Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Predicting Employee Attrition Using Logistic Regression with Feature Selection

Fitri Herinda Wardhani[1)*], Kemas Muslim Lhaksmana[2)]
[1,2)]Telkom University, Indonesia
[1)]fitriherinda@student.telkomuniversity.ac.id, [2)]kemasmuslim@telkomuniversity.ac.id

Submitted : Sep 9, 2022 | **Accepted** : Sep 12, 2022 | **Published** : Oct 1, 2022

Abstract: Employee attrition is a reduction in employees that happens gradually. Employee attrition can damage the organization of a company, including the projects and its employee structure. This study aims to predict employee attrition in a company using the logistic regression method. Employee attrition can be predicted using machine learning because the machine learning approach is not biased due to human interference. In addition, human resources in a company need to know the most influential factors that cause the occurrence of employee attrition. In this study, we proposed feature selection methods to identify those influential factors and simplify the data training. Our approach is to predict employee attrition with three kinds of feature selection methods, namely information gain, select k-best, and recursive feature elimination (RFE). The 10-fold cross-validation was performed as an evaluation method. Prediction of employee attrition using the logistic regression method without applying feature selection gets an accuracy value of 0.865 and an AUC score of 0.932. However, by applying the RFE feature selection showed the highest evaluation result than information gain and select k-best, with an accuracy value of 0.853 and an AUC score of 0.925

Keywords: employee; attrition; logistic regression; classification; prediction

INTRODUCTION

In a company, human resources (HR) are a valuable asset that affects the company's success (Atul Chanodkar, Ravi Changle, 2019). One of the problems in managing human resources is the occurrence of employee attrition, which is a reduction in employees that happens gradually. A company is said to have a high attrition rate if there are a lot of employees deciding to leave the company. In general, there are three factors that cause employees to choose to leave the company. Those factors are personal characteristics, company system factors, and work environment (Harvida & Wijaya, 2020). Companies have to avoid the occurrence of employee attrition because it can have a destructive impact on the company system. When employees leave the company, they must recruit new employees for replacements, which will take more time and money (Alduayj & Rajpoot, 2019). The reduction in employees will also impact the projects and the structure of the company system so that the company's performance could not be maximal (Fallucchi et al., 2020). Therefore, the HR division needs to know the most influential factors that cause employee attrition. One way to deal with a high employee attrition rate is to build machine learning models to predict it. In this study, we predict employee attrition by applying the feature selection method using the logistic regression classification method. This research is expected to help HR managers in companies reduce attrition rates by predicting them early. The machine learning approach is not biased due to human interference, so the HR division can take the proper steps in the future to minimize employee attrition (Ghazi et al., 2021).

A few studies have focused on comparing several classification and data sampling methods. However, the application of several feature selection methods such, as information gain and select k-best, has rarely been studied directly. Therefore, in this study, we proposed feature selection methods to identify the most influential features in employee attrition and to simplify data training. We implemented a classification technique using the logistic regression algorithm by applying the feature selection method which includes supervised learning by finding patterns in the data using data train and data tests (Fallucchi et al., 2020). In this study, we proposed four main scenarios. The first scenario predicted employee attrition using the logistic regression classification method without applying feature selection. The second scenario predicted employee attrition using the logistic regression classification method by applying the information gain feature selection method. The third scenario predicted

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

employee attrition using the logistic regression classification method by applying the select k-best feature selection method. The fourth scenario predicted employee attrition using the logistic regression classification method by applying the Recursive Feature Elimination (RFE) feature selection method. These four approaches were used to determine which feature selection method is better than the others by comparing the evaluation results in terms of accuracy, precision, recall, f1-score, and AUC score. Furthermore, this study employed the feature selection method to determine which features that become the most influential factors in employee attrition. Implementing feature selection can also simplify data training. The evaluation method used to obtain the test results was k-fold cross validation with a k value of 10. K-fold cross-validation divides the data into k parts, using one part as test data and the remaining part as training data (Pal & Patel, 2020).

