Classification of Employee Competency Assessment Using Naïve Bayes and K-Nearest Neighbor (KNN) Algorithms

Kurnia Gusti Ayu [1], Dwi Wulandari Sari [1], Ida Farida [2], Djoko Harsono [3], Ratih Widayanti Kosaman [4], Ibnu Hadi Sumitro [3], Fauzi Nur Iman [2], Mansuri [5], Emil R. Kaburuan [2,*], and Malinda Alya Fitri [2]

[1] Information System Department, Computer Science Faculty, Mercu Buana University, Jakarta, Indonesia
[2] Informatics Engineering Department, Computer Science Faculty, Mercu Buana University, Jakarta, Indonesia
[3] Information System Department, Computer Science Faculty, Borobudur University, Jakarta, Indonesia
[4] Informatics Management Department, Computer Science Faculty, Borobudur University, Jakarta, Indonesia
[5] Computer System Department, Computer Science Faculty, Borobudur University, Jakarta, Indonesia
Email: kurnia.gusti@mercubuana.ac.id (K.G.A.); dwi.wulandari@mercubuana.ac.id (D.W.S.); dae.farida@mercubuana.ac.id (I.F.); djokoharsono@borobudur.ac.id (D.H.); ratihkosaman@borobudur.ac.id (R.W.K.); ibnusumitro@borobudur.ac.id (I.H.S.); fauzi@mercubuana.ac.id (F.N.I.); mansuri@borobudur.ac.id (M.); emil.kaburuan@mercubuana.ac.id (E.R.K.); malindaalya26@gmail.com (M.A.F.)
*Corresponding author

***Abstract*—Employees are one of the most important resources for every company. The company will run well if it has good employees. One way to find out whether the employees' is worthy of working or not in a company is to conduct an employees' competency assessment. However, many companies sometimes conduct assessments inappropriately because of the many parameters that need to be considered. For some companies that have thousands of employees, of course, assessing employees' competence is not an easy thing if it must be done manually. Therefore, this research was made to facilitate the assessment team in assessing employees' competence by predicting classification using the Naïve Bayes and K-Nearest Neighbor (KNN) algorithms. This research is also expected to help companies analyze employee competence and performance. The dataset used in this research is 3,634 employees' data with parameters, assessment scores, and learning journey scores. This research will do a comparison to see which model produces better accuracy. The results obtained show that KNN is superior with an accuracy of 99.45% with a comparison of training data and testing data 70:30, 99.33% with a comparison of 75:25, and 99.44% with a comparison of 80:20. While Naïve Bayes obtained an accuracy of 98.44% with a comparison of training data and testing data of 70:30, 98.45% with a comparison of 75:25, and 98.48% with a comparison of 80:20.**

***Keywords*—assessment, classification method, employee competencies, Naïve Bayes, K-Nearest Neighbor (KNN)**

I. Introduction

For every company, employees are one of the most important resources that play a major role in the company's success. Employees are people who work for an institution (office, company, and so on) by getting a salary (wage) or people who sell services (mind and energy) and get compensation, the amount of which has been determined in advance [1, 2]. With human resources, the company can run well. The success of a company will be realized if employees have good competence. Therefore, companies need to see and assess the performance and competence of their employees.

Competency is an expertise that everyone has in terms of knowledge, skills, and abilities to do something. The competencies possessed by employees can be used as a benchmark to improve the performance of these employees so that they can contribute to the company or organization's success. Many factors can affect the competence of an employee, such as skills possessed or training that has been followed, age, gender, and educational background [3–5]. In general, employees' competency assessment is carried out by the Human Resource team in each company. But sometimes, the selection of the best employees is still inaccurate because of the many assessment criteria or the assessment process that is still done manually coupled with the number of employees who are also not small. This research is intended to help make it easier for companies to assess employees' competence using data mining. This research will test and process the data using the Naïve Bayes and K-Nearest Neighbor (KNN) algorithms.

Naïve Bayes and K-Nearest Neighbor (KNN) are algorithms that classify data. Data classification is an example of supervised machine learning, where the learning process is done by labeling the training data. Naive Bayes is a statistical classification method whose basic concept is Bayes' theorem, which is used to calculate the probability of a class from each group of criteria/features and can determine which class is the

Manuscript received September 8, 2023; revised November 2, 2024; accepted February 1, 2024; published July 29, 2024.

most optimal [4–9]. Naive Bayes is a simple probabilistic-based prediction technique based on applying Bayes' theorem (Bayes' rule) with a strong (naive) independence assumption [7, 10, 11]. In other words, the independent feature model is used in Naive Bayes. Meanwhile, the KNN algorithm works by labeling new data according to the label of the closest data. This means that data that tends to be similar will be close to each other.

