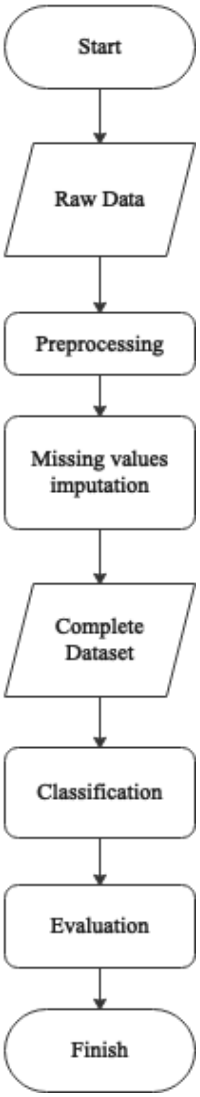

Figure 1 Research Design

C. *Preprocessing*

Preprocessing is a series of process to make sure the dataset is "clean" enough to be processed. The raw dataset may contain inconsistent data, noise, incorrect data format, duplicate or redundant data, or missing values. therefore, those situations can degrade the effectiveness and the trustworthiness of a data mining result because the dataset is not conditioned to be ready for further processing. To make sure the data is free from the unwanted condition, the preprocess stages is necessary to be carried on. Missing values is also problems that handled in the preprocessing stages, but because it needed the data is clean from noise and such disruption then the missing values imputation steps is separated from the other preprocessing steps and will be done later. The condition of the dataset which needed to be address can be seen in Table II.

TABLE II. NOISES IN THE DATASET SAMPLES

Education_Level	*GPA*	*Achievement_target_1*	*achievement_ target_2*
level_4	40	?	?
level_3	?	achiev_100%-150%	achiev_< 50%
?	?	achiev_50%-100%	achiev_< 50%
level_4	?	?	?
level_4	?	achiev_50%-100%	Tercapai_< 50%
level_4	?	?	?

As shown in Table II, the missing values numbers from that small samples are already numerous and there are some irregular data such as GPA which have values of 40 which is doesn't make sense because the maximum value of GPA is 5 internationally yet in Indonesia the GPA scores is 4 at maximum value. There is also inconsistent data format on the last 2 rows of *achievement_target_2* attribute.

The outlier data of the samples will be deleted and let it become a missing-values, the duplicates will be filtered so there is no duplication anymore, the inconsistent data will be corrected into the desirable and appropriate format. The outlier detection is done using box-plot to see the boundaries of the data distribution. The box-plot samples of several attributes can be seen in Figure 2 and Figure 3.

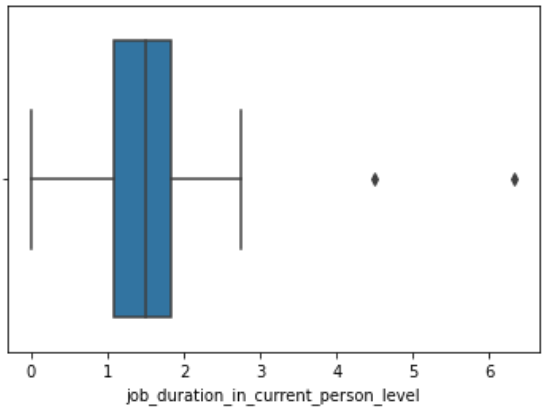

Figure 2 Box-Plot on Job_duration_in_current_person_level attribute

Can be seen in Figure 2, that the attribute has several outliers in it, so the dataset will be filtered by the maximum number of the upper limit of the plot which less than 3.

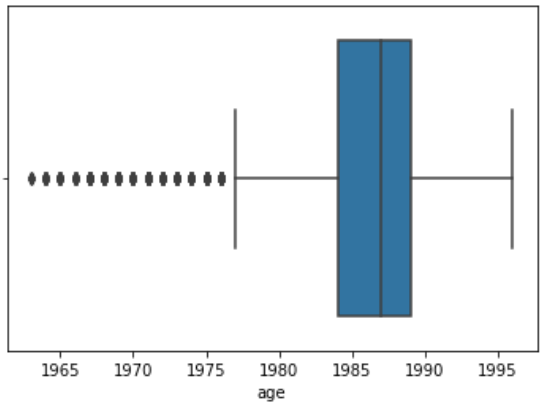

Figure 3 box-plot on age attribute

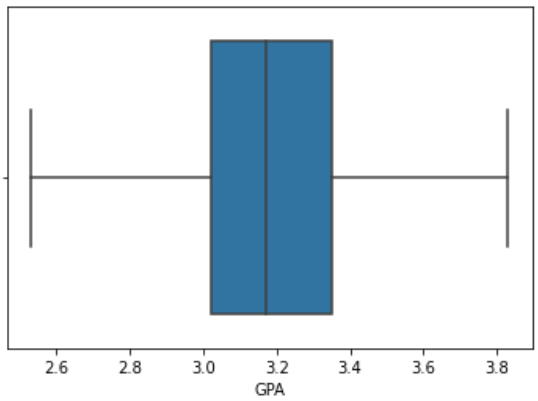

Figure 4 Box Plot on GPA

On Figure 3 can be seen more outlier values on age attribute, the process is same for all outlier values detected which is deleting all the outlier and let it become a missing - values. After those noise overcame, the missing values imputation step takes place. It will estimate the original missing values from the raw data and the deleted values due to outlier. The outlier is treated by such process is due to the value of the overall instance is too precious, the number of instances which is also the number of information is maintained. By just deleting the outlier value and not the instances will retain the other valuable information for the data mining process.