This study is based on several previous related research. There is a research conducted to investigate employee churn. It specifically aims to predict employee turnover that causes high turnover rates. This study proposed popular classification methods: decision tree, logistic regression, support vector machine, k nearest neighbor, random forest, and Naive Bayes. In addition, this study employed the recursive feature elimination method to compare the results with those of the classification without using the feature selection method (Yiğit & Shourabizadeh, 2017). Another study proposed a model for predicting employee turnover using nine classification methods. This study models a feature prediction framework to determine the most influential factors in the occurrence of employee turnover. The results showed that the most significant features affecting employee turnover are age; monthly income; behavioral features such as over time, years at company, and total working years; and attitudinal factors such as environment satisfaction and job satisfaction (Ghazi et al., 2021). There is another research which develops an employee turnover prediction model with a case study of the IT industry in India using the support vector machine method. The results showed that the model performs better at predicting employees who will leave the industry than those who will not (Khera & Divya, 2019). Another study develops a model to evaluate several machine learning methods for predicting employee churn. This study also identifies whether an employee would leave the company based on the details of the employee's work and work environment (Sisodia et al., 2018). The last research develops a model to predict employee attrition with three types of experiments. First, using imbalance training data and applying the classification using support vector machine, random forest, and k nearest neighbor methods. Second, using an adaptive synthetic approach (ADASYN) to overcome the imbalance class and classifying it using the same method as the previous experiment. Third, using manual under-sampling to overcome imbalanced data and applying the same classification as the method in the previous experiment (Alduayj & Rajpoot, 2019).

METHOD

Working on this study, consists of several stages, including preparing the dataset and then data preprocessing. The data preprocessing consists of data encoding, data scaling, and data sampling, followed by applying the feature selection method. Then the classification model uses the logistic regression algorithm and performs the evaluation method using the 10-fold cross-validation method and calculating the ROC AUC value. The research methods flow shown in Figure 1.

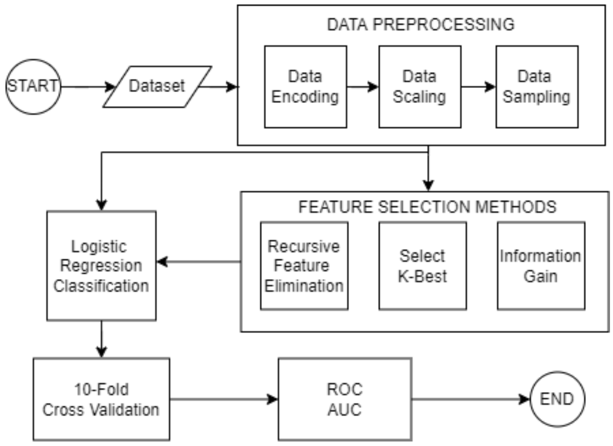

Figure 1. The Research Method Flow

Dataset

We used the IBM HR Employee Attrition dataset obtained from Kaggle (Pavansubhash, 2017). It consists of 35 features and 1470 rows. Each row has a label in the Attrition column or target class containing Yes and No values. The number of Yes classes is 1233, and No classes are 237. The features description is shown in Table 1.

*name of corresponding author

Table 1. Features and Their Data Types

Features	Data Type	Features	Data Type
Age	Numerical Value	MonthlyIncome	Numerical Value
Attrition	Categorical Value	MonthlyRate	Numerical Value
BusinessTravel	Categorical Value	NumCompaniesWorked	Numerical Value
DailyRate	Numerical Value	Over18	Categorical Value
Department	Categorical Value	OverTime	Categorical Value
DistanceFromHome	Numerical Value	PercentSalaryHike	Numerical Value
Education	Numerical Value	PerformanceRating	Numerical Value
EducationField	Categorical Value	RelationshipSatisfaction	Numerical Value
EmployeeCount	Numerical Value	StandardHours	Numerical Value
EmployeeNumber	Numerical Value	StockOptionLevel	Numerical Value
EnvironmentSatisfaction	Numerical Value	TotalWorkingYears	Numerical Value
Gender	Categorical Value	TrainingTimesLastYear	Numerical Value
HourlyRate	Numerical Value	WorkLifeBalance	Numerical Value
JobInvolvement	Numerical Value	YearsAtCompany	Numerical Value
JobLevel	Numerical Value	YearsInCurrentRole	Numerical Value
JobRole	Categorical Value	YearsSinceLastPromotion	Numerical Value
JobSatisfaction	Numerical Value	YearsWithCurrManager	Numerical Value
MaritalStatus	Categorical Value		

