Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification

Sindi Fatika Sari, Kemas Muslim Lhaksmana*

Informatics, S1 Informatics, Telkom University, Bandung, Indonesia
Email: [1]sindifatikas@student.telkomuniversity.ac.id, [2*]kemasmuslim@telkomuniversity.ac.id
Submitted: **12/08/2022**; Accepted: **23/08/2022**; Published: **30/08/2022**

Abstrak−Employee attrition adalah hilangnya karyawan dalam suatu perusahaan yang disebabkan oleh beberapa faktor, yaitu karyawan mengundurkan diri, pensiun, atau faktor lainnya. Employee attrition dapat berdampak negatif pada suatu perusahaan jika tidak ditangani dengan baik, antara lain penurunan produktivitas. Perusahaan juga membutuhkan lebih banyak waktu dan tenaga untuk merekrut dan melatih karyawan baru untuk mengisi posisi yang kosong. Prediksi attrition ini bertujuan untuk membantu bagian sumber daya manusia (SDM) pada perusahaan untuk mengetahui faktor-faktor apa saja yang memengaruhi terjadinya attrition karyawan. Penelitian ini mengimplementasikan Random Forest dengan membandingkan metode seleksi fitur Information Gain, Select K Best, dan Recursive Feature Elimination untuk mencari seleksi fitur mana yang menghasilkan performasi terbaik. Penerapan metode-metode tersebut mengungguli penelitian sebelumnya dalam hal akurasi, presisi, recall, dan skor f1. Dalam perancangan penelitian ini, penulis pertama mengumpulkan dataset, membuat program, dan menyusun jurnal. Penulis kedua membantu penulis pertama dalam memprogram dan menyiapkan jurnal. Dari hasil pengujian yang telah dilakukan, Information Gain menghasilkan nilai akurasi tertinggi yaitu sebesar 89.2%, sedangkan Select K Best menghasilkan nilai akurasi sebesar 87.8% dan Recursive Feature Elimination menghasilkan nilai akurasi sebesar 88.8%.

Kata Kunci: Klasifikasi; Employee Attrition; Seleksi Fitur; Information Gain; Random Forest

Abstract−Employee attrition is the loss of employees in a company caused by several factors, namely employees resigning, retiring, or other factors. Employee attrition of employees can have a negative impact on a company if it is not handled properly, including decreased productivity. The company also requires more time and effort to recruit and train new employees to fill vacant positions. This attrition prediction aims to help the human resources (HR) department in the company to find out what factors influence the occurrence of employee attrition. This research implements Random Forest while comparing Information Gain, Select K Best, and Recursive Feature Elimination feature selection methods to find which feature selection produces the best performance. The implementation of the aforementioned methods outperforms previous research in terms of accuracy, precision, recall, and f1 scores. In preparing this research, the first author collects data sets, makes programs, and compiles journals. The second author assists the first author in programming and preparing the journal. From the results of the tests that have been carried out, Information Gain produces the highest accuracy value of 89.2%, while Select K Best produces an accuracy value of 87.8% and Recursive Feature Elimination produces an accuracy value of 88.8%.

Keywords: Classification; Employee Attrition; Feature Selection; Information Gain; Random Forest

1. INTRODUCTION

With the rapid development of the economy and industry, the phenomenon of employee attrition has gradually become popular in recent years [1]. In a company or agency, attrition often occurs or the process of reducing employees is caused by various factors. Employee attrition is one part of people analytics to help make more appropriate human resource (HR) decisions [2]. Employees are an important element in a company to fulfill the vision and mission to be achieved by the company. By having superior employees, the company has a competitive advantage over other companies [3]. Therefore, we need a system that can manage human resources effectively and efficiently.

The reduction of employees can have a negative impact on the company because it brings new problems if not handled properly. When a company changes employees frequently, it can be said that the attrition level of the company is very high. The level of attrition itself is measured based on the number of employees who stop working within a certain period of time. If the attrition level is high, it can cause problems for the company, including recruitment time to recruit, train, and develop new employees to fill vacant job positions [4], Productivity declines, and new employees have to re-adapt. This makes performance not optimal.

Prediction of employee attrition is carried out to determine what factors can affect employee attrition and can provide initial information about employee reductions that may occur soon so that the company can take appropriate action against the situation. In this final project, the employee attrition prediction is made using the IBM HR Analytics dataset via the *Kaggle.com* site [5].

In this study, the authors will compare the use of the Information Gain, Select K Best, and Recursive Feature Elimination (RFE) selection features to find out what factors can affect the occurrence of attrition and provide initial information to the company regarding the possibility of employee attrition that will occur. Then, compare the performance results of the three feature selections using the Random Forest classification method. The Random Forest classification method is used because this method is very suitable for developing predictive models [6].

