Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix

Ellya Helmud[1]*, Ellya Helmud[2], Fitriyani[3], Parlia Romadiana[4]
Sistem Informasi, Fakultas Teknologi Informasi ISB Atma Luhur[1], [2], [3], [4]
Pangkalpinang, Indonesia
ellyahelmud@atmaluhur.ac.id [1], ellyahelmud@students.undip.ac.id [2], fitriyani@atmaluhur.ac.id [3],
parliaromadiana@atmaluhur.ac.id [4]

Abstract*— The classification method is part of data mining which is used to predict existing problems and also as predictions for the future. The form of dataset used in the classification method is supervised data. The random forest classification method is processed by forming several decision trees and then combining them to get better and more precise predictions. while a decision tree is the concept of changing a pile of data into a decision tree that presents the rules of a decision. From these two classification methods, researchers will compare the level of accuracy of predictions from both methods with the same dataset, namely the employee dataset in India, to predict the level of accuracy of employees who leave their jobs or still remain to work at their company. The number of records available is 4654 records. Of the existing data, 90% was used as training data and 10% was used as test data. From the results of testing this method, it was found that the accuracy level of the random forest method was 86.45%, while the decision tree method was 84.30% accuracy level. Then, by using the confusion matrix, you can see the magnitude of the distribution of experimental validity visually to calculate precision, recall and F1-Score. The random forest algorithm obtained precision of: 96.7%, sensitivity of: 84.7%, specificity of: 91.4%, and F1-Score of: 90.2%. Meanwhile, the decision tree algorithm obtained precision of: 95.7%, sensitivity of: 82.9%, specificity of: 88.4%, and F1-Score of: 88.8%.

Keywords*— Data Mining, Classification, Random Forest, Decision Tree, Confusion Matrix

I. INTRODUCTION

Data Mining is a process of knowledge discovery in databases. Data Mining will be carried out extraction of important information or patterns in large data [1]. Data mining work can be done using prediction, association and segmentation methods [2], where prediction methods are divided into three parts which include classification, regression, and time series. The classification of algorithms used includes : Decision Tree, Random Forest, Neural Network, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, dan GA. Classification has the characteristics of grouping data based on the attachment of data to sample data, where in grouping the data must have label or target attributes. From several algorithms in the classification method, we will compare 2 (two) algorithms in the classification method, namely decision tree and random forest. A decision tree is a basic classifier that has two learning steps and classification [3]. In the learning phase, the learning decision tree generates a decision tree from a set of classified training samples [3]. In the classification phase, the decision tree obtained from the learning phase is used to classify the unclassified data. random forest proposed by [4] There are several decision trees in the random forest classifier, and the average accuracy of these decision trees is used to increase the precision in a collection of several decision trees, taking an estimate of each tree and predicting the final outcome based on majority votes. Based on several previous studies :

a. "Metode Pembelajaran Mesin untuk Memprediksi Persetujuan Pinjaman dengan Membandingkan Algoritma *Random Forest* dan *Decision Tree*" said that the random forest method was found to have a 79.4490 percent success rate in this investigation. When compared to a decision tree
b "Komparasi Tingkat Ketepatan *Random Forest* dan *Decision Tree* C4.5 Pada Klasifikasi Data Penyakit Infertilitas Agung" said that from the test results of these two algorithms, different results were obtained with significant differences in accuracy in the classification of fertility data. The results of random forest testing and Decision tree C4.5 in predicting the success rate and it can be concluded that the test results using random forest obtained an accuracy of 87.20%, then decision tree C4.5 obtained an accuracy of 85.90%. [5]
c. "Komparasi Algoritma *Random Forest* Dan *Decision Tree* Untuk Memprediksi Keberhasilan Immunotheraphy" [5] said that with testing using cross-validation, an accuracy value of 84.4% was obtained using the decision tree method and then an accuracy value of 85.5% using the

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024

random forest method. Using employee datasets in India to predict employees who are still working or out of work. After obtaining the results of the prediction of the level of accuracy from these two algorithms, then the calculation of accuracy, precision, recall and F1-Score will be carried out to see the distribution of experimental validity visually using a confusion matrix. The dataset used was 4654 records, and then disaggregated for training data and trial data with a ratio of 9 : 1

II. RESEARCH METHODOLOGY

This study was conducted to compare the performance level of random forest and decision tree algorithms using employee datasets in India, namely predicting employees who leave their jobs or still stay in an agency/company. The stages carried out to compare these two methods can be illustrated in the following figure.