Data Preprocessing

The purpose of data preprocessing is to change the data to facilitate the processing and improve the quality of the data so that data can suit the needs. The quality of the data affects the model that will be built. Poor data quality will result in poor model quality as well because the quality of the data is more important than the model. Even the best model will not be able to produce good results if the quality of the data is poor. We perform three steps of data preprocessing. First, is data encoding. Data encoding is converting categorical data into numerical data. The data encoding is conducted so the machine learning model can process the data correctly. The approach to use is a label encoder. In the label encoder, categorical data is converted into consecutive numbers. Second, is data scaling. The approach taken is using min-max scaler. The process of min-max scaler is to make several variables have the same range of values. The purpose of this step is to simplify statistical analysis. Last, is data sampling. Data sampling is dealing with unbalanced class values. Data sampling uses the SMOTE-ENN (Synthetic Minority Oversampling Technique-Edited Nearest Neighbors) method. SMOTE-ENN is a combination of over-sampling using the SMOTE method and under-sampling using the ENN method. The SMOTE method adds synthetic data to the minority class by interpolating the original data so that the result of synthetic data varies. They also combine the ENN method to under-sampling the majority class by eliminating the majority class sample with different labels on adjacent data (Indrawati, 2021).

Feature Selection

Feature selection is the process of obtaining a subset of all the features in the dataset, and the selection is conducted with specific criteria. Feature selection aims to compress the data processing scale. The irrelevant features will be removed. The results of good feature selection can simplify the results of training data (Cai et al., 2018). In this study, we implement three kinds of feature selection methods. First, information gain. Information gain feature selection sorting the features by calculating the entropy value of each class. Entropy calculation is the basis for calculating information gain. Here, we show the entropy equation in Formula (1).

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (1)$$

Where n is the number of criteria in the class and pi is the ratio of the number of samples in class i to the total samples in the dataset. After obtaining each class's entropy value, then continued calculating the information gain using Formula (2).

$$Gain\ (S, A) = Entropy\ (s) \sum_{values\ (A)} - \left(\frac{S_v}{S}\right) Entropy\ (S_v) \quad (2)$$

Where A is the feature, v is the probability value of feature A. Values (A) is the probability value of the set A. Sv is total sample data and Entropy (Sv) is entropy example of v value. After performing the gain calculation, each feature is sorted according to the highest gain value. The feature with highest gain value is the dataset's

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

most appropriate and relevant feature (Hasibuan & Marji, 2019). The second feature selection method is select k-best. Select k-best feature selection is a type of univariate feature selection. Univariate feature selection works by selecting the best features based on univariate statistical tests. The select k-best method selects features based on the highest k value from statistical testing (Rahmansyah et al., 2018). The third feature selection method is Recursive Feature Elimination (RFE). The approach taken by the RFE method is to recursively select the optimal subset of features based on their importance to the prediction process. In each iteration, features will be ranked, and non-optimal or irrelevant features will be removed.

Employee Attrition Prediction Model Using Logistic Regression Classification

Classification with logistic regression is obtained from the data values and sends them to the sigmoid function. Logistic regression included in the regression analysis used when the dependent variable is categorical value (Ponnuru, 2020). A simple logistic regression model is a model for one independent variable X with a variable dichotomous Y. When the value of the variable Y is 1 indicates the existence of a characteristic, and Y is 0 indicates the absence of a characteristic (Paelongan et al., 2018). Logistic regression using the sigmoid function in its curve. The equation of the sigmoid function model shown in Formula (3).

$$f(z) = \frac{1}{1+e^{-z}} \quad (3)$$

f(z) is the probability of the logistic regression model (Ramadhy & Sibaroni, 2022). The parameter z is defined in Formula (4). Where w is the model's learned weights, b is bias, and x is the feature values.

$$z = b + w_1x_1 + w_2x_2 + \cdots w_nx_n \quad (4)$$

10-Fold Cross Validation

Cross Validation is an evaluation model that divides data into ten equal parts and conducts ten learning times. In each iteration, one part of the data is used as testing data, and the remaining nine parts are used as training data. The cross-validation evaluation value is obtained by calculating the average of the ten test results (Julianto et al., 2022).