The difference from previous studies is in research [7] Random Forest produces a good accuracy value of 0.85, but it produces low precision, recall, and f1 score values. Precision value is 0.60, recall is 0.28, and an f1 score is 0.39. Therefore, this study develops previous research to seek more optimal results. In research [8]

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

compared five classification methods, namely SVM, Decision Tree, Random Forest, KNN, and Naive Bayes, the selection of features used is based on the correlation value. Meanwhile, in this study, we compare three feature selection methods: Information Gain, select K Best, and RFE. Another research on the prediction of employee attrition [4] comparison of 8 classification models. But there is no more detailed theoretical understanding and explanation of these methods, and all models produce low precision, recall, and f1 scores. Another research [9] compared four classification methods to evaluate employee performance, the average monthly and annual working hours spent in the company as a feature to predict whether a certain employee will leave or not, but there is no more detailed explanation of the dataset used. Research [11] predicts employees will leave the company, which may occur using deep neural networks and using the ADASYN sampling method. In contrast, this study uses SMOTE-ENN to compare the sampling method results with previous studies.

The author chose Random forest because this method produces a low error, high classification accuracy, can handle very large amounts of training data, and is effective for handling incomplete data [9]. To measure the performance of this research method is to use a confusion matrix.

2. RESEARCH METHODOLOGY

2.1 Research Stages

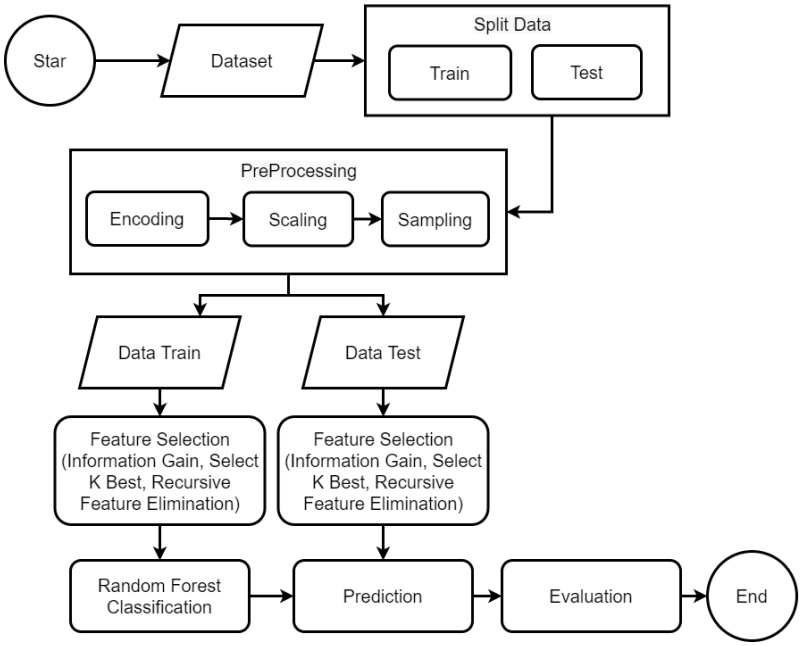


Figure 1. System design flowchart

Figure 1 shows how the flow of the system built in this study. The first step is to prepare the dataset. The dataset is IBM HR Analytics data via the *Kaggle.com* site [5]. The second stage divides the dataset into train and test data. The third stage is preprocessing, which consists of encoding, scaling, and sampling to change the raw data to make it easier to understand and ready to be processed by the system at the next stage. The fourth stage is feature selection to find the features that influence the prediction process most. The fifth stage is to perform a classification model on the train data. The sixth stage is making predictions. The last stage is an evaluation to calculate the accurate value of the classification process and the processes carried out.

2.2. Dataset

The dataset used in this study is the IBM HR Analytics dataset sourced from the kaggle.com site [5]. The dataset consists of 1470 data, 35 attributes, and uses English. After analyzing the data content of the dataset attribute, the 'EmployeeCount' attribute is dropped because the data content is sequential data only. Then drop the feature on the 'StandardHours', 'Over18', and 'EmployeeNumber' features because the data only contains the same values. After removing the feature, the total dataset used consists of 1470 data and 31 attributes. The attributes used in this dataset are listed in Table 1.