D. Missing Values Imputation

It is described in the previous section that missing values imputation will be done after the data is "cleaned". The missing values imputation method used in this study is a Random Forest based imputation or called missForest imputation and the technical implementation of missForest imputation method is carried on Python. missForest is relatively new approach to handle missing values on a dataset. After the missing values imputation step done, then the data is ready to be processed further into the classification step.

E. Classification

After the data cleansing steps or so-called preprocessing steps is done. The main stage of the study will be carried on, the classification method used in this study is as follows:

Decision Tree, k-NN, Random Forest, Stacking and Bagging methods.

Decision tree is a commonly used classification method. This method is already established and included in almost all data mining tools. The decision tree changes the tabular dataset into a tree-like model to represent the rule conditions of decision[17]. The main requirement of decision tree method is a labeled dataset, or can be said that decision tree is a supervised learning method.

Random Forest is a classifier that contain a bunch of tree-structured classifiers that are identically independent and distributed random vectors. Each tree casts a unit vote for most popular class at input[18].

Nearest Neighbor classification, also known as K-nearest neighbors (KNN), is based on the idea that the nearest patterns to a target pattern x, for which we seek the label, deliver useful label information[19]. The K-NN method is widely used in data mining field as one of the simplest methods to do a classification.

Stacking is an ensemble method that stack individual learners into a single powerful learner. The general principle of stacking is as follows: given *d* different learning algorithms, evaluate each of them on the predictor matrix *X*, given outcome vector *y* in a *k*-fold cross-validation. Save the out-of-fold predictions and combine them to a new data matrix *Z*. *Z* now has *d* columns and the same number of rows as *X*. Then, estimate a weighted scheme for each column of *Z* to combine to a final prediction[20].

Bagging predictor is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting class[21].

F. Evaluation

For each classification method, the measurement metric on how good the classification result is carried on by the k-fold cross validation. The evaluation method is said to be fair and square method to measure the classifiers performance due to the nature of its method that split the data into parts and do the training-testing process using those parts. The first fold is treated as a validation set and the rest is fitted toward the method. So, the number parts that fitted toward the method is k-1 fold[22].

III. RESULT AND DISCUSSION

The initial dataset is preprocessed and treated using K-NN Imputation method to estimate any missing values on the dataset. K-NN Imputation method can work on both numerical and categorical data. Yet in this study the missing values processed onto an encoded value of the dataset. After the data is ready to be processed further, then the classification methods take place. The illustration of the missing values imputation result can be seen int Table III and Table IV.

TABLE III. PRE IMPUTATION DATASET

GPA	*Year_raduated*	*Avg_Achievement_%1*
NaN	NaN	NaN
NaN	NaN	35.3433
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN

TABLE IV. IMPUTED DATASET

GPA	*Year_raduated*	*Avg_Achievement_%1*
3.246	2012	25.453
3.312	2013	35.3433
3.238	2015	28.753
3.402	2014	30.1647
3.082	2012	30.238

From Table IV can be seen that the process of the missing values imputation is done as no more missing values in the samples. Therefore, the next stage of data mining can be started.

The classification methods used in this study are K-NN, Decision tree, Random Forest, plus two ensemble techniques such as Bagging and Stacking. Then the classification result evaluated using 10-fold cross validation to make sure the results are fair and because the dataset number is not that big. The K-NN method parameter used in this study is k = 5, and produced accuracy of 0.86 with deviation 0.0. Decision Tree implemented and produced 0.87 accuracy with 0.02 deviation.

The accuracy of the single classifier method can be seen in Table III below.

TABLE V. SINGLE CLASSIFIER RESULT

No	Methods	Accuracy
1	K-NN	86%
2	Decision Tree	87%
3	Random Forest	88%

As shown in Table IV among the single classifiers, Random forest produced biggest accuracy with 88%. To know more of the potential model for the dataset, some of the ensemble techniques are implemented. Bagging K-NN and the Stacking of Random Forest and Logistic Regression are used.

The accuracy comparison of all classification methods can be seen in Table V below.

TABLE VI. CLASSIFICATION RESULT OF ALL USED CLASSIFIER

No	Methods	Accuracy
1	Decision Tree	87%
2	Random Forest	88%
3	K-NN	86%
4	Bagging	88%
5	Stacking	88%

From Table V can be inferred that there is no accuracy improvement from the Bagging and Stacking method toward the random Forest method. The situation caused by several possible problem, the first is the class distribution of the dataset that is not yet to be measured. If it is in an imbalanced condition, surely the situation should be treated beforehand. Because imbalanced dataset is already become a specific problem that may affect the classification result.