In this research, the author created a system to classify employees' competency assessments using assessment data and the learning journey of each employee. The assessment system applies data mining using Naïve Bayes and K-Nearest Neighbor (KNN) algorithms. Both algorithms classify employees based on the accumulation of several assessment elements. The system results will classify employees into four groups: Very Good, Good, Fair, and Less.

The Naïve Bayes algorithm is a classification method that refers to Bayes' theorem or a simplification technique based on the Bayesian algorithm first proposed by Thomas Bayes, an English scientist intended to build classifiers that are essentially conditional probability models. This method utilizes the probability theorem to find the best chance by predicting future probabilities based on previous information [12–16]. The Naive Bayes algorithm uses the Bayes Theorem formula to calculate conditional probabilities. In general, the Naive Bayes algorithm calculation formula is as follows:

$$P(C|X) = \frac{P(X|C)\ P(C)}{P(X)}$$

X: Sample data that has an unknown class (label); C: Hypothesis that X is class (label) data; P(C): Probability of hypothesis C; P(X): Probability of the observed sample data (probability C); P(X|C): Probability based on the conditions in the hypothesis.

The K-Nearest Neighbor (KNN) algorithm calculates a new object based on its (K) nearest neighbors. KNN is a supervised learning algorithm where the result of a new query instance is classified based on most categories in KNN [17–19]. The class that appears the most will be the class of classification results. Several techniques can be used to measure the class distance K of this KNN algorithm, namely Euclidean Distance, Hamming Distance, Manhattan Distance, and Minkowski Distance [20]. In this research, distance measurement is done using the Euclidean distance method.

Euclidean Distance is a distance metric measured between two vectors by calculating the square root of the sum of the squared differences between them. This calculation concept is almost similar to the calculation concept in the Pythagorean theorem. Here is the general formula for Euclidean Distance:

$$d\ (x, y) = \sqrt{\sum_{i=1}^{n}(x_{training}^{i} - y_{testing}^{i})^2}$$

$d(x,y)$: Distance; $x_{training}^{i}$: Data training; $y_{testing}^{i}$: Data testing; i: Variable data; n: Data dimension.

In previous research, Laga [3] states that the K-NN method has the highest accuracy rate of 90.13%, precision rate of 91%, and recall rate of 98.95% in classifying employees' performance with the best ratio of training data and test data, namely 75% and 25%. Meanwhile, the SVM algorithm found an accuracy of 88.85% and a precision level of 89.71%.

Another study by Cholil *et al*. [17] produced an accuracy value of 90.5% using the KNN algorithm to classify scholarship acceptance selection. Meanwhile, Senika *et al*. [9] produced an accuracy value of 91.67% to assess employees' performance using RapidMiner.

With good accuracy in previous studies, the authors used the Naive Bayes and K-Nearest Neighbor (KNN) algorithms to classify employees' performance appraisals in this study. The research objectives are as follows:

- To see the accuracy level and the effectiveness of the Naïve Bayes and K-Nearest Neighbor algorithms in helping the employees' competency assessment process.
- Facilitate the human resource team in conducting employees' competency assessments.
- To determine which algorithm better classifies employees' competence between Naïve Bayes and K-Nearest Neighbor.

II. Research Methods

A. Research Type

The type of research used in this research is quantitative research. Quantitative research is research that focuses on analyzing numerical data (numbers). The data will be processed using data mining so that it can produce new information that is useful and can be used for analysis following the problems in the research.

B. Research Stages

Fig. 1 is the flow of this research.

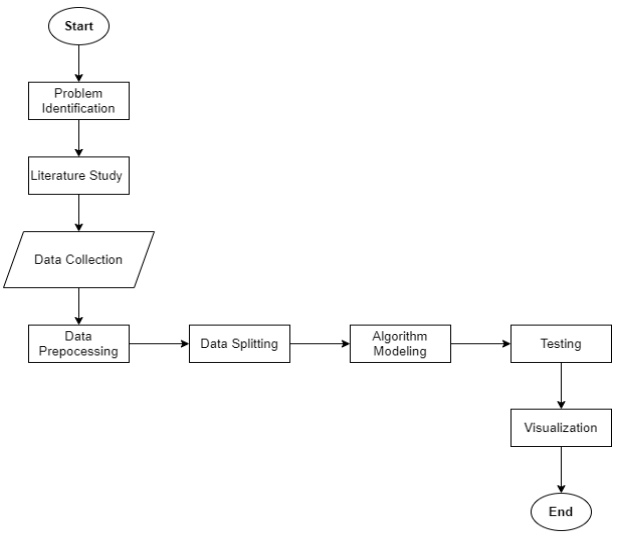

Fig. 1. Research stages.