Table 1. Dataset features

Features Name	Type	Features Name	Type
Age	Numerical	MonthlyIncome	Numerical
Attrition	Categorical	MonthlyRate	Numerical
BusinessTravel	Categorical	NumCompaniesWorked	Numerical

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

Features Name	Type	Features Name	Type
DailyRate	Numerical	OverTime	Categorical
Departemen	Categorical	PercentSalaryHike	Numerical
DistanceFromHome	Numerical	PerformanceRating	Numerical
Education	Numerical	RelationshipSatisfaction	Numerical
EducationField	Categorical	StockOptionLevel	Numerical
EnvironmentSatisfaction	Numerical	TotalWorkingYears	Numerical
Gender	Categorical	TrainingTimesLastYear	Numerical
HourlyRate	Numerical	WorkLifeBalance	Numerical
JobInvolvement	Numerical	YearsAtCompany	Numerical
JobLevel	Numerical	YearsInCurrentRole	Numerical
JobRole	Categorical	YearsSinceLastPromotion	Numerical
JobSatisfaction	Numerical	YearsWithCurrManager	Numerical
MaritalStatus	Categorical		

2.3. Split Data

In this data sharing process, the dataset is divided into train data and test data, which is 70% train data and 30% test data. Data sharing is done by using random states whose purpose is to make the values consistent when the system is run. The number of datasets after splitting is 1029 trains and 441 test data.

2.4. Preprocessing

Preprocessing is an important step before classification. Preprocessing aims to process raw data into data that is ready to use and facilitate the classification process. Preprocessing in this study consists of several stages as follows:

a. Data Encoding
At this step, the dataset variable of type "categorical" is changed to "numerical" to equate all data types with making modeling easier. Data conversion is done using 'LabelEncoder'.

b. Feature Scaling
In an HR dataset, the data for each feature generally has a different scale [10]. For example, the age range of employees in the dataset ranges from 20 to 50 years, and earnings range from $1000 to $15,000. However, the presence of significant scale gaps between features usually slows down optimization algorithms [10]. This study normalized and standardized the original data set after the dataset type conversion step. Normalization can be done using the MinMax scaler with the equation (1).

$$x' = \frac{x - \min(x)}{max(x) - \min(x)} \quad (1)$$

Where x' is a new value, x is the old value, min (x) is the minimum value, and max (x) is the max value.

c. Sampling

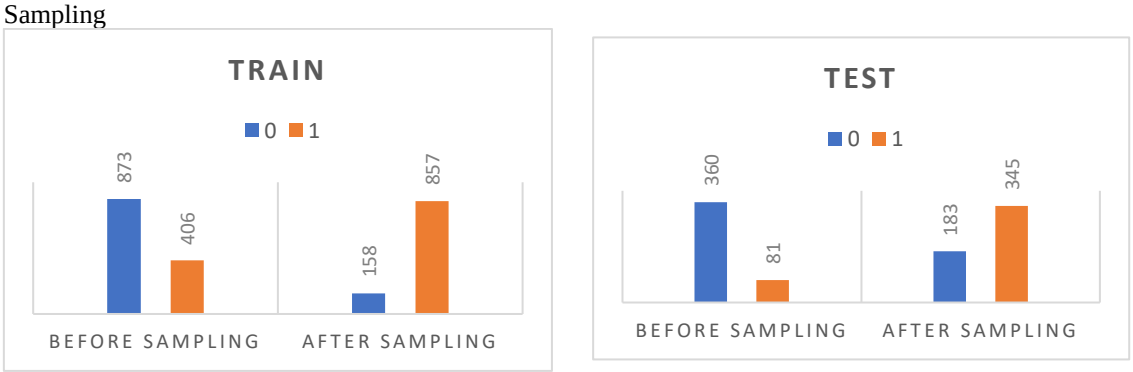


Figure 2. Sampling Train and Test

At this stage, resampling of the dataset is carried out to overcome unbalanced classes using the SMOTE-ENN method. SMOTE-ENN is a sampling technique that combines over and under sampling techniques in the minority class [11]. SMOTE-ENN works by first looking for the k-nearest neighbor of each observation, then checking whether the majority class of the k-nearest neighbor observation is the same as the observation class. If different, then both are deleted. The number of k used in the ENN is the default value of K = 3.

2.5. Feature Selection

After the preprocessing stage, the next step is selecting the dataset's features. This feature selection aims to eliminate or reduce considered unnecessary features and find out what features most affect employee attrition. In this study, we will compare three feature selections, namely:

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

a. Data Encoding
Information Gain is a simple and efficient feature selection method [12]. Determination of the features in the Information Gain is based on the most informative features in a particular class [13]. The best feature is determined by first calculating the entropy value using equation 2.

$$Entropy(S) = \sum_i^c - p_i \, log_2 \, p_i \quad (2)$$

Where c is the number of values in the classification class, and pi is the number of samples in class i. After that, the Information Gain is calculated using equation 3.