ROC AUC

ROC (Receiver Operating Characteristic) curve is a graph showing the classification model's performance. The graph shows the relationship between the False Positive Rate (FPR) on the X-axis and the True Positive Rate (TPR) on the Y-axis. FPR and TPR are defined in Formulas (5) and (6).

$$FPR = \frac{FP}{FP+TN} \quad (5)$$

$$TPR = \frac{TP}{TP+FN} \quad (6)$$

AUC (Area Under ROC Curve) measures the area under the ROC Curve. The area under the curve provides an overview of the overall measurement of the model's suitability (Aswan Supriyadi Sunge, 2018). The category of AUC values is shown in Table 2.

Table 2. AUC Value Category

AUC Value	Category
0.90 – 1.00	Excellent Classification
0.80 – 0.89	Good Classification
0.70 – 0.79	Fair Classification
0.60 – 0.69	Poor Classification
0.50 – 0.59	Failure

RESULT

Data Preprocessing

In data preprocessing, the first thing to do is check the value of each data. The result shows that EmployeeCount, StandardHours, and Over18 contain a constant value. In other words, the value does not change from one employee to another. Furthermore, the value of EmployeeNumber is just a regular sequential number. So, we dropped that four features and processed the remaining 31 features in the next stage. The following process is data encoding which aims to convert categorical data into numerical data. Data encoding is essential, so machine learning models can precisely process the data. The approach we take is to use label encoder. Label

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

encoder converts categorical data into sequent numbers. The next step is the data scaling process. Data scaling is making several variables have the same range of values. Data scaling is necessary in order to simplify statistical analysis. The next stage is data sampling. We handle class imbalance by using the SMOTE-ENN approach. Before sampling, the value of class target Attrition is 1233 for Yes classes and 237 for No Classes. After sampling, the number of Yes classes is 1202, and No classes are 645.

Feature Selection

In this study, we compare the effect of three feature selection methods, information gain, select k-best, and RFE, on the logistic regression algorithm. Each feature selection is taken by 10 features and 20 features. In information gain, the gain value is calculated first in feature selection using the information gain method. After obtaining the gain value, the features are sorted based on the highest gain value, and then the top 20 features are selected, as shown in Table 3.

Table 3. Information Gain Selected Features

Features	Gain	Features	Gain
Age	0.338	YearsSinceLastPromotion	0.233
TotalWorkingYears	0.332	JobLevel	0.215
YearsAtCompany	0.310	Education	0.211
DistanceFromHome	0.300	TrainingTimesLastYear	0.207
PercentSalaryHike	0.291	EducationField	0.204
HourlyRate	0.281	EnvironmentSatisfaction	0.202
YearsInCurrentRole	0.270	RelationshipSatisfaction	0.182
YearsWithCurrManager	0.265	StockOptionLevel	0.171
JobRole	0.242	JobSatisfaction	0.169
NumCompaniesWorked	0.242	WorkLifeBalance	0.167

In feature selection using the select k-best method, the determined value of k is 10 and 20. After that, the features are sorted with the highest k value. The features obtained are shown in Table 4.

Table 4. Select K-Best Selected Features

Feature Count	Features
10 Features	Age, JobLevel, MaritalStatus, Monthly Income, OverTime, StockOptionLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWith CurrManager.
20 Features	Age, DailyRate, Department, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, OverTime, PercentSalaryHike, StockOptionLevel, TotalWorkingYears, TrainingTimes LastYear, YearsAtCompany, YearsInCurrentRole, YearsSince LastPromotion, YearsWithCurrManager.

In the RFE feature selection, the iteration is recursively based on the importance of features to the prediction process. Furthermore, the features will be ranked, and the top 10 features and top 20 features are selected. The selected features are shown in Table 5.