C. Problem Identification

At this stage, the authors analyze the problems and what needs to be improved from one of the systems in a company that can be resolved using the data mining process so that the author can determine the topic and what data is needed in this study.

D. Literature Study

Literature study is done by searching and collecting all information related to the research topic taken to help the author during the research process.

E. Data Collection

The author's data collection is done by searching for information and data related to the research topic. The data taken by the author are Employee ID, Total assessment score, and total learning journey score of each employee. The total data used in this study is 3,634 employee data.

F. Data Preprocessing

In this research, the preprocessing carried out is data cleansing and data encoding. Cleansing data is done by removing unused columns. Meanwhile, data encoding is done to convert non-numeric labels into numeric labels.

G. Data Splitting

In this research, the dataset will be divided into 2 parts, namely training data and testing data. The division of the dataset into 2 parts like this is done so that the system can learn the data through training data before testing using testing data. In this study, the data will be divided into 3 parts, namely 70:30, 75:25, and 80:20.

H. Algorithms Modelling

After pre-processing and dividing the dataset, the next step is to perform modeling according to the algorithm used, namely naïve bayes and K-Nearest Neighbor (KNN).

For naïve bayes modeling, the system only needs to import the naïve bayes function, namely GaussianNB, followed by entering training data. As for KNN modeling, besides importing the KNN function, we also need to determine the number of "k" and the distance metric method to be used. Distance metric itself is a method used to calculate the distance from the "k" value itself.

I. Testing

In this research, the testing stage is carried out using the confusion matrix method. Confusion matrix is one of the methods used to test the accuracy of machine learning. The confusion matrix works by comparing the predicted data with the actual data. In the confusion matrix, the values displayed are precision, recall, and F1-Score.

J. Visualization

At this stage, the accuracy level of machine learning will be visualized. This visualization aims to help readers better understand the results of this research.

III. Results

A. Dataset

In this study, the dataset used was taken directly from a retail company. The data is obtained from the database team in the Human Capital department. The data used in this study is learning journey data and employee assessments during 2022 which amounted to 3634 data and is divided into 4 labels, namely Less (total score below 61), Fair (total score between 61–75), Good (total score between 76–87), and Very Good (total score above 87). Table I is a sample dataset table that has been preprocessed:

TABLE I. Dataset

Employee ID	Religion	Marital Status	Assessment Final Score	Journey Final Score	Status
201013545	Moslem	1	100.00	90.00	Very Good
202111876	Christian	0	100.00	87.50	Very Good
202112343	Moslem	0	100.00	87.50	Very Good
202200475	Catholic	1	60.00	90.53	Fair
202203689	Budha	0	75.00	90.92	Good
202000684	Christian	1	100.00	92.00	Very Good
202208585	Moslem	1	100.00	88.67	Very Good

B. Data Preprocessing

The first step of data cleansing is to delete unused columns (Table II). In this case, the unused columns are Employee ID, Religion, and Marital Status (Table III).

After removing unused columns, the next step in pre-processing in this study is to divide the dataset into 2 parts, namely X and Y. Part X contains data with the Assessment Final Score and Journey Final Score columns (Table IV). While part Y only contains data for the Status column. This is done to make it easier for the system to learn data patterns for classification (Table V).