$$Gain(S, A) = \sum_{values\,(A)} \frac{|\,S_v|}{|S|} \; entropy\,(S_v) \quad (3)$$

Where *A* is attribute, *v* is the possible value of attribute *A*, Value (*A*) is the set of possible values of *A*, $|S_v|$ is the number of sample values *v*, |*S*| is the sum of all data samples, and $entropy\,(S_v)$ is the sample entropy with a value of v.

b. Select K Best
SelectKBest is a module in the sci-kit learns library that selects the k features with the top scores. Scores were calculated based on univariate statistical analysis, ie, variables were analyzed one by one [14]. In this study, selecting the k best features uses the 'import SelectKBest' library and the SelectKBest() function.

c. Recursive Feature Elimination.
Recursive feature elimination (RFE) is a recursive process that sorts the features according to their importance to the prediction process [15]. In each iteration, feature importance is measured, and less relevant features are removed. The advantages of RFE are that it is easy to configure, use, and can efficiently select features to predict target variables.

2.6. Feature Selection

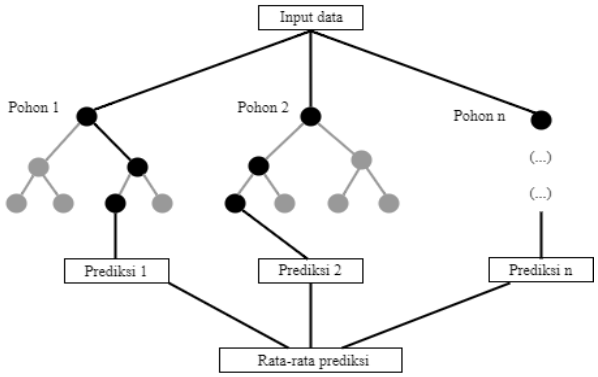


Figure 3. Random Forest method step

Random forest is one of the Machine Learning Supervised classification techniques invented by Leo Breiman and Adele Cutler in 2000 [6] and developed to improve the decision tree method, which is prone to overfitting [16]. In its development, this method has become one of the most popular methods in machine learning [16]. Random forest is a method that can be used to develop predictive models [6]. A Random Forest consists of many decision trees, from the 1st tree to the nth tree, where n is the total number of trees in the Random Forest [17].

In Figure 5, the Random Forest method combines each tree from the best decision tree model, then combines them into a model. The more trees used, the better the accuracy. Determination of the classification is formed based on the voting results of the formed tree.

This method is used to take data attributes randomly in accordance with applicable regulations and build a decision tree consisting of root nodes, internal nodes, and leaf nodes. The root node is the top node or is the input commonly called the root of the decision tree. An internal node is a branch node that has at least two outputs and only one input. The leaf node is the last node that has only one input and no output. The decision tree first calculates the Gini value as a branch determinant at the node. Calculation of the Gini value using the equation (4).

$$GINI\,(t) = 1 - \sum_j [\, p\left(\frac{j}{t}\right)]^2 \quad (4)$$

Where $p\left(\frac{j}{t}\right)$ is the relative frequency of class j at node t. After that, when the node p is divided into k partitions (children), the quality of the split is calculated using the equation (5).

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} \; GINI\,(i) \quad (5)$$

Where n_i is the number of records in child *i* and *n* is the number of records in node *p*.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

The lowest Gini index value will be the best split value for the attribute. After the class from each decision, a tree is formed, and voting is carried out for each class in the sample data. Then, the votes from each class are combined, and the most votes are obtained.

2.7. Evaluation

This system is used to measure the modeling performance of the classification predictions that have been built using the Confusion Matrix. The confusion matrix table can be seen in Table 2.

Table 2. Confusion Matrix

	Positive Actual	Negative Actual
Positive Prediction	TP	FP
Negative Prediction	FN	TN

The confusion matrix is divided into two classes, namely positive class and negative class [18]. Then there are 4 categories, namely positive (TP), true negative (TN), false positive (FP), dan false negative (FN).

a. Accuracy
Accuracy is the total number of correct predictions [18]. Accuracy can be formulated as follows:

$$\text{ACC (\%)} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

b. Precision
Precision is the accuracy of the predicted data correctly [18]. Precision can be formulated as follows:

$$\text{PREC (\%)} = \frac{TP}{FP+TP} \quad (7)$$

c. Recall
Recall is the accuracy of correctly identified data [18]. Recall can be formulated as follows:

$$\text{REC (\%)} = \frac{TP}{FN+TP} \quad (8)$$

d. F-Measure
F-Measure is a process to optimize the value of Precision and Recall [18]. F-Measure can be formulated as follows:

$$\textit{F-Measure (\%)} = \frac{2\ x\ REC\ x\ PREC}{REC+PREC} \quad (9)$$

2.8. AUROC Curve

ROC curve is a curve that researchers widely use to evaluate predictive results [19]. The ROC curve is divided into two dimensions, where the TP level is plotted on the Y axis, and the FP level is plotted on the X axis. To calculate the area under the ROC curve is to use the AUC (Area Under the ROC) method. AUC is a fraction of a square unit area, and its value always ranges from 0.0 to 1.0, so the greater the AUC value, the stronger the resulting classification [20].

