ISSN : 1978-8282, Online ISSN: 2655-4275

Performance Analysis of Tree-Based Algorithms in Predicting Employee Attrition

Musthofa Galih Pradana[*1], I Wayan Rangga Pinastawa[2], Nurhuda Maulana[3], Wahit Desta Prastowo[4]
[1,2,3]Informatika, University Pembangunan Nasional Veteran Jakarta, Jakarta
[4]Informatika, University Alma Ata, Yogyakarta
E-mail: [*1]**musthofagalihpradana@upnvj.ac.id**, [2]rangga@upnvj.ac.id,
[3]nurhudamaulana@upnvj.ac.id, [4]wahitdesta@almaata.ac.id

Abstract

Based on data throughout 2022, there have been many reductions in employees both globally and Indonesia. The reduction was made due to adjustments with developments to keep the business afloat in increasingly fierce competition. However, reducing the number of employees is not an easy decision to make. This decision can have an impact on many aspects of the development and course of a business or company. To make a decision especially related to the aspect of termination of employment, it is necessary to consider carefully and thoroughly. Assessment and decision-making cannot be based on just one aspect, other aspects need to be seen to be taken into consideration. Additional aspects that can be selected to strengthen decision-making can be taken from the data. Data will not have any value without processing it with various approaches, one of which is the prediction process. Starting from the data, the prediction results will be more appropriate to make a decision. This study made a comparison of 3 decision tree algorithms, and produced a comparison of the three methods in terms of accuracy. The results of this study are the best accuracy for each algorithm C.45 = 83.44; Random Forests = 85.85; LMT = 88.29 with a linear precision value, and the best algorithm model with the highest accuracy is the Logistic Model Tree (LMT) algorithm.

Keywords *— Prediction, Attrition, C.45, Random Forest, Logistic Model Tree*

1. INTRODUCTION

Based on data throughout 2022, there have been many employee reductions. Reporting from Layoffs.fyi as of May 2022, a property company from the United States Better.com has terminated the employment of 3,000 of its employees. There have also been layoffs at other companies, such as Peloton which has laid off 2,800 employees, and Carvana with 2,500 employees. The reason for reducing the number of employees or employees is based on adjustments to business needs and the focus of the company[1]. The same thing happened to Indonesia, man layoff with fs, according to Kata Data, the biggest impact was the impact of the Covid-19 pandemic with data for 2022 of 4.15 million unemployed people that lay off employees such as shopee, and tani hub. Based on data, the highest number of dismissals occurred in November 2022 with a total of 45,000 job terminations. The dismissals occurred at Meta, Amazon, Twitter, GoTo, and SIRCLO employees. Dismissals that occurred at meta companies reached 11,000 employees, and Amazon as many as 10,000 employees[2].

In running a business or a company, of course, you don't always experience conditions that satisfy all parties, there are times when there are positions or conditions in which you have to make decisions by sacrificing several things or decisions that don't satisfy all parties. This is unavoidable because the flow of competition in the world of work or the industrial world is fluctuating and increasingly competitive. One of the things that might happen is the condition or decision that must be taken by the company by taking action to reduce the number of employees. This condition certainly results in employees' dissatisfaction or discomfort because they risk being terminated.

Observing what happened, termination of employment can be a vital decision in the course of an organization or business. Decision-making cannot be based solely on instinct or subjective judgment, a clear basis for making decisions is also needed. Therefore, the role of data is very important here, the data that is owned can be considered in making decisions taken. Based on the data, predictions can be made of employees or employees who will be evaluated and at the end of the day termination of employment will be carried out. One predictive approach that can be done is using a tree-based algorithm. In this study, predictions will be made using a tree-based algorithm to predict employees who will be terminated. The tree-based algorithm used is C.45, Random Forest, and Logistic Model Tree (LMT). These three algorithms will be analyzed for the performance that is produced in carrying out the prediction process related to data which in the end is related to the termination of employment for employees.

2. RESEARCH METHOD

The flow carried out in this study is as follows:

1. Literature Study
   This process is carried out by reviewing and observing previous research that is relevant to the research to be carried out. Referred research is up to date and has a reputation for maintaining the quality of referrals used.
2. Collecting Data
   The second process is collecting data, the data used is in the form of an employee dataset with 35 attributes with a total of 1470 data lines.
3. Prediction
   The prediction process in the third stage is carried out by carrying out the prediction process into 3 algorithms namely C.45, Random Forest, and Logistic Model Tree (LMT). The results of these three algorithms will be compared, and the results of the comparison can be analyzed further such as the value of accuracy, precision, and recall of each algorithm.
4. Testing
   The fourth process is using data testing, this process is continuous with the third process because this process is needed to produce performance values for each of the algorithms used.
5. Result
   The final stage of this research is to get the final results based on test scenarios that have been determined, to do the interpretation of the results.