Table 5. RFE Selected Features

Feature Count	Features
10 Features	MaritalStatus, TotalWorkingYears, PercentSalaryHike, OverTime, MonthlyIncome, YearsSinceLast Promotion, JobSatisfaction, Job Involvement, YearsWithCurrManager, EnvironmentSatisfaction.
20 Features	Age, YearsInCurrentRole, YearsAtCompany, WorkLifeBalance, TrainingTimes LastYear, TotalWorkingYears, Percent SalaryHike, OverTime, NumCompaniesWorked, MonthlyIncome, YearsSince LastPromotion, JobSatisfaction, MaritalStatus, JobLevel, Department, DistanceFromHome, EnvironmentSatisfaction, YearsWithCurrManager, JobInvolvement, Gender

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation Result

After performing the feature selection process, the following step is classification modeling using the logistic regression method. The evaluation model uses the 10-fold cross-validation method where the data is divided into 10 parts, one part of the data is used as testing data, and the remaining nine data is used as training data. The evaluation results of all the scenario study are shown in Table 6. The results show that by applying feature selection methods, the accuracy value is not higher than without applying the feature selection method. Still, implementing feature selection can simplify data training. However, RFE with selecting 20 features is the best feature selection method than information gain and select k-best for predicting employee attrition with the logistic regression method with feature selection in terms of accuracy.

Table 6. Evaluation Result

Feature Selection Methods	Number of Features	Accuracy	Precision	Recall	F1-Score
Without Feature Selection	31 Features	**0.865**	0.883	0.914	0.898
Information Gain	10 Features	0.743	0.754	0.898	0.819
	20 Features	0.801	0.808	0.910	0.856
Select K-Best	10 Features	0.775	0.810	0.855	0.831
	20 Features	0.843	0.866	0.899	0.882
Recursive Feature	10 Features	0.811	0.838	0.879	0.858
Elimination (RFE)	20 Features	**0.853**	0.873	0.879	0.889

ROC AUC

We also calculate the ROC graph and the AUC value from the fourth main scenarios. This value is a parameter of whether a prediction is good or not. Without applying the feature selection method, it gets an AUC value of 0.932, which is included in the excellent classification category. The ROC curve shown in Figure 2.

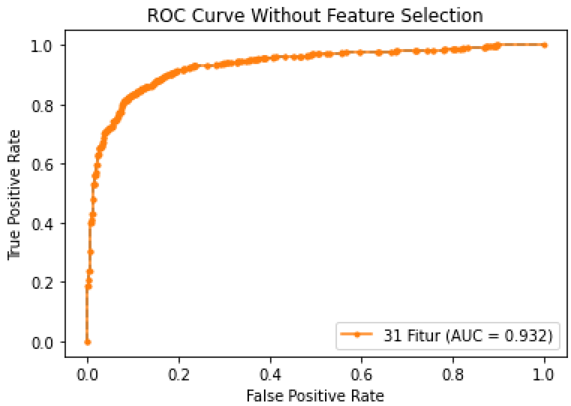

Figure 2. ROC Curve of Without Feature Selection

Using the information gain feature selection, when the ten features are selected, they get an AUC value of 0.771, which is included in the category fair classification. However, when the 20 features selected, it gets an AUC value of 0.861, which is included in the category good classification. The curve shown in Figure 3.

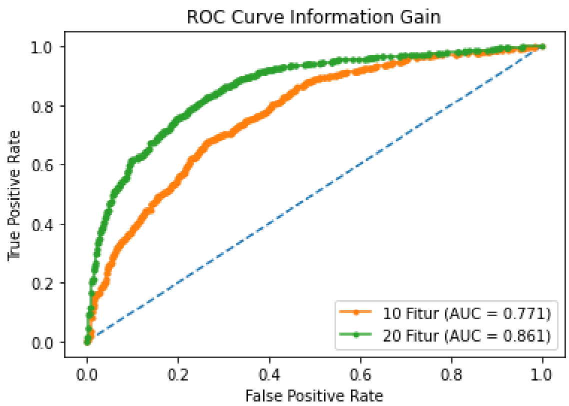

Figure 3. ROC Curve of Information Gain

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Select k-best gets an AUC value of 0.861 when ten features are selected; it is included in the category good classification. However, when the 20 features are selected, it gets an AUC value of 0.922 which means it is included in the category excellent classification. The curve shown in Figure 4.