Table 3. AUC Score

AUC Score	Classification
0.9 – 1.0	Best
0.8 – 0.9	Good
0.7 – 0.8	Medium
0.6 – 0.7	Low
0.5 – 0.6	Fail

3. RESULTS AND DISCUSSION

In this study, we will use a dataset that has gone through the resampling stage using SMOTE-ENN to predict employee attrition and has two target classes, namely yes (1) and no (0). The data train has 1,263 data consisting of 857 yes data and 406 no data. The test data has 528 data consisting of 345 yes data and 183 no data. There are three test scenarios in this study, the first scenario is to see the performance of using Information Gain feature selection, the second scenario is to see the performance results from using the Select K Best feature selection, and the third scenario is to see the performance results from the use of Recursive Feature Elimination (RFE). The last scenario is to see performance results without using feature selection. The four scenarios use the Random Forest classification method and will compare the performance results of the three feature selections.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

3.1 Information Gain Feature Selection Performance with Random Forest

In this scenario, Information Gain feature selection is used by sorting the features based on calculating the highest score and using the Random Forest classification method. The results of the feature ranking based on the calculation of the Information Gain value can be seen in Table 4 the features have been sorted by the highest score value. Testing will be carried out by selecting the number of features to be tested, namely 10, 15, 20, and 25. Then, comparing the accuracy values generated from each number of features tested.

Table 4. Feature rating based on Information Gain

Ranking	Features Name	Score	Ranking	Features Name	Score
1	TotalWorkingYears	0.306	16	RelationshipSatisfaction	0.194
2	YearsAtCompany	0.302	17	EducationField	0.189
3	DistanceFromHome	0.294	18	EnvironmentSatisfaction	0.180
4	YearsWithCurrManager	0.291	19	StockOptionLevel	0.174
5	Age	0.287	20	WorkLifeBalance	0.166
6	PercentSalaryHike	0.256	21	MaritalStatus	0.153
7	JobRole	0.256	22	JobInvolvement	0.134
8	NumCompaniesWorked	0.253	23	BusinessTravel	0.100
9	YearsInCurrentRole	0.239	24	MonthlyIncome	0.085
10	TrainingTimesLastYear	0.215	25	Overtime	0.073
11	JobSatisfaction	0.211	26	Department	0.062
12	HourlyRate	0.209	27	Gender	0.035
13	Education	0.199	28	DailyRate	0.016
14	JobLevel	0.198	29	MonthlyRate	0.004
15	YearsSinceLastPromotion	0.197	30	PerformanceRating	0.000

Table 5. Information Gain Results with Random Forest

Number of features	Performance			
	Accuracy	Precision	Recall	F1 Score
10	75.9%	73.4%	72.8%	73.1%
15	80.1%	78.0%	79.1%	78.4%
20	83.5%	81.6%	82.9%	82.1%
25	89.2%	87.8%	88.6%	88.2%

The performance results of the Information Gain feature selection test using the Random Forest classification method are shown in Table 5. Based on Table 5, the highest accuracy value was obtained using 25 features, which is 89.2%. This is probably caused by the selection of the number of features used in the classification modeling is very influential. In this scenario, it can be concluded that the more features used, the higher the accuracy, precision, recall, and f1 score value. In this scenario, the AUC value is also displayed which can be seen in Figure 4. The highest AUROC value is obtained by using 25 features, which is 0.953.

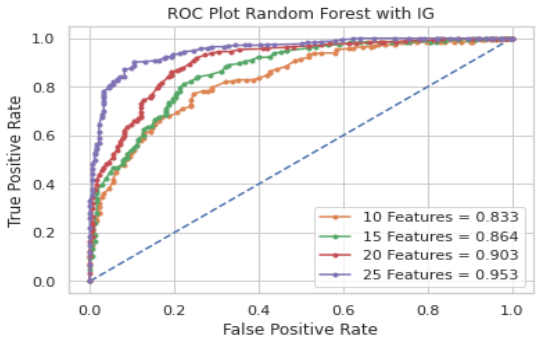


Figure 4. IG AUROC

3.2 Select K Best Feature Selection Performance with Random Forest

Table 6. Select K Best Result

Number of features	Features Name
10	'Age', 'JobLevel', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager'.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

Number of features	Features Name
15	'Age', 'DailyRate', 'DistanceFromHome', 'EnvironmentSatisfaction', 'Gender', 'JobLevel', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager'.
20	'Age', 'DailyRate', 'DistanceFromHome', 'EnvironmentSatisfaction', 'Gender', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'
25	'Age', 'DailyRate', 'Department', 'DistanceFromHome', 'EducationField', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'PercentSalaryHike', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'.