The chart of the research flow is shown in Figure 1.

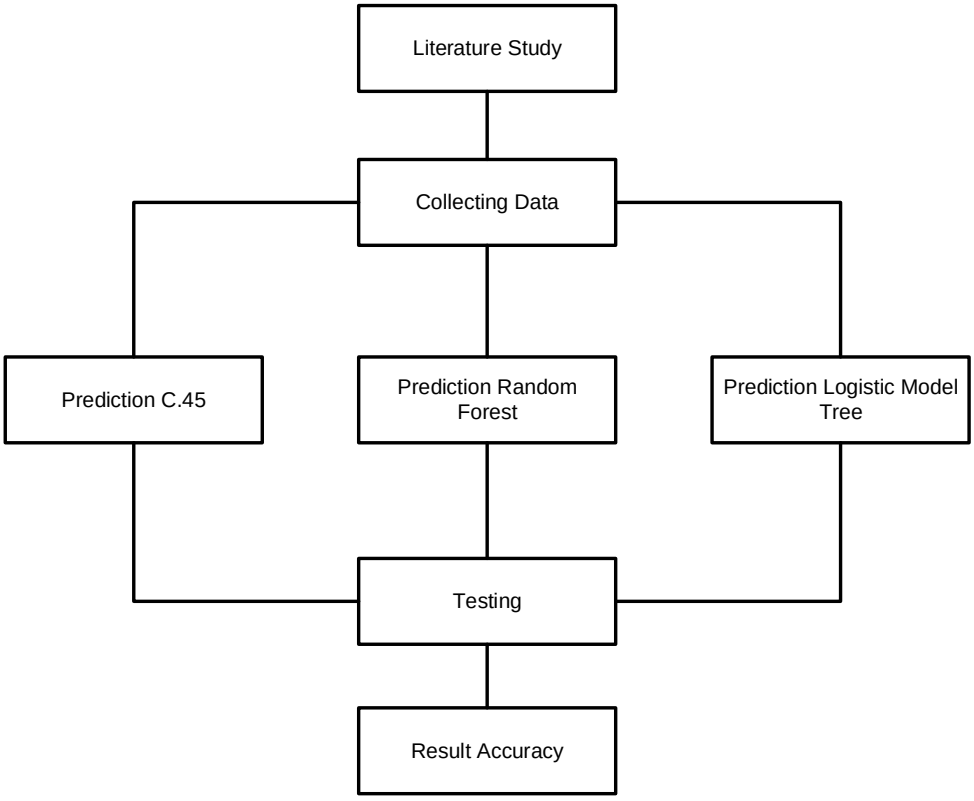

Figure 1. Research Flow

2.1. Literature Review

Relevant reference research from previous research is to apply the C.45 algorithm in classifying customer service quality at the BTN Pematangsiantar bank, there are 7 attributes used in this study such as age, occupation, tangible and other variables. The results of applying the C.45 algorithm are obtaining an accuracy of 77.78%, which means that banks need to improve service quality to increase customer satisfaction[3].

The application of the C.45 algorithm can also be used to predict graduation time. Apart from implementing the C.45 algorithm, this study also applies another algorithm, namely Naïve Bayes. The results of identification in this study are to obtain factors that influence graduation, namely the cumulative grade point average in the first semester, three, and four. Graduation patterns can be analyzed until students enter the second year[4].

Naïve Bayes and C.45 have also been written by Saruni Dwiasnati in predicting insurance customer decisions. This study uses variables such as price, quality, number of products, place, and service. The results of this study resulted in an accuracy of the Naive Bayes algorithm of 90.11% and the C4.5 algorithm of 75.27%[5].

Other relevant research applying random forest to make predictions were written by Jaime Lynn Speiser with the final result in this study being that this algorithm can assess variable selection techniques that have high variation with good results[6].

Random Forest which was written in the International Journal of Machine Learning and Cybernetics by Lakshmanaprabu with prediction results using a random forest that can classify data, with the best result value being a precision value of 94.2%. To verify the effectiveness of the proposed method, different performance measures are analyzed and compared with existing methods[7].

Random forest comparisons have also been written using a comparison technique with a support vector machine in Mohammadreza's research which resulted in the conclusion in this study is a comparison of descriptive and quantitative data with a total of 251 papers where the database contains 251 papers that meet the requirements where 42% and 68% of the database including RF and SVM implementations[8].