TABLE II. Dataset before Deleting Unused Columns

Employee ID	Religion	Marital Status	Assessment Final Score	Journey Final Score	Status
201013545	Moslem	1	100.00	90.00	Very Good
202111876	Chatolic	0	100.00	87.50	Very Good
202112343	Moslem	0	100.00	87.50	Very Good
200900583	Christian	1	100.00	95.20	Very Good
202112012	Moslem	1	100.00	87.26	Very Good

TABLE III. DATASET AFTER DELETING UNUSED COLUMNS

Employee ID	Assessment Final Score	Journey Final Score	Status
201013545	100.00	90.00	Very Good
202111876	100.00	87.50	Very Good
202112343	100.00	87.50	Very Good
202200475	60.00	90.53	Fair
202203689	75.00	90.92	Good
202000684	100.00	92.00	Very Good
202208585	100.00	88.67	Very Good

TABLE IV. DATA X FROM THE DATA SET

Assessment Final Score	Journey Final Score
100.00	90.00
100.00	87.50
100.00	87.50
…	…
100.00	95.00
100.00	92.00
100.00	88.67

TABLE V. DATA Y FROM THE DATA SET

Status
Very Good
Very Good
Very Good
…
Very Good
Very Good
Very Good

C. Algorithms Modeling

The modeling process begins by dividing the data into 2 parts: training data and testing data. The system will use the training data to learn the data pattern. The testing data will be used to test the classification on machine learning. In this study, the dataset is divided into 3 categories, namely 70:30, 75:25, and 80:20. Below is the script for each data test classification. Fig. 2 shows the script used to create 70% of data training and 30% of data testing of algorithms modeling.

```
[17] from sklearn.model_selection import train_test_split
     x_train70, x_test30, y_train70, y_test30 = train_test_split(x, y, test_size=0.3, random_state=1) #Data training 70% Data testing 30%

[18] len(x_train70)
     2543

[19] len(x_test30)
     1091

[20] len(y_train70)
     2543

[21] len(y_test30)
     1091
```

Fig. 2. Data splitting 70:30.

Fig. 3 shows the script used to create 75% of data training and 25% of data testing of algorithms modeling.

```
[ ] from sklearn.model_selection import train_test_split
    x_train75, x_test25, y_train75, y_test25 = train_test_split(x, y, test_size=0.25, random_state=1) #Data training 75% Data testing 25%

[ ] len(x_train75)
    2725

[ ] len(x_test25)
    909

[ ] len(y_train75)
    2725

[ ] len(y_test25)
    909
```

Fig. 3. Data splitting 75:25.

Fig. 4 shows the script used to create 80% of data training and 20% of data testing of algorithms modeling.

```
[ ] from sklearn.model_selection import train_test_split
    x_train80, x_test20, y_train80, y_test20 = train_test_split(x, y, test_size=0.2, random_state=1) #Data training 80% Data testing 20%

[ ] len(x_train80)
    2907

[ ] len(x_test20)
    727

[ ] len(y_train80)
    2907

[ ] len(y_test20)
    727
```

Fig. 4. Data splitting 80:20.

1) Naïve Bayes

The Naïve Bayes algorithm is modeled by importing the Gaussian NB library available on Google Collabs. The following is the modeling process for Naïve Bayes.

Fig. 5 shows the Gaussian NB script used to process 70% of data training and 30% of data testing of algorithms modeling.

```
[ ] from sklearn.naive_bayes import GaussianNB
    classifier70 = GaussianNB()
    data_predict = classifier70.fit(x_train70, y_train70)
```

Fig. 5. Naive Bayes Modeling in 70:30.

Fig. 6 shows the Gaussian NB script used to process 75% of data training and 25% of data testing of algorithms modeling.

```
[ ] from sklearn.naive_bayes import GaussianNB
    classifier75 = GaussianNB()
    data_predict = classifier75.fit(x_train75, y_train75)
```

Fig. 6. Naive Bayes Modeling in 75:25.

Fig. 7 shows the Gaussian NB script used to process 80% of data training and 20% of data testing of algorithms modeling.

```
[ ] from sklearn.naive_bayes import GaussianNB
    classifier80 = GaussianNB()
    data_predict = classifier.fit(x_train80, y_train80)
```

Fig. 7. Naive Bayes Modeling in 80:20.

2) K-Nearest Neighbor (KNN)

For modeling with the KNN algorithm, we need to determine the value of k and the type of distance metric to calculate the distance from the k value. In this research, the distance metric used is Euclidean, which calculates the distance between 2 points by calculating the square root of the sum of the square differences between the two. This Euclidean metric calculation has the same concept as the calculation of the Pythagorean theorem. As for the value of k itself, the author determines that k is 1, 3, and 5. The following is the modeling process of the K-Nearest Neighbor algorithm.

Fig. 8 shows the KNN script used to process 70% of data training and 20% of data testing of algorithms modeling.

```
[22] from sklearn.neighbors import KNeighborsClassifier
     knn = KNeighborsClassifier(n_neighbors=1, metric = 'euclidean', p=2)
     knn.fit(x_train70, y_train70)

     KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=1)
```

Fig. 8. KNN modeling in 70:30.