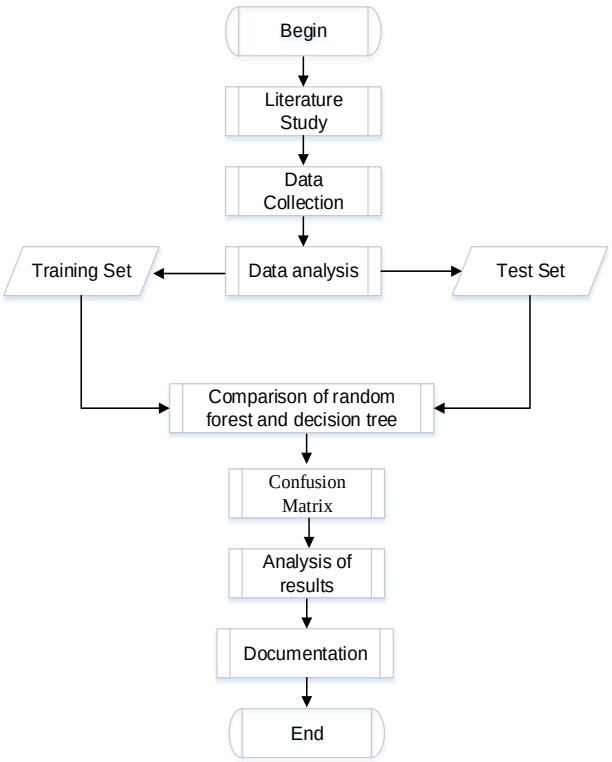

Figure 1
Research Methods

2.1. Literature Study

Research should refer to a literature study. Literature study which is a series of activities that include collecting library data, reading and recording, and managing research materials. Literature study is an activity that must be carried out in research, more specifically research related to academics where the goal is to explore the conceptual and technical sides. Literature studies are conducted by researchers with the intention of obtaining basic theories, frameworks of thinking, and obtaining temporary estimates called research hypotheses. Thus researchers can create groups, allocate organize, and various library developments to reset their research. The existence of literature studies allows researchers to deeply and broadly dig deeper into the problem to be studied. With literature studies researchers can determine the topic in the research, the problem formulated then they can go directly to the intended research place for the required data collection.

2.2 Data Collection

When data collection activities are carried out, namely searching for datasets which are data sources used to compare classifications. The dataset taken for this research is in the form of public data sourced from https://www.kaggle.com/ website, namely the employee dataset. This dataset contains employee data that at the time of testing will influence the employee to leave his job or still work. The data is 4654 records where there are 8 attributes and 1 attribute used as a target label. Before the dataset is processed for testing, it needs to be looked at (data preparation).

Data preparation is a stage for cleaning data as well as transforming raw data before processing and analyzing which includes all the steps needed to obtain, prepare, precise, and supervise data resources in the organization. It is important for researchers to do before the data is processed and format the data, check the data first, and combine collection data for data exploration. Before analyzing the data, researchers need to collect data from various sources, delete, or fill in Null data, duplicate data, or update data into the correct rules. Training data is a subset of datasets used to train machine learning models.

The goal is to study patterns and relationships between features (independent variables) and targets (dependent variables).Data training measures usually have a larger proportion than testing data. Most datasets are allocated for model training. The model process learns through iterative iterations (epochs) of training data, adjusting its internal parameters to achieve good performance on this data. Data testing is a subset of datasets that are not used during the model training process and are stored to test model performance after training. The goal is to measure the extent to which the model can generalize information from data not seen during training.

The size of the testing data is usually a small part of the overall dataset, but it is representative of the overall data distribution. The process after the model is trained, its performance is tested on testing data. This helps assess the extent to which the model can cope with new data and provides good predictions. The division between training data and testing data is usually done randomly. The comparison used in this study was 9:1 The Importance of Division Separating the data into these two subsets is important to avoid overfitting. If the model is only treated on training data, it may only memorize that data and may not generalize well to the new data.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024

The following sample dataset has been prepared to test the accuracy of prediction of random forest and decision tree methods, can be seen in the figure below :