The random forest and logistic regression research published by Kailin made a comparison of these two algorithms with the results that the random forest has a higher value than the logistic regression and produces a higher false positive value in the dataset with an increase in the noise variable as well[9].

Another algorithm is the Logistic Model Tree (LMT) in predicting landslide susceptibility, the result is the performance of the algorithm using the area under the curve (AUC). The highest value of the LMT model is 84.7% and the Logistic Regression model at a value of 76.5%, from these results it can be concluded that the decision to apply these two algorithms is quite good and increases the predictive power of the landslide model[10].

The prediction of the Logistic Model Tree (LMT) algorithm in predicting Student Academic Performance in supporting an intelligent decision support system (IDSS) with research results got the highest accuracy value at a percentage of 83.48%. The results of this prediction are expected to provide an overview for parents, institutions, and management to make decisions[11].

Sheikh Amir Fayaz in his writing also researches the Logistic Model Tree (LMT) algorithm in predicting the accuracy of meteorological data. From the data, predictions of rainfall apply the LMT algorithm with good performance achieving an accuracy of 87.23%. In previous studies, Sheik Amir Fayaz has also made comparisons and the results obtained are that the LMT algorithm has accurate results[12].

The prediction of landslides applies the LMT algorithm written by Ha Thi Hang using performance measurement indices such as sensitivity (SST), specificity (SPF), accuracy (ACC), the area under the ROC curve (AUC), RMSE, and index k. The result is the best model RSSLMT performance (AUC: 0.816). The data used is open-source data on Vietnam's National Highway 6[13]. Predictions using random forests are applied as a measure for managing forest fires, in China by increasing awareness and preventive action in forest fires that occur. The results of this study were able to show an accuracy of 70% up to the highest value of 91.4%[14]. Forest fires are also related to air pollution caused, which can cause many types of diseases that affect humans, with particles less than 2.5 (PM 2.5), this can be predicted by PM 2.5 concentrations with random forest algorithms, XGBoost, and Deep Learning. The result is using Xboost to get the most optimal performance[15]. The performance

of the Random Forest algorithm can be improved as in Angshuman's research by removing iterations of several features that are not important and can dynamically change the size of the trees in the random forest[16]. The approach of using random forests can be applied in estimating populations, with research that carries out geographic implementations of random forests called Geographical Random Forests (GRF). Using Geographical Random Forest (GRF) can reduce the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values[17]. Predictions using LMT in cases of obesity show that this LMT algorithm has the best results compared to SVM and Random Forest with a yield of 96.65%[18]. Predictions using LMT also produce superior final results compared to other algorithms in predicting student graduation scores with the results of the LMT algorithm producing an accuracy rate of 71% better than Decision Tree J48[19]. Various types of algorithms in the decision tree can be evaluated because they have their respective abilities in predicting and classifying[20].

In the literature review conducted by Mohammad Reza, the use of Random Forest as a classification algorithm exceeds the use of SVM which was previously widely used[21]. This is an example of its use in detecting credit card fraud by applying the Random Forest algorithm with good results on small data sets, there are still some problems such as unbalanced data[22]. One of the tree-based algorithms, namely C45, its application in classifying the understanding of STIKOM students has an accuracy rate of 87.50%[23]. C45 can also achieve an accuracy value of 88.75% with an AUC value of 0.744 and testing on applications made in BCA customer satisfaction[24]. The application of the C45 algorithm can be optimized, such as the addition of swarm particle optimization which achieves an accuracy of 97.13% in measuring the readiness of junior high school students to face the national exam[25]. Another result of the C45 algorithm in predicting customer or customer loyalty is reflected in 2 similar results, namely the C.45 algorithm has good accuracy[26] compared to other algorithms such as Naïve Bayes[27], although in certain cases the results are sometimes contradictory, the resulting accuracy is still relatively good[28]. Apart from the C45 and Random Forest algorithms, the LMT algorithm is one of the algorithms that can be used for the classification process, such as in Natuthe ral Language Processing-based Mental Health Risk Prediction study, the most accurate prediction results were achieved in the DASA dataset using the sentiment dictionary and the LMT and SVM algorithms[29]. From the various research references that have been carried out, this research is positioned to seek results from the best accuracy of the three algorithms with different test data conditions.

3. RESEARCH RESULTS AND DISCUSSION

3.1. Datasets

The dataset used in this study is employee data with various attributes, the details of which are shown in Figure 2.

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipS
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	

5 rows × 35 columns

Figure 2. Datasets

While the details of the attributes used are as many as 35 attributes with attribute descriptions shown in Table 1.