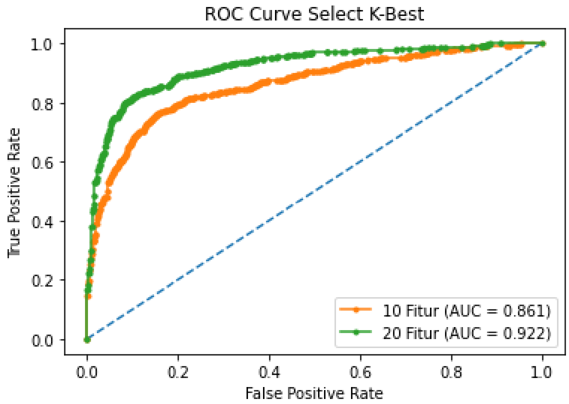

Figure 4. ROC Curve of Select K-Best

10 features selected by RFE get an AUC value of 0.904, and when the 20 features are selected, it gets an AUC value of 0.925 which means both are classified as excellent classification. The curve is shown in Figure 5.

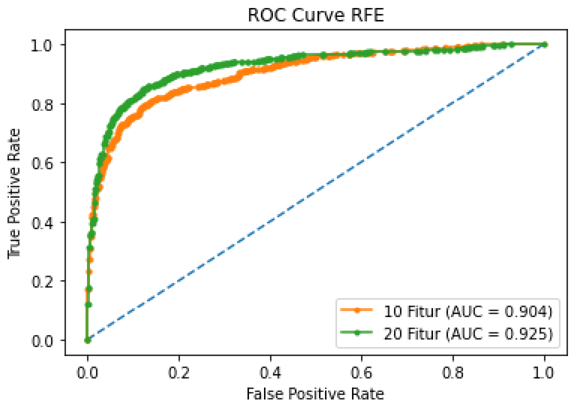

Figure 5. ROC Curve of RFE

DISCUSSION

The first scenario is predicting employee attrition using the logistic regression classification method without applying the feature selection method. From this scenario, it gets accuracy value of 0.865. The second scenario applying the information gain feature selection method by experimenting with selecting 10 features and 20 features. When using 10 selected features, it gets an accuracy value of 0.743 while when using 20 selected features, it gets an accuracy value of 0.801. The third scenario is using select k-best method. By selecting 10 features, the accuracy value is 0.775 and by selecting 20 features, the accuracy value is 0.843. The fourth scenario is using RFE method. When using only 10 selected features, it gets an accuracy value by 0.811. When using 20 selected features, the accuracy value is 0.853. The evaluation result show that by applying the feature selection method, namely information gain, select k-best and RFE, the accuracy value obtained is higher when selecting 20 features than only 10 features. Without applying the feature selection method, in other words using all the features in the dataset, which are 31 features, the accuracy value obtained is higher than by applying the feature selection method. Without implementing the feature selection, the accuracy value obtained is 0.865. Still, implementing feature selection can simplify data training. With implementing feature selection method, RFE is the best feature selection method than information gain and select k-best. It gets an accuracy value by 0.853. The evaluation result using 10-fold cross-validation can work quickly with structured sampling. In addition, we can see the model get the best accuracy because the data is divided into 10 partitions randomly so we can find out the composition of the best model and the best evaluation result.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11783
e-ISSN : 2541-2019
p-ISSN : 2541-044X

CONCLUSION

This study aims to predict employee attrition using the logistic regression method with feature selection. The three feature selections were compared to the logistic regression classification method, specifically information gain, select k-best, and recursive feature elimination. Prediction of employee attrition using logistic regression without applying feature selection method gets high accuracy value of 0.865 and an AUC value of 0.932. By applying feature selection methods, the accuracy value is not higher than without applying the feature selection method. Still, implementing feature selection can simplify data training. In applying feature selection, the method that gets the best evaluation results to predict employee attrition using the logistic regression classification method is applying the RFE method with selecting the top 20 features. We get a high accuracy value of 0.853, and the AUC value is 0.925, which means it is classified as excellent classification. The top 20 features were selected by the RFE method as the most influential features. The 20 features are Age, YearsInCurrentRole, YearsAtCompany, WorkLifeBalance, TrainingTimeLastYear, TotalWorkingYears, PercentSalaryHike, OverTime, NumCompanies Worked, MonthlyIncome, YearsSinceLastPromotion, JobSatisfaction, MaritalStatus, JobLevel, Department, DistanceFromHome, EnvironmentSatisfaction, YearsWithCurrManager, JobInvolvement, and Gender.