Fig. 9 shows the KNN script used to process 75% of data training and 25% of data testing of algorithms modeling.

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    knn1_75 = KNeighborsClassifier(n_neighbors=1, metric = 'euclidean', p=2)
    knn1_75.fit(x_train75, y_train75)

    KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=1)
```

Fig. 9. KNN modeling in 75:25.

Fig. 10 shows the KNN script used to process 80% of data training and 20% of data testing of algorithms modeling.

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    knn1_80 = KNeighborsClassifier(n_neighbors=1, metric = 'euclidean', p=2)
    knn1_80.fit(x_train80, y_train80)

    KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=1)
```

Fig. 10. KNN modeling in 80:20.

D. Data Visualization

In this study, the authors used bar charts to visualize the research results to make it easier for readers to understand. The data displayed in the diagram is a comparison of data classification between training data and testing data after prediction using the Naïve Bayes algorithm and also KNN.

Table VI compares the original data and the prediction data from the classification using the Naïve Bayes algorithm where for the Very Good status in the prediction data amounted to 1,004 employees while in the testing data amounted to 1,010 people. Fair status in the prediction data amounted to 60 employees while in the testing data amounted to 47 employees. For employees who have a Good status in the prediction data, there are 18 people while in the testing data there are 20 people. And for the Less status in the prediction data amounted to 9 people while in the testing data amounted to 14 people.

TABLE VI. COMPARISON OF PREDICTION DATA AND TESTING DATA WITH NAÏVE BAYES

Status	Prediction	Testing
Very Good	1004	1010
Fair	60	47
Good	18	20
Less	9	14

As for the comparison between the original data and the prediction data from the classification using the KNN algorithm shown in Table VII, namely for the Very Good status in the prediction data totaling 1008 employees while in the testing data totaling 1010 people, fair status in the prediction data amounted to 50 employees while in the testing data amounted to 47 employees. For employees with Good status in the prediction data, there are 19 people, while in the testing data, there are 20 people. The Less status in the prediction data amounted to 14 people, while in the testing data, it amounted to 14 people.

TABLE VII. COMPARISON OF PREDICTION DATA AND TESTING DATA WITH KNN

Status	Prediction	Testing
Very Good	1008	1010
Fair	50	47
Good	19	20
Less	14	14

E. Testing

Testing in research is very important to measure the performance of machine learning models. With testing, we can determine how much the model can make accurate and relevant predictions. In this study, the author conducted testing using a confusion matrix.

F. Result Analysis

Based on Table VIII in the testing stage and the graph in the data visualization stage, the author compares the accuracy, precision, and recall values of the Naïve Bayes algorithm and the KNN algorithm to determine the results of this study. The accuracy, precision, and recall values are obtained using the confusion matrix. The following is a comparison of the accuracy values of the two algorithms.

TABLE VIII. CONFUSION MATRIX NAIVE BAYES

Categories	treeTraining	0 (Less)	1 (Fair)	2 (Good)	3 (Very Good)
70:30	0 (Less)	7	7	0	0
	1 (Fair)	0	46	0	1
	2 (Good)	2	0	18	0
	3 (Very Good)	0	7	0	1003
75:25	0 (Less)	7	7	0	0
	1 (Fair)	0	36	0	0
	2 (Good)	2	0	17	0
	3 (Very Good)	0	5	0	835
80:20	0 (Less)	7	5	0	0
	1 (Fair)	0	25	0	0
	2 (Good)	2	0	14	0
	3 (Very Good)	0	4	0	670
70:30	0 (Less)	13	1	0	0
	1 (Fair)	0	46	0	1
	2 (Good)	1	0	19	0
	3 (Very Good)	0	3	0	1007
75:25	0 (Less)	13	1	0	0
	1 (Fair)	0	35	0	1
	2 (Good)	1	0	18	0
	3 (Very Good)	0	3	0	837
80:20	0 (Less)	11	1	0	0
	1 (Fair)	0	25	0	0
	2 (Good)	1	0	15	0
	3 (Very Good)	0	2	0	672

From Tables IX, it can be concluded that the accuracy value produced by the KNN algorithm is greater than the accuracy value of the Naïve Bayes algorithm. In the KNN algorithm, we can also see that the highest level of accuracy falls on the value of k is 5. This is because, in the classification process of the KNN algorithm, the system will classify the data based on the nearest neighbor. Therefore, when creating a model in the KNN algorithm, we must set the neighbor distance or k value.