Table 1. Attribute Dataset

No	Attribute	No	Attribute
1	Age	19	Monthly Income
2	Attrition	20	Monthly Rate
3	Business Travel	21	Num Companies Worked
4	Daily Rate	22	Over 18
5	Department	23	Over Time
6	Distance from Home	24	Percent Salary Hike
7	Education	25	Performance Rating
8	Education Field	26	Relationship Satisfaction
9	Employee Count	27	Standard Hours
10	Employee Number	28	Stock Option Level
11	Environment Satisfaction	29	Total Working Years
12	Gender	30	Training Times Last Year
13	Hourly Rate	31	Work-Life Balance
14	Job Involvement	32	Years at Company
15	Job Level	33	Years in Current Role
16	Job Role	34	Years Since Last Promotion
17	Job Satisfaction	35	Years with Current Manager
18	Marital Status		

The table above shows that several indicators or variables can determine the prediction results. One of the important variables is the length of time at the company, performance, and also the current position.

3.2. Data Visualization

The dataset used can be visualized in the form of a chart to carry out the initial mapping and observation of the data. The following are some of the visualization results of the data in graphical form.

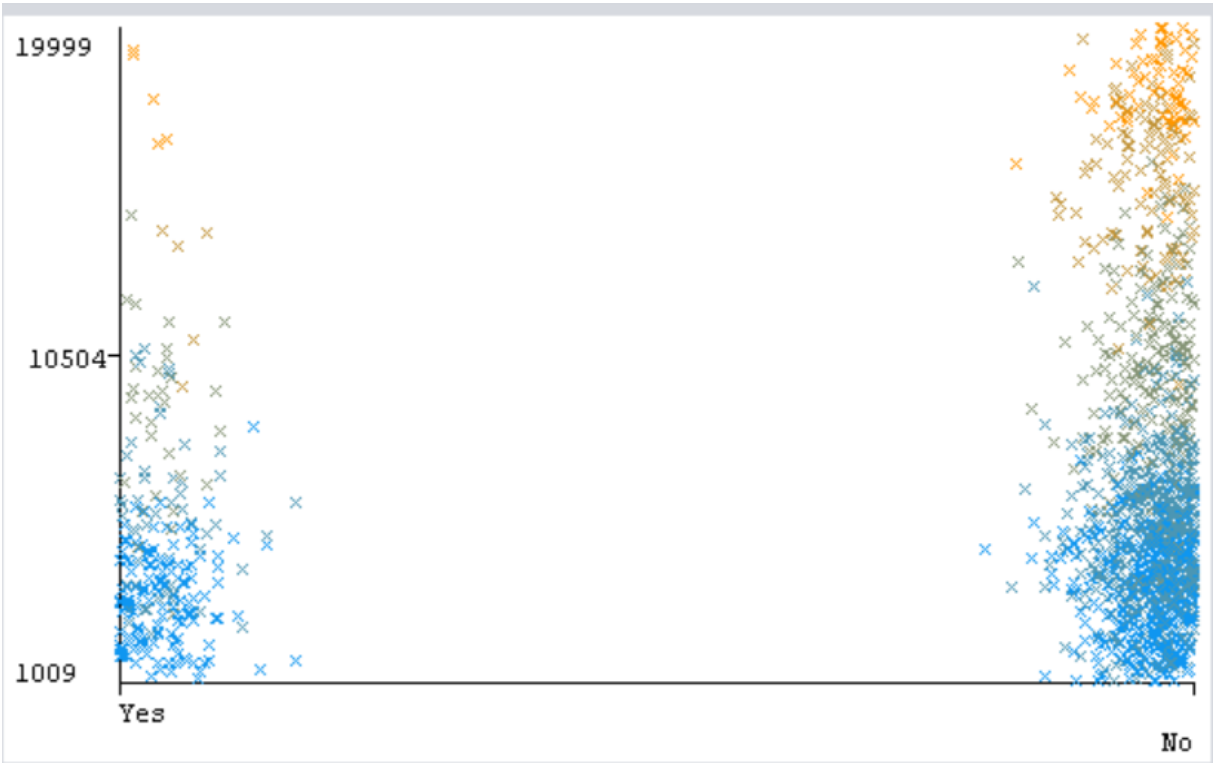

Figure 3. Visualization Job Level

The graphic above shows the level of work in blue for the lowest level, gray for the middle, and gold for the highest level of work, while the numbers on the graphic on the left show the amount of salary. The distribution of data, shows that the distribution is the most for the lowest level who experience attrition with a lower salary level.