In this scenario, the Select K Best feature selection is used by sorting the features based on calculating the k value and using the Random Forest classification method. The results of feature selection based on the calculation of the value of k can be seen in Table 6, which will be tested using the number of features 10, 15, 20, and 25. Then, compare the accuracy values generated from each number of features tested.

Table 7. Select K Best Results with Random Forest

Number of features	Performance			
	Accuracy	Precision	Recall	F1 Score
10	84.6%	83.0%	83.0%	83.0%
15	85.4%	83.6%	84.9%	84.2%
20	87.8%	86.3%	86.7%	86.8%
25	87.1%	85.5%	86.6%	86.0%

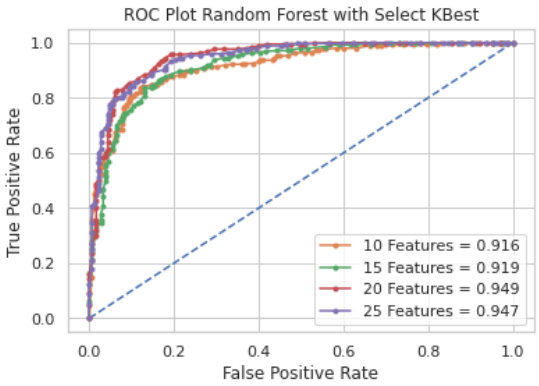


Figure 5. KBest AUROC

The performance results of the Select K Best feature selection test using the Random Forest classification method are shown in Table 7. Based on Table 7, the values of accuracy, precision, recall, and f1 score increased in the tests of 10, 15, and 20 features but experienced a slight decrease in the tests of 25 features. The highest accuracy value in this scenario is obtained when using 20 features, which is 87.8%. In this scenario, the AUC value is displayed which can be seen in Figure 5. The highest AUROC value is obtained by using 20 features, which is 0.949.

3.3 Recursive Feature Elimination Performance with Random Forest

Table 8. RFE Result

Number of features	Features Name
10	'Age', 'DistanceFromHome', 'JobLevel', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsInCurrentRole'.
15	'Age', 'DailyRate', 'DistanceFromHome', ‘EnvironmentSatisfaction’, 'HourlyRate', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole'.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

Number of features	Features Name
20	'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'HourlyRate', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'OverTime', 'PercentSalaryHike', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager'
25	'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EducationField', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime', 'PercentSalaryHike', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager'.

In this scenario, Recursive Feature Elimination is used by sorting the features based on the ranking of the most important features and using the Random Forest classification method. The top 10, 15, 20, and 25 features are taken in this scenario testing. Then, compare the accuracy values generated from each feature to be tested. The results of feature selection based on Recursive Feature Elimination can be seen in Table 8.

Table 9. RFE Results with Random Forest

Number of features	Performance			
	Accuracy	Precision	Recall	F1 Score
10	83.3%	81.5%	83.3%	82.1%
15	86.3%	84.7%	85.7%	85.1%
20	87.6%	86.1%	87.3%	86.6%
25	88.8%	87.2%	88.7%	87.9%

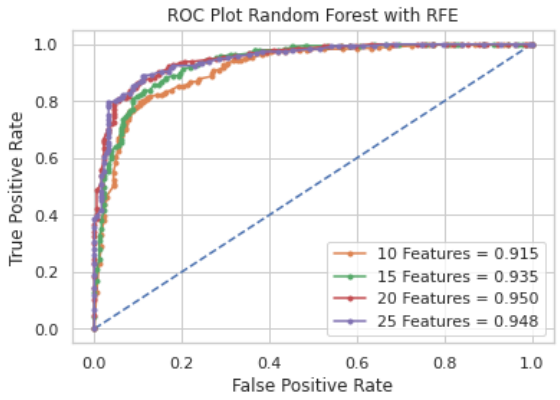


Figure 6. RFE AUROC

The results of the Recursive Feature Elimination test performance using the Random Forest classification method are shown in Table 9. Based on Table 9, the more features used, the higher the accuracy value. The highest value of accuracy, precision, recall, and f1 score is obtained when using 25 features. The accuracy value is 88.8%. In this scenario, the AUC value is displayed which can be seen in Figure 6. The highest AUROC value is obtained by using 20 features, which is 0.950.