	A	B	C	D	E	F	G	H	I
1	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
2	Bachelors	2017	Bangalore	3	34	Male	No	0	0
3	Bachelors	2013	Pune	1	28	Female	No	3	1
4	Bachelors	2014	New Delhi	3	38	Female	No	2	0
5	Masters	2016	Bangalore	3	27	Male	No	5	1
6	Masters	2017	Pune	3	24	Male	Yes	2	1
7	Bachelors	2016	Bangalore	3	22	Male	No	0	0
8	Bachelors	2015	New Delhi	3	38	Male	No	0	0
9	Bachelors	2016	Bangalore	3	34	Female	No	2	1
10	Bachelors	2016	Pune	3	23	Male	No	1	0
11	Masters	2017	New Delhi	2	37	Male	No	2	0
12	Masters	2012	Bangalore	3	27	Male	No	5	1
13	----------								
4647	Masters	2017	Pune	2	31	Female	No	2	0
4648	Bachelors	2013	Bangalore	3	25	Female	No	3	0
4649	Bachelors	2016	Pune	3	30	Male	No	2	0
4650	Bachelors	2013	Bangalore	3	26	Female	No	4	0
4651	Masters	2013	Pune	2	37	Male	No	2	1
4652	Masters	2018	New Delhi	3	27	Male	No	5	1
4653	Bachelors	2012	Bangalore	3	30	Male	Yes	2	0
4654	Bachelors	2015	Bangalore	3	33	Male	Yes	4	0

Figure 2
Dataset Employee

2.3 *Random Forest*

Random Forest is a supervision learning algorithm where there is training data and also testing data proposed by Breiman in 2001 (Louppe, 2014) [6]. Random forests are commonly used to solve problems that have to do with classification, regression, These algorithms build multiple decision trees during training and combine the results to improve the performance and robustness of the model and others. Why this algorithm is called random is because :

a. Each tree grows on a different bootstrap instance, taken from random training data.
b. In each split node during decision tree formation, a sample portion of the variable m is selected from its original data collection after which it will best be used in that node.

This algorithm is a combination of several tree predictors or can be called decision trees where each tree relies on random vector values to be used as free and equitable examples on all trees in the forest [7]. The prediction results from random forests are obtained from the most results from each decision tree (voting for classification and average for regression). For RF consisting of N trees it is formulated as:

$$l(y) = argmax_c(\sum_{n=1}^{n} Ih_n(y) = c) \quad \text{... (2.1)}$$

2.4 *Decision tree*

A decision tree is a predictive model that maps decision problems based on a set of conditional decisions. This model can be used for classification and regression tasks. Decision trees take the form of hierarchical structures similar to trees, with each node representing decisions based on certain features. Decision trees are used to form decision trees with very strong predictions. The decision tree method converts very large facts into decision trees that represent rules that can be easily understood [8]. The stages in the decision tree are:

Calculate the entropy value of each attribute :

a. Calculate the entropy value of each attribute :

$$Entropy(S) \ \sum_{i=1}^{n} -pi * log_2 pi \quad \text{... (2.2)}$$

b. Calculation of the value of obtaining information on each attribute :

$$Info\ Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{[S1]}{|S|} * Entropy\ (S_1) \quad \text{... (2.3)}$$

c. Calculation of separate information values for each attribute:

$$SplitInfo_a(D) = -\sum_{j=1}^{n} \frac{[D1]}{|D|} * log_2(\frac{Dj}{D}) \quad \text{... (2.4)}$$

d. Calculate the value of the gain ratio for each attribute

$$Gain\ Ratio(A) = \frac{InfoGain\ (A)}{SplitInfo\ (A)} \quad \text{... (2.5)}$$

e. The attribute that has the highest profit ratio is selected to be the root (splitting attribute) and the attribute that has a profit ratio value lower than the root (root) is selected to be a branch.
f. Recalculate the value of the profit ratio of each attribute by excluding the attribute selected as root in the previous step
g. Attributes that have the highest Gain Ratio will be made branches. Then iterations for steps 4 and five obtained a value of Gain of 0 from all remaining attributes

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024

2.5 *Confusion Matrix*

Calculation of classification performance from each method test with Confusion Matrix to obtain accuracy, precision, and recall results. The confusion matrix is used to obtain an estimate of how well classified unequal class detection [19]. The Confusion Matrix table can be seen in the Table

Table 1.
Table Confusion Matrix

Actual Value	Assigned classes	
	Positive	Negative
positive	True Positive	False Negative
Negative	False Positive	True Negative

Sokolova's opinion (2009) obtained several matrices to be calculated in the confusion matrix, namely accuracy, precision, recall, specificity, and F1-Score. In the calculation obtained the formula :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \dots (2.6)$$

$$Precision = \frac{TP}{TP+FP} \quad \dots (2.7)$$

$$Recall = \frac{TP}{TP+FN} \quad \dots (2.8)$$

$$Specificity = \frac{TN}{TN+FP} \quad \dots (2.9)$$

$$F1-Score = \frac{2TP}{2TP+FP+FN} \quad \dots (2.10)$$

III. DISCUSSION

3.1. The testing process and results of testing random forest and decision tree algorithms using the rapid miner tool are as follows :