Figure 4. Visualization Working Time

While the distribution of data in the graph above shows that with a working duration of 0 years in blue, gray for 20 years of work, and 40 years of work in gold in the range on the left, while the graph on the left shows the performance rating. The results show that attrition is often done at work levels that are still new, namely 0 years.

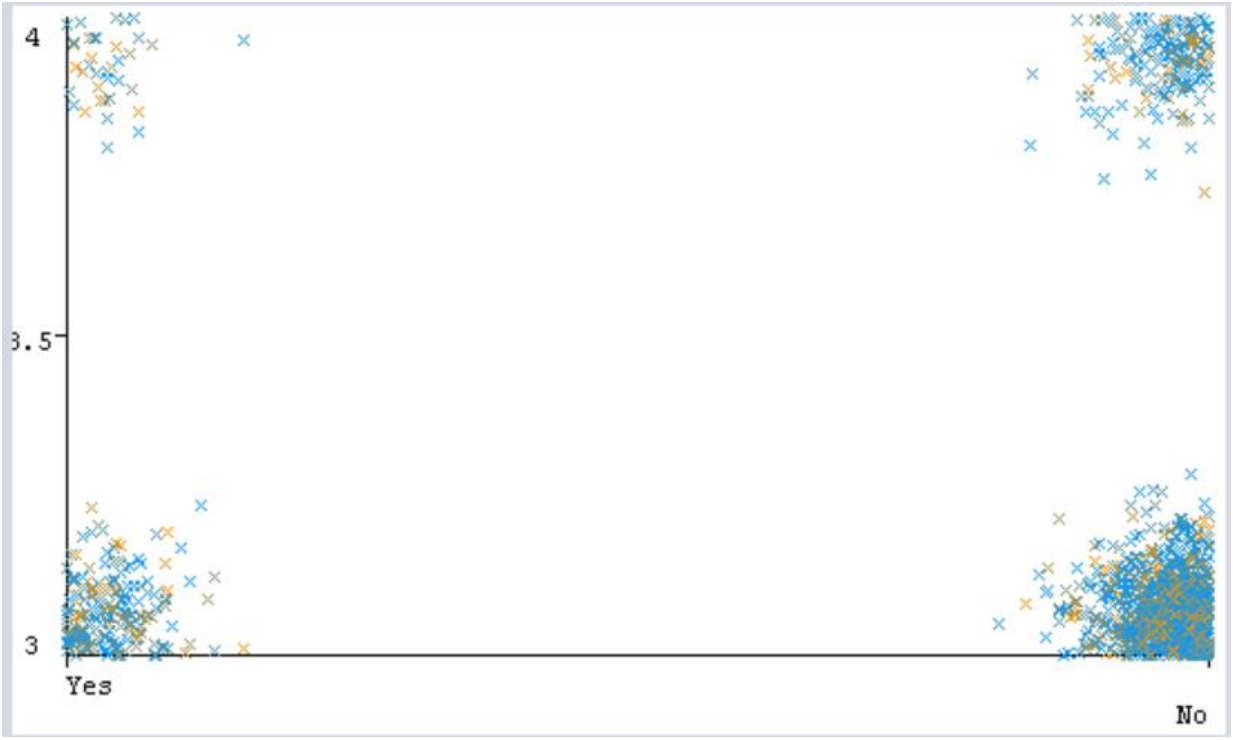

Figure 5. Visualization Distance

From the graph above, it shows that employees with attrition are not too significant with the distance from home to the office, it is shown that the blue color, which means having the proximity of their residence to the office, experiences quite a lot of attrition.

3.3. Data Analysis

For data with attributes that are owned, the prediction process is carried out using the three algorithms. After analyzing the data, the results obtained from the three algorithms will be analyzed for each performance produced. The first scenario used is K-Fold Validation testing by looking for the best accuracy, and the second is analyzing the performance of the algorithm from the precision value.

Scenario 1 : Accuracy value with k-fold

The results of each test are 4 times for each algorithm with the results shown in Figure 6.