3.4 Random Forest Performance without Feature Selection

In this scenario, testing is carried out using all the features in the dataset. This test will display the performance value using all the number of features, namely as many as 30 features using the Random Forest classification method. Performance results can be seen in Table 10.

Table 10. Result without Feature Selection

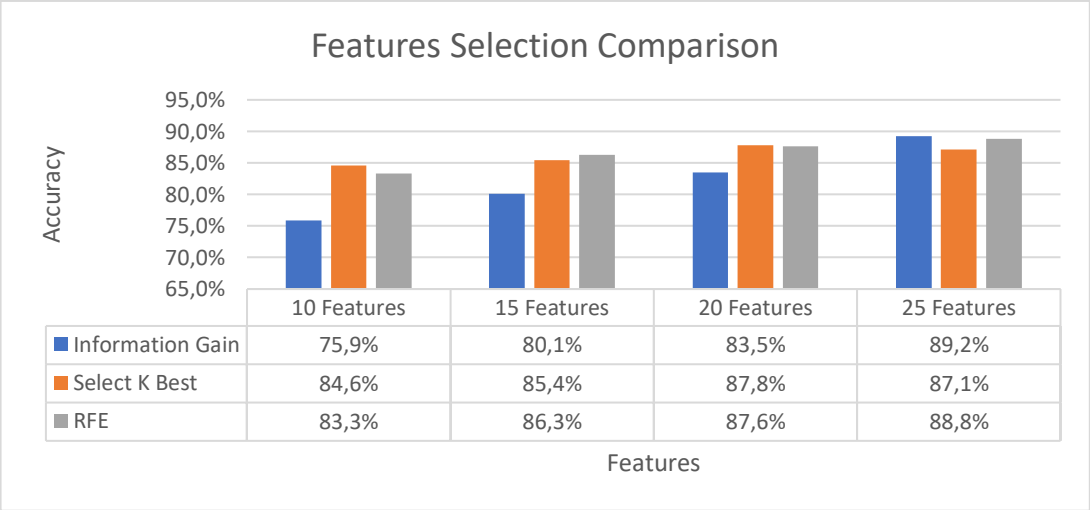
Accuracy	Precision	Recall	F1 Score
88%	86.5%	87.7%	87%

3.5 Implementation

The results of the feature selection comparison are shown in Figure 7. Information gain feature selection accuracy has increased. By using the Information Gain feature selection method, the highest accuracy value was obtained

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

from the use of 25 features, which is 89.2%. By using the Select K Best feature selection method, the accuracy value has increased in the 10, 15, and 20 feature experiments. Then experience a decrease in the value of accuracy in the experiment of 25 features. 20 features obtained the highest accuracy value using the Select K Best feature selection method is 87.8%. In the use of the Recursive Feature Elimination method, the accuracy increases with each trial the number of features used. The highest accuracy value in the use of Recursive Feature Elimination was obtained from the use of 25 features is 88.8%.



	10 Features	15 Features	20 Features	25 Features
Information Gain	75,9%	80,1%	83,5%	89,2%
Select K Best	84,6%	85,4%	87,8%	87,1%
RFE	83,3%	86,3%	87,6%	88,8%

Figure 7. Feature Selection Comparison

In the experiment using 10 features, the highest accuracy was obtained by using the Select K Best feature selection method, which was 84.6%. The experiment used 15 features the highest accuracy was obtained by using the Recursive Feature Elimination method, which was 86.3%. The experiment used 20 features the highest accuracy was obtained by the Select K Best method, which was 87.8. The last experiment used 25 features the highest accuracy was obtained by the Information Gain feature selection method, which was 89.2%.

From the test results in Figure 3.9, there is an increasing and decreasing accuracy each time a different feature is used. However, this does not mean that this experiment is not successful because it is based on research statements [21] that if the accuracy increases, the number of features is the optimal number of features, and if the accuracy decreases, then the number of features is not optimal.

4. CONCLUSION

This study aims to detect employee attrition in a company by implementing the Random Forest classification modeling by comparing feature selection of Information Gain, Select K Best, and Recursive Feature Elimination. Based on the test scenario in this study, Information Gain obtains the highest accuracy value when using 25 features, which is 89.2%. The Select K Best feature selection method gets the highest accuracy value when using 20 features of 87.8%. The Recursive Feature Elimination method obtains the highest accuracy value when using 25 features of 88.8%. Meanwhile, without using feature selection, the accuracy value is 88%. The Information Gain method produces better performance based on comparing the accuracy values of the three feature selections and without feature selection. Information Gain also produces better precision, recall, and f1 scores than previous research. The precision value is 87.8%, recall value is 88.9%, and f1 score value is 88.2%. Therefore, using feature selection can affect the classification accuracy of the Random Forest method to predict employee attrition. The number of features used greatly affects the accuracy of each feature selection. The more features used can increase the value of accuracy, precision, recall, and f1 score. Further research suggests comparing other feature selection or classification methods to find which technique can perform better than the method used in this study.