TABLE IX. ACCURACY, PRECISION, AND RECALL VALUE OF NAIVE BAYES AND KNN

Algorithm	Categories	Accuracy	Precision	Recall
Naive Bayes	70:30	98.44%	88.50%	79.00%
	75:25	98.45%	88.25%	84.50%
	80:20	98.48%	87.75%	86.00%
KNN (K = 1)	70:30	99.45%	95.75%	95.75%
	75:25	99.33%	95.00%	95.50%
	80:20	99.44%	95.00%	95.75%
KNN (K = 3)	70:30	99.54%	98.00%	96.50%
	75:25	99.55%	98.25%	97.00%
	80:20	99.44%	96.75%	96.50%
KNN (K = 5)	70:30	99.63%	98.50%	96.50%
	75:25	99.55%	98.25%	97.00%
	80:20	99.58%	97.25%	97.50%

IV. CONCLUSION AND SUGGESTIONS

Based on the results and discussion in the previous chapter, the conclusions that can be drawn from this study as below:

The implementation of the Naïve Bayes and KNN algorithms in classifying employees' competency assessments is by studying patterns from available datasets where the data contains examples that have been classified correctly. In this study, the dataset is divided into 2 parts, namely training data and testing data, with a ratio of 70:30, 75:25, and 80:20. For the Naïve Bayes algorithm, the prediction process for classification is done by calculating the probability of the appearance of each class in the dataset. The classification process in the KNN algorithm compares the testing data with the training data that has been classified and classifies the testing data based on the majority of its nearest neighbor classes.

The accuracy rate produced by the Naïve Bayes algorithm in classifying employee competency assessments is 98%. While the accuracy rate produced by the KNN algorithm is 99%.

Based on the points above, the better algorithm in classifying employee competency assessment is the K-Nearest Neighbor (KNN) algorithm because it has a greater accuracy rate of 99%.

With these accurate results, the Naive Bayes and K-Nearest Neighbor algorithms are well used to classify employees' performance competency assessments using employees' assessment and learning journey data. With a total data of 3634 employees' data.

This research is suitable for companies that want to conduct employees' performance appraisals and still do it manually or subjectively. This research can provide a reference for companies to implement machine learning into one of their business processes. This research can also be a recommendation for companies to determine the parameters that can be used to assess employees' performance.

During this research, the author also realizes that this research has weaknesses and limitations. Therefore, the authors have several suggestions to improve the weaknesses and shortcomings in this study, namely: Additional parameters are needed to measure and classify employees' competence; other techniques are expected to be added to the testing process so that the accuracy produced by the two algorithms can be compared in more detail; and it is hoped that further research can develop a good user interface. So that users can use the system more easily. This research does not have a user interface.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

ERK and MAF prepared the concept of the research; KGA and DWS worked on the data collection; IF and FNI worked on the creation of algorithm models; DH and HIS worked-on data cleaning; RWK worked on the data visualization; Mansuri wrote the paper; all authors had approved the final version.

REFERENCES

[1] C. Naya, "Analysis of naïve bayes algorithm for classification of exemplary employees at pt. Toyoseal Indonesia," *SIGMA –Jurnal Teknologi Pelita Bangsa*, vol. 10, no. 2, 2019. (in Indonesian)

[2] C. E. A. Pah, "Decision support model for employee recruitment using data mining classification," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 1511–1516, May 2020. doi: 10.30534/ijeter/2020/06852020

[3] S. A. Laga, "Comparison of K-NN and SVM methods based on employee performance," *J. Sist. Komput. Dan Inform. JSON*, vol. 4, no. 3, 420, Mar. 2023. doi: 10.30865/json.v4i3.5816 (in Indonesian)

[4] A. Koda, P. Rahayu, A. Pratama, A. Rafly, and Kaslani, "Employee bonus determination using k-nearest neighbor algorithm," *KOPERTIP J. Ilm. Manaj. Inform. Dan Komput.*, vol. 4, no. 1, pp. 14–20, Jun. 2022. doi: 10.32485/kopertip.v4i1.115 (in Indonesian)

[5] A. P. Widyassari and P. E. Suryani, "Comparison of Naïve Bayes and SAW methods for selection of employee incentive receipt," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 02, pp. 149–159, Nov. 2021. doi: 10.46772/intech.v3i02.555 (in Indonesian)

[6] N. Giarsyani, "Comparison of machine learning and deep learning algorithms for named entity recognition: Case study of disaster data," *Indones. J. Appl. Inform.*, vol. 4, no. 2, 138, Aug. 2020. doi: 10.20961/ijai.v4i2.41317 (in Indonesian)

[7] S. K. Setianto and D. Jatikusumo, "Employee turnover analysis using comparison of decision tree and naive bayes prediction algorithms on k- means clustering algorithms at PT. AT," *Jurnal Mantik*, vol. 4, no. 3, 2020.