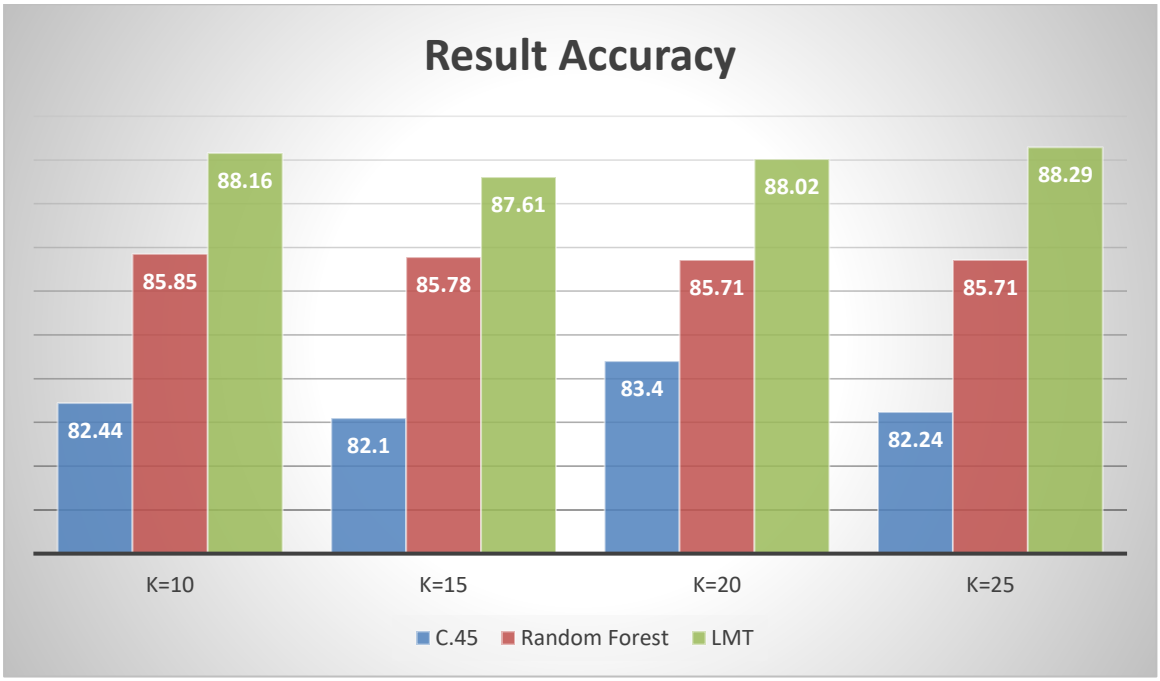

Figure 6. Result Accuracy

From the resulting values, the best results can be visualized for each algorithm in Figure 7.

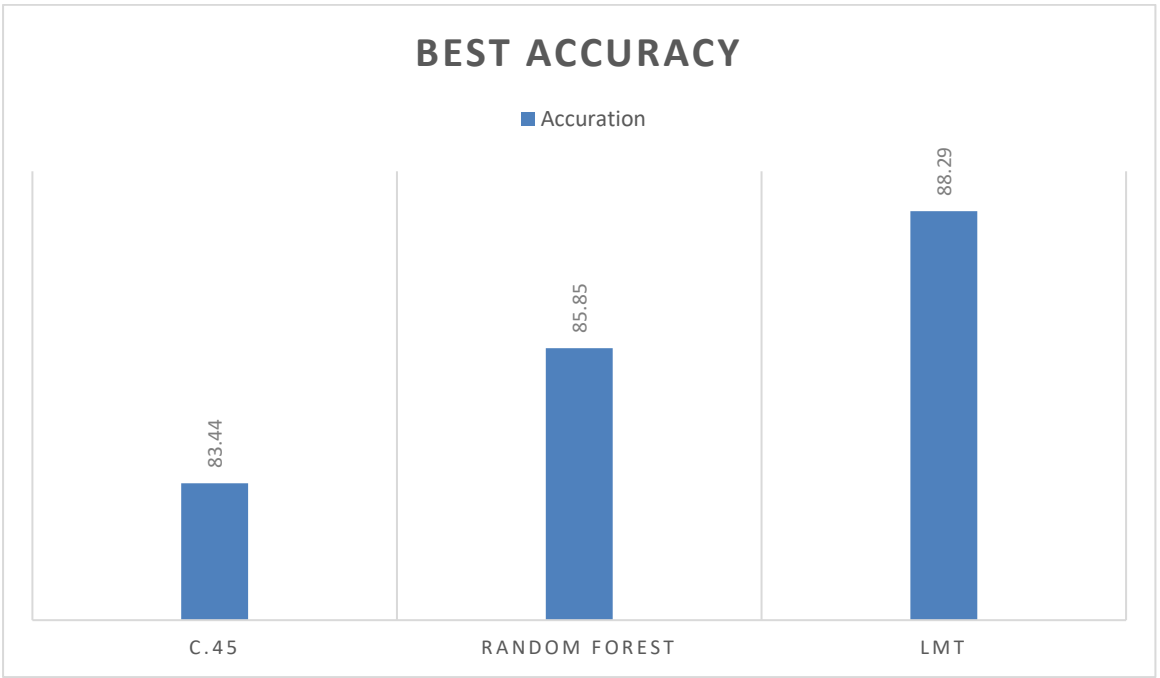

Figure 7. Best Accuracy

Scenario 2 : Precision Value

In this second scenario, the precision value will be observed. Precision is the comparison between True Positive (TP) and the amount of data that is predicted to be positive. The precision value of each algorithm is as follows.