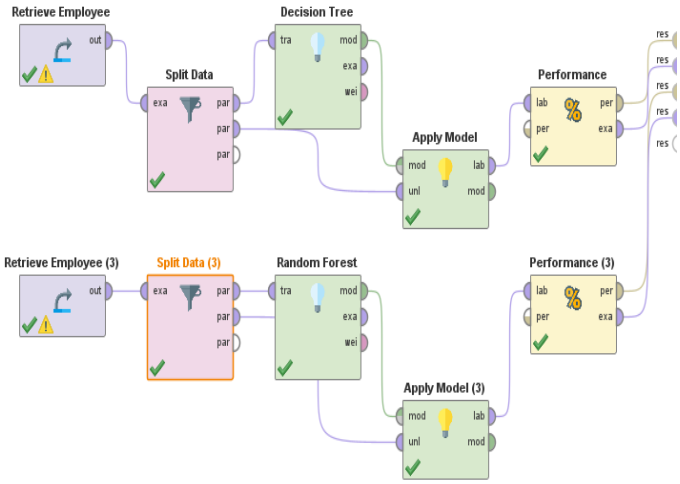

Figure 3
Testing Process of random forest and decision tree algorithms

Row...	LeaveOrNot	prediction...	confidence(Leave)	confidence(Not)	Education	Joining...	City	PaymentTier	Age	Gend...	EverBenched	Exp
1	Not	Not	0.281	0.719	Masters	2016	Bangalore	3	27	Male	No	5
2	Leave	Leave	0.953	0.047	Bachelors	2016	Pune	3	23	Male	No	1
3	Leave	Leave	0.752	0.248	Masters	2017	New Delhi	2	28	Male	No	4
4	Leave	Leave	0.801	0.199	Bachelors	2014	Bangalore	1	30	Female	No	3
5	Leave	Leave	0.939	0.061	Bachelors	2014	Pune	3	30	Male	Yes	4
6	Leave	Leave	0.744	0.256	Masters	2013	New Delhi	3	35	Male	No	2
7	Leave	Leave	0.873	0.127	Bachelors	2016	Bangalore	3	34	Male	No	4
8	Not	Not	0.490	0.510	Masters	2017	Bangalore	3	40	Female	No	2
9	Not	Not	0.024	0.976	Bachelors	2018	Bangalore	3	23	Female	No	1
...	...	...	...	...	...	...	...	...	...	...	...	...
461	Leave	Leave	0.829	0.171	Bachelors	2015	Bangalore	3	28	Male	No	1
462	Not	Not	0.007	0.993	Bachelors	2018	Bangalore	3	31	Male	No	0
463	Not	Leave	0.768	0.232	Bachelors	2012	Bangalore	3	23	Female	No	1
464	Leave	Leave	0.873	0.127	Bachelors	2013	Bangalore	3	34	Male	No	4
465	Leave	Leave	0.844	0.156	Bachelors	2012	Bangalore	3	39	Female	No	1

Figure 4
Prediction Results from the Random Forest Algorithm

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024

Row...	LeaveOrNot	prediction...	confidence(Leave)	confidence(Not)	Education	JoiningYear	City	PaymentTier	Age	Gend...	EverBench...	Experien
1	Leave	Leave	0.882	0.118	Bachelors	2016	Bangalore	3	39	Male	No	2
2	Leave	Leave	0.882	0.118	Bachelors	2012	Bangalore	3	37	Male	No	0
3	Leave	Leave	0.882	0.118	Bachelors	2015	Bangalore	3	23	Male	No	1
4	Leave	Leave	0.816	0.184	Masters	2017	New Delhi	2	30	Female	No	2
5	Leave	Leave	0.814	0.186	Bachelors	2014	Bangalore	1	30	Female	No	3
6	Leave	Leave	0.816	0.184	Masters	2017	New Delhi	2	34	Male	No	0
7	Not	Not	0	1	Bachelors	2015	Pune	2	26	Female	No	4
8	Leave	Leave	0.814	0.186	Bachelors	2014	Bangalore	1	23	Female	No	1
9	Leave	Leave	0.821	0.179	Bachelors	2017	Bangalore	3	23	Female	No	1
⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣	⎣
461	Leave	Leave	1	0	Masters	2017	New Delhi	3	40	Male	No	2
462	Leave	Leave	0.882	0.118	Bachelors	2016	Pune	3	36	Male	Yes	3
463	Not	Not	0	1	Bachelors	2014	New Delhi	2	23	Female	No	1
464	Leave	Leave	1	0	Masters	2015	Bangalore	3	38	Female	No	1
465	Not	Not	0.193	0.807	Bachelors	2012	Pune	3	25	Female	No	3

Figure 5
Prediction Results from Decison Tree Algorithm

From employee data available in India, the results of algorithm classification predictions using Random Forest with a total of 4654 records in spit for training data of 90% and testing data of 10%. shows the accuracy level of the random forest algorithm of 86.45%. While the accuracy level of the decision tree algorithm is 84.30%.