REFERENCES

Alduayj, S. S., & Rajpoot, K. (2019). Predicting Employee Attrition using Machine Learning. *Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018*, 93–98. https://doi.org/10.1109/INNOVATIONS.2018.8605976

Aswan Supriyadi Sunge. (2018). Prediksi Kompetensi Karyawan Menggunakan Algoritma C4 . 5 (Studi Kasus : PT Hankook Tire Indonesia). *Seminar Nasional Teknologi Informasi Dan Komunikasi 2018 (SENTIKA 2018)*, *2018*(Sentika), 23–24.

Atul Chanodkar, Ravi Changle, D. M. (2019). *Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms*.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 1–17. https://doi.org/10.3390/computers9040086

Ghazi, A. H., Elsayed, S. I., & Khedr, A. E. (2021). A proposed model for predicting employee turnover of information technology specialists using data mining techniques. *International Journal of Electrical and Computer Engineering Systems*, *12*(2), 113–121. https://doi.org/10.32985/IJECES.12.2.6

Harvida, D. A., & Wijaya, C. (2020). Faktor Yang Mempengaruhi Turnover Karyawan dan Strategi Retensi Sebagai Pencegahan Turnover Karyawan : Sebuah Tinjauan Literatur. *Jurnal Ilmu Administrasi Negara*, *16*(1), 13–23.

Hasibuan, M. R., & Marji. (2019). *Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)*. *3*(11), 10435–10443.

Indrawati, A. (2021). Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. *JIKO (Jurnal Informatika Dan Komputer)*, *4*(1), 38–43. https://doi.org/10.33387/jiko.v4i1.2561

Julianto, I. T., Kurniadi, D., Nashrulloh, M. R., & Mulyani, A. (2022). Comparison of Data Mining Algorithm For Forecasting Bitccoin Crypto Currency Trends. *Jurnal Teknik Informatika (Jutif)*, *3*(2), 245–248. https://doi.org/10.20884/1.jutif.2022.3.2.194

Khera, S. N., & Divya. (2019). Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. *Vision*, *23*(1), 12–21. https://doi.org/10.1177/0972262918821221

Paelongan, P. L., Palupi, I., Si, S., & Si, M. (2018). *Desain Model Prediksi Penderita Kanker Paru-Paru menggunakan Regresi Logistik Linier*. 3–15.

Pal, K., & Patel, B. V. (2020). Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques. *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, *Iccmc*, 83–87. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00016

Pavansubhash. (2017). *IBM HR Analytics Employee Attrition & Performance*. Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv

Ponnuru, S. R. (2020). Employee Attrition Prediction using Logistic Regression. *International Journal for Research in Applied Science and Engineering Technology*, *8*(5), 2871–2875. https://doi.org/10.22214/ijraset.2020.5481

Rahmansyah, A., Dewi, O., Andini, P., Hastuti, T., Ningrum, P., & Suryana, M. E. (2018). Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine. *Seminar*

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Nasional Aplikasi Teknologi Informasi (SNATi), 1907–5022.

Ramadhy, I. F., & Sibaroni, Y. (2022). Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression. *Jurnal Riset Komputer)*, *9*(1), 2407–389.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2018). Evaluation of machine learning models for employee churn prediction. *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, *Icici*, 1016–1020. https://doi.org/10.1109/ICICI.2017.8365293

Yiğit, I. O., & Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining. *IDAP 2017 - International Artificial Intelligence and Data Processing Symposium*. https://doi.org/10.1109/IDAP.2017.8090324

*name of corresponding author