Table 2. C.45

No	Algorithm	K	Precision	Recall	F-Measure
1	C.45	10	0.798	0.824	0.808
2	C.45	15	0.795	0.821	0.805
3	C.45	20	0.810	0.834	0.818
4	C.45	25	0.800	0.822	0.809

The best value is in the third experiment with a value of K = 20 with a precision value of 0.810. It is linear with the best accuracy value on the third try.

Table 3. Random Forest

No	Algorithm	K	Precision	Recall	F-Measure
1	Random Forest	10	0.858	0.859	0.814
2	Random Forest	15	0.852	0.858	0.814
3	Random Forest	20	0.849	0.858	0.816
4	Random Forest	25	0.849	0.857	0.814

The best value in the first experiment with a value of K = 10 with a precision value of 0.858. It is linear with the best accuracy value on the first try.

Table 4. Logistic Model Tree

No	Algorithm	K	Precision	Recall	F-Measure
1	Logistic Model Tree	10	0.872	0.882	0.864
2	Logistic Model Tree	15	0.863	0.876	0.859
3	Logistic Model Tree	20	0.869	0.880	0.864
4	Logistic Model Tree	25	0.873	0.883	0.867

The best value is in the fourth experiment with a value of K = 25 with a precision value of 0.873. It is linear with the best accuracy value on the fourth try.

3.4. Comparative

After carrying out two test scenarios, the last step is to make comparisons and draw conclusions from each experiment that has been carried out. As for the results of the comparison of the two test scenarios, the order of the performance of the three algorithms is shown in Table 5.

Table 5. Result

Algorithm	Rank
Logistic Model Tree (LMT)	1
Random Forest	2
C.45	3

From the experiments that have been carried out, the best accuracy is the Logistic Model Tree (LMT) algorithm, which in every scenario is always superior to the other 2 algorithms.

4. CONCLUSION

The conclusions that can be drawn from this study are as follows:

1. The results of the three algorithms can be used in predicting employee attrition, with different accuracy results, this can help the identification process that can provide an opportunity description based on existing data cases.
2. Using 2 test scenarios, all show the same order as the best accuracy algorithm model in the Logistic Model Tree (LMT)
3. The best accuracy of each algorithm is C.45 = 83.44; Random Forests = 85.85; LMT = 88.29.
4. The precision value in the second scenario is linear with the best accuracy value, the better the accuracy value, the higher the precision value.

5. REFERENCES

[1] KataData, "Ada Gelombang PHK , Ini Startup yang Terbanyak Memecat Karyawan," p. 2022, 2022.

[2] Katadata, "Gelombang PHK Startup Makin Tinggi pada November 2022," no. November, p. 2022, 2022.

[3] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch," J. Phys. Conf. Ser., vol. 1255, no. 1, pp. 1–6, 2019, doi: 10.1088/1742-6596/1255/1/012002.

[4] A. Wibowo, D. Manongga, and H. D. Purnomo, "The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation," Sci. J. Informatics, vol. 7, no. 1, pp. 99–112, 2020, doi: 10.15294/sji.v7i1.24241.

[5] S. Dwiasnati and Y. Devianto, "Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C.45 and Naive Bayes Algorithms," J. Phys. Conf. Ser., vol. 1179, no. 1, 2019, doi: 10.1088/1742-6596/1179/1/012023.

[6] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Syst. Appl., vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.

[7] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random forest for big data classification in the internet of things using optimal features," Int. J. Mach. Learn. Cybern., vol. 10, no. 10, pp. 2609–2618, 2019, doi: 10.1007/s13042-018-00916-z.

[8] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, pp. 19–25, 2017, doi: 10.14569/ijacsa.2017.080603.

[9] K. ; Kirasich, T. ; Smith, and B. Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Sci. Rev., vol. 1, no. 3, p. 9, 2018, [Online]. Available: https://scholar.smu.edu/datasciencereviewAvailableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/9http://digitalrepository.smu.edu.

[10] W. Chen et al., "Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree," Geocarto Int., vol. 34, no. 11, pp. 1177–1201, 2019, doi: 10.1080/10106049.2019.1588393.