The assumption of this situation is either the datamining methods is just do not fit for the dataset, or it is the dataset that has the bad quality in terms of statistical values such as normality, or correlation, or it is because the nature of the dataset that have multi class in it. In terms of classification method results, stacking technique produced lowest accuracy score while individual learner such as Random Forest and Decision Tree produced better result. Even though, the score is still close to each other and may be have no significant difference.

The second possible problem that may affect the classification's result is the number of the dataset. The 22.000 dataset may not enough to build a good model for the class.

IV. CONLUSSION

This research is emphasizing on the discovery of the classification method that can do a great job for the dataset. Several classification methods have been tried out and produced pretty much different result. The best result carried out by Random Forest, Bagging and Stacking method with 88% accuracy score.

The future work of this research is to check out the class distribution to measure the potential imbalanced dataset situation and overcome if it is any of that situation occurred.

REFERENCES

[1] V. Dhar, "Data Science and Prediction," *Commun. ACM*, vol. 56, no. 12, pp. 64–73, 2013, doi: 10.1145/2500499.

[2] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artif. Intell. Rev.*, vol. 50, no. 2, pp. 283–339, 2018, doi: 10.1007/s10462-017-9545-7.

[3] D. A. Neu, J. Lahann, and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," *Artif. Intell. Rev.*, no. 0123456789, 2021, doi: 10.1007/s10462-021-09960-8.

[4] S. S. Bashar, M. S. Miah, A. H. M. Z. Karim, and M. A. Al Mahmud, "Extraction of Heart Rate from PPG Signal: A Machine Learning Approach using Decision Tree Regression Algorithm," *2019 4th Int. Conf. Electr. Inf. Commun. Technol. EICT 2019*, no. February, pp. 1–5, 2019, doi: 10.1109/EICT48899.2019.9068845.

[5] K. Nugroho, "Javanese Gender Speech Recognition Based on Machine Learning Using Random Forest and Neural Network," no. February, 2021, doi: 10.24167/Sisforma.

[6] F. Fallucchi, M. Coladangelo, and R. Giuliano, "Predicting Employee Attrition Using Machine Learning Techniques," pp. 1–17, doi: 10.3390/computers9040086.

[7] J. Vasa and K. Masrani, "Foreseeing Employee Attritions using Diverse Data Mining Strategies," no. 3, pp. 620–626, 2019, doi: 10.35940/ijrte.B2406.098319.

[8] Y. Zhao, M. K. Hryniewicki, and F. Cheng, *Employee Turnover Prediction with Machine Learning : A Reliable Approach*, vol. 1. Springer International Publishing, 2019.

[9] G. Chhabra, V. Vashisht, and J. Ranjan, "A review on missing data value estimation using imputation algorithm," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 7 Special Issue, pp. 312–318, 2019.

[10] M. Data and I. Methods, "Statistical Minute," vol. 131, no. 5, pp. 1419–1420, 2020.

[11] S. Gorard, "Handling missing data in numeric analyses," *Int. J. Soc. Res. Methodol.*, vol. 23, no. 6, pp. 651–660, 2020, doi: 10.1080/13645579.2020.1729974.

[12] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, 2012, doi: 10.1016/j.jss.2012.05.073.

[13] A. Foley and A. E. Foley, "Using Machine Learning to Predict Employee Resignation in the Swedish Armed Forces Using Machine Learning to Predict Employee Resignation in the Swedish Armed Forces," 2019.

[14] R. Fahlapi, H. Hermanto, A. Y. Kuntoro, L. Effendi, R. O. Nitra, and S. Nurlela, "Prediction of Employee Attendance Factors Using C4.5 Algorithm, Random Tree, Random Forest," *Semesta Tek.*, vol. 23, no. 1, pp. 39–53, 2020, doi: 10.18196/st.231254.

[15] J. You, S. A. S. van der Klein, E. Lou, and M. J. Zuidhof, "Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision feeding system," *Comput. Electron. Agric.*, vol. 175, no. February, p. 105526, 2020, doi: 10.1016/j.compag.2020.105526.

[16] M. A. Hossen, E. Hossain, A. K. Z. Ishwar, and F. Siddika, "Ensemble method based architecture using random forest importance to predict employee's turn over," *J. Phys. Conf. Ser.*, vol. 1755, no. 1, 2021, doi: 10.1088/1742-6596/1755/1/012039.

[17] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[18] L. Breiman, "Random Forests," *Int. J. Adv. Comput. Sci. Appl.*, 2001.

[19] O. Kramer, "K-Nearest Neighbors," in *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23.

[20] C. F. Kurz, W. Maier, and C. Rink, "A greedy stacking algorithm for model ensembling and domain weighting," *BMC Res. Notes*,

vol. 13, no. 1, p. 70, 2020, doi: 10.1186/s13104-020-4931-7.

[21] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.

[22] G. Casella, S. Fienberg, and I. Olkin, "An Introduction to Statistical Learning," *Springer Texts Stat.*, p. 618, 2013.