REFERENCES

[1] H. Zhang, L. Xu, X. Cheng, K. Chao dan X. Zhao, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning," 2018 18th International Symposium on Communications and Information Technologies (ISCIT), pp. 371-376, 2018.

[2] M. T. Bodie, M. A. Cherry, M. L. Mccormick dan J. Tang, "The Law and Policy of People Analytics," 2016.

[3] C. Stephanie dan R. Sarno, " Classification Talent of Employee Using C4.5,," 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 388-393, 2019.

[4] F. Fallucchi, M. Coladangelo, R. Giuliano dan E. W. De Luca, "Predicting employee attrition using machine learning techniques," Predicting employee attrition using machine learning techniques, vol. 9, no. 4, pp. 1-17, 2020.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 410−419
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2099

[5] P. "IBM HR Analytics Employee Attrition & Performance," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset. [Diakses 2022].

[6] J. L. Speiser, M. E. Miller, J. Tooze dan E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," vol. 134, pp. 93-101, 2019.

[7] A. Qutub, A. Al-Mehmadi, M. Al-Hssan dan R. Aljohani, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods," International Journal of Machine Learning and Computing, vol. 11, pp. 110-114, 03 2021.

[8] D. S. Sisodia, S. Vishwakarma dan A. Pujahari, "Evaluation of Machine Learning Models for Employee Churn Prediction," Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), pp. 1016-1020, 2017.

[9] A. Primajaya dan B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," pp. 27-31, 2018.

[10] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani dan H. S. Alghamdi, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods," International Journal of Machine Learning and Computing, vol. 11, no. 2, pp. 110-114, 2021.

[11] M. Muntasir Nishat, F. Faisal, I. Jahan Ratul, A. Al-Monsur, A. M. Abdullah, S. M. Nasrullah, M. T. Reza dan M. R. H. Khan, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," Scientific Programming, vol. 2022, 2022.

[12] A. E. Irsad, "SELEKSI FITUR INFORMATION GAIN UNTUK KLASIFIKASI INFORMASI TEMPAT TINGGAL DI KOTA MALANG BERDASARKAN TWEET MENGGUNAKAN METODE NAIVE BAYES DAN PEMBOBOTAN TF-IDF-CF," 2019.

[13] S. H. A. Aini, Y. A. Sari dan A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," pp. 2546-2554, 2018.

[14] T. Desyani, A. Saifudin dan Y. Yulianti, "Feature Selection Based on Naive Bayes for Caesarean Section Prediction," IOP Conference Series: Materials Science and Engineering, 2020.

[15] M. S. Wibawa, K. Dwi dan P. Novianti, "Reduksi Fitur Untuk Optimalisasi Klasifikasi Tumor," Konferensi Nasional Sistem & Informatika 2017, pp. 73-78, 2017.

[16] R. D. L. P, C. Fatichah dan D. Purwitasari, "Deteksi Gempa Berdasarkan Data Twitter Menggunakan Decision Tree, Random Forest, dan SVM," Jurnal Teknik ITS, vol. 6, pp. 153-158, 2018.

[17] D. G. I. Desantha dan K. M. Lhaksmana, "Aplikasi Sistem Seleksi Pelamar Kerja dengan menggunakan Metode," eProceedings of Engineering, vol. 7, pp. 9731-9738, 2020.

[18] R. M. Rifqi P, "KLASIFIKASI KEPRIBADIAN PADA SELEKSI PELAMAR KERJA DI PT. TELKOM INDONESIA MENGGUNAKAN PENDEKATAN SUPERVISED MACHINE LEARNING," 2020.

[19] N. Hadianto, H. B. Novitasari dan A. Rahmawati, "KLASIFIKASI PEMINJAMAN NASABAH BANK MENGGUNAKAN METODE NEURAL NETWORK," Jurnal Pilar Nusa Mandiri, vol. 15, no. 2, pp. 163-70, 2019.

[20] A. A. Rahayuningsih dan R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining," Analisis Perbandingan Algoritma Klasifikasi Data Mining, vol. VI, no. 1, 2018.

[21] T. K. Amory, A. dan W. Astuti, Jurnal Tugas Akhir Fakultas Informatika, 2021.