[11] F. Aman, A. Rauf, R. Ali, F. Iqbal, and A. M. Khattak, "A Predictive Model for Predicting Students Academic Performance," 10th Int. Conf. Information, Intell. Syst. Appl. IISA 2019, pp. 1–4, 2019, doi: 10.1109/IISA.2019.8900760.

[12] S. A. Fayaz, M. Zaman, and M. A. Butt, "An application of logistic model tree (LMT) algorithm to ameliorate prediction accuracy of meteorological data," Int. J. Adv. Technol. Eng. Explor., vol. 8, no. 84, pp. 1424–1440, 2021, doi: 10.19101/IJATEE.2021.874586.

[13] H. T. Hang et al., "Spatial prediction of landslides along National Highway-6, Hoa Binh province, Vietnam using novel hybrid models," Geocarto Int., vol. 37, no. 18, pp. 5201–5226, 2022, doi: 10.1080/10106049.2021.1912195.

[14] W. Ma, Z. Feng, Z. Cheng, S. Chen, and F. Wang, "Identifying forest fire driving factors and related impacts in china using random forest algorithm," Forests, vol. 11, no. 5, 2020, doi: 10.3390/F11050507.

[15] S. T. Mehdi Zamani Joharestani, Chunxiang Cao, Xilian Ni, Barjeece Bashir, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," Atmosphere (Basel)., no. 1992, pp. 6425–6432, 2019.

[16] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," IEEE Trans. Image Process., vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.

[17] S. Georganos et al., "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," Geocarto Int., vol. 36, no. 2, pp. 121–136, 2021, doi: 10.1080/10106049.2019.1595177.

[18] D. Molina, A. De-La-Hoz, and F. Mendoza, "Classification and features selection method for obesity level prediction," J. Theor. Appl. Inf. Technol., vol. 99, no. 11, pp. 2525–2536, 2021.

[19] M. F. Maulana and M. Defriani, "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period," PIKSEL Penelit. Ilmu Komput. Sist. Embed. Log., vol. 8, no. 1, pp. 39–48, 2020, doi: 10.33558/piksel.v8i1.2018.

[20] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020, 2020, doi: 10.1109/ic-ETITE47903.2020.242.

[21] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[22] C. J. Shiyang Xuan, Gaunjun Liu. Zhenchuan Li, Lutao Zheng, Shuo Wang, "Random Forest for Credit Card Fraud Detection Shiyang," IEEE, 2018.

[23] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject," J. Phys. Conf. Ser., vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012005.

[24] A. Lia Hananto, S. Sofiah Hilabi, and D. Noviani, "Design of Customer Satisfaction Application at BCA Kcp Rengasdengklok Using C.45 Algorithm Method," Buana Inf. Technol. Comput. Sci. (BIT CS), vol. 3, no. 1, pp. 11–16, 2022, doi: 10.36805/bit-cs.v3i1.2048.

[25] A. Suherman, D. Kurnaedi, and R. Darmawan, "Junior Class Preparedness Classification Faces A National Exam Using A C.45 Algorithm With A Particle Swarm Optimization Approach," bit-Tech, vol. 3, no. 1, pp. 11–20, 2020, doi: 10.32877/bt.v3i1.169.

[26] R. Muttaqien, M. G. Pradana, and A. Pramuntadi, "Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office," Int. J. Comput. Inf. Syst., vol. 2, no. 3, pp. 64–68, 2021, doi: 10.29040/ijcis.v2i3.36.

[27] M. G. Pradana and P. H. Saputro, "Komparasi Metode Naïve Bayes Dan C4.5 Dalam Klasifikasi Loyalitas Pelanggan Terhadap Layanan Perusahaan," Indones. J. Bus. Intell., vol. 3, no. 1, p. 20, 2020, doi: 10.21927/ijubi.v3i1.1205.

[28] Y. Alkhalifi, A. Zumarniansyah, R. Ardianto, N. Hardi, and A. E. Augustia, "Comparison of Naive Bayes Algorithm and C.45 Algorithm in Classification of Poor Communities Receiving Non Cash Food Assistance in Wanasari Village Karawang Regency," J. Techno Nusa Mandiri, vol. 17, no. 1, pp. 37–42, 2020, doi: 10.33480/techno.v17i1.1191.

[29] D. Van Le, J. Montgomery, K. C. Kirkby, and J. Scanlan, "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting," J. Biomed. Inform., vol. 86, pp. 49–58, 2018, doi: 10.1016/j.jbi.2018.08.007.