3.2 Then the results of the large distribution of validity of the random forest algorithm and decision tree using the confusion matrix to determine precision, recall and F1-score for each method are as follows:

accuracy: 86.45%

	true Leave	true Not	class precision
pred. Leave	295	53	84.77%
pred. Not	10	107	91.45%
class recall	96.72%	66.88%	

Figure 6
Confusion Matrix from Random Forest Alforithm

accuracy: 84.30%

	true Leave	true Not	class precision
pred. Leave	292	60	82.95%
pred. Not	13	100	88.50%
class recall	95.74%	62.50%	

Figure 7
Confusion Matrix from the Decision Tree Algorithm

IV. CONCLUSION

From the test results of these 2 classification algorithms, it was found that

The accuracy level of the random forest algorithm is 86.45% higher than the decision tree algorithm which is 84.30%. Precision : The model has the ability to classify data of employees who leave their jobs correctly when the model is predicted for the Random Forest algorithm by 96.7% and the decision tree algorithm by 95.73%. Recall : The model has the ability to classify data of employees who leave their jobs correctly for random forest algorithm by 84.7% and decision tree algorithm by 82.9% Specificity : The model has the ability to correctly classify employee data that continues to work in the company for random forest algorithm by 91.4% and decision tree algorithm by 88.4%. F1-Score : A measure of the balance between precision and recall for the random forest algorithm at 90.2% and the decision tree algorithm at 88.8%

Because it can be concluded from the results of the testing process that the algorithm uses random forest better than the decision tree algorithm

REFERENCES

[1] P. Han, Kamber, *Data Mining Concepts and Techniques*. 2012.

[2] F. A. Hermawati, "*Data Mining*," no. January, 2018.

[3] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, "ScienceDirect A comparative study of decision tree , random forest , and convolutional neural network for spread-F identification," *Adv. Sp. Res.*, vol. 65, no. 8, pp. 2052–2061, 2020, doi: 10.1016/j.asr.2020.01.036.

[4] L. Breiman, "Random Forest," pp. 1–33, 2001.

[5] A. Prabowo, S. Wardani, R. W. Dewantoro, and W. Wesly, "Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas," vol. 4, no. 1, pp. 218–224, 2023, doi: 10.30865/klik.v4i1.1115.

[6] C. Science, "U niversity of L ième," no. July, 2014.

[7] C. Curtis, C. Liu, T. J. Bollerman, and O. S. Pianykh, "Machine Learning for Predicting Patient Wait Times and Appointment Delays," *J. Am. Coll. Radiol.*, no. Ml, pp. 1–7, 2017, doi: 10.1016/j.jacr.2017.08.021.

[8] P. Bhargav and K. Sashirekha, "A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms .," vol. 10, pp. 1803–1813, 2023.

[9] N. Sunanto and G. Falah, "Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes," Rabit J. Teknol. dan Sist. Inf. Univrab, vol. 7, no. 2, pp. 208–216, 2022, doi: 10.36341/rabit.v7i2.2435.

[10] R. Estian Pambudi, Sriyanto, and Firmansyah, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45," Ijccs, vol. x, No.x, no. x, pp. 1–5, 2022.

[11] M. Ardiansyah, A. Sunyoto, and E. T. Luthfi, "Analisis Perbandingan Akurasi Algoritma Naïve Bayes Dan C4.5 untuk Klasifikasi Diabetes," Edumatic J. Pendidik. Inform., vol. 5, no. 2, pp. 147–156, 2021, doi: 10.29408/edumatic.v5i2.3424.

[12] Saifullah, Muhammad Zarlis, Zakaria Zakaria, Rahmat Widia Sembiring, "Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Prosesing Data," J-SAKTI (Jurnal Sains Komputer & Informatika), Vol 1, No 2 (2017)

[13] Svetnik V 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling J. Journal of Chemical Information & Computer Sciences 1 43

[14] Alvita Izana Kusumarini, Pandu Ananto Hogantara, Muammar Fadhlurohman, Nurul Chamidah, "Perbandingan Algoritma Random

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024

Forest, Naive Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E