[8] P. Arsi, B. A. Kusuma, and A. Nurhakim, "Naive bayes classifier-based capital move sentiment analysis," *J. Inform. Upgris*, vol. 7, no. 1, Jun. 2021. doi: 10.26877/jiu.v7i1.7636 (in Indonesian)

[9] A. Senika, R. Rasiban, and D. Iskandar, "Implementation of Naïve Bayes method in rating sales marketing performance at PT. Pachira Distrinusa," *J. Media Inform. Budidarma*, vol. 6, no. 1, 701, Jan. 2022. doi: 10.30865/mib.v6i1.3331 (in Indonesian)

[10] A. R. Wicaksono, S. H. Wijoyo, and N. Y. Setiawan, "Analysis of the performance evaluation results of the functional position group of tax auditors of the PMA dua tax service office using the naïve bayes method," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 9, pp. 4086–4094, 2021. (in Indonesian)

[11] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment analysis about e-commerce from tweets using decision tree, k-nearest neighbor, and naïve bayes," in *Proc. 2018 International Conference on Orange Technologies (ICOT)*, Nusa Dua, Bali, Indonesia, IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/ICOT.2018.8705796 (in Indonesian)

[12] A. D. Wibisono, S. D. Rizkiono, and A. Wantoro, "Filtering email spam using naive bayes method," *Telefortech J. Telemat. Inf. Technol.*, vol. 1, no. 1, Jul. 2020. doi: 10.33365/tft.v1i1.685 (in Indonesian)

[13] E. R. Kaburuan, Y. S. Sari, and I. Agustina, "Sentiment analysis on product reviews from Shopee marketplace using the naïve

bayes classifier," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 3, 150, Nov. 2022. doi:10.24843/LKJITI.2022.v13.i03.p02

[14] E. Peranginangin, E. R. Kaburuan, and Y. Andrea, "Comparison of naïve bayes and LIWC for sentiment analysis of Gojek (Goto financial) user satisfaction," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 22, 2023.

[15] I. Ranggadara, G. Wang, and E. R. Kaburuan, "Applying customer loyalty classification with RFM and Naïve Bayes for better decision making," in *Proc. 2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, IEEE, Sep. 2019, pp. 564–568. doi: 10.1109/ISEMANTIC.2019.8884262

[16] D. Dedy and A. Cherid, "Data mining data processing of prospective Indonesian migrant workers (PMI) with the application of k-means clustering method and K-Nearest Neighbor (KNN) classification method: Case study of PT. SAM," *Format J. Ilm. Tek. Inform.*, vol. 9, no. 2, 166, Jan. 2021. doi: 10.22441/format.2020.v9.i2.008 (in Indonesian)

[17] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementation of K-Nearest Neighbor (KNN) classification algorithm for scholarship recipient selection classification," *IJCIT Indones. J. Comput. Inf. Technol.*, vol. 6, no. 2, Dec. 2021. doi: 10.31294/ijcit.v6i2.10438 (in Indonesian)

[18] I. L. Emmanuel-Okereke and S. O. Anigbogu, "KNN and SVM machine learning to predict staff due for promotions and training," *The International Journal of Engineering and Science (IJES)*, vol. 11, issue 4, pp. 26–34, 2022.

[19] S. Suprayogi, C. A. Sari, and E. H. Rachmawanto, "Sentiment analyst on twitter using the K-Nearest Neighbors (KNN) algorithm against COVID-19 vaccination," *J. Appl. Intell. Syst.*, vol. 7, no. 2, pp. 135–145, Sep. 2022. doi: 10.33633/jais.v7i2.6734

[20] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using Naive Bayes, KNN and random forest," in *Proc. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, IEEE, Mar. 2019, pp. 679–684. doi: 10.1109/ICCMC.2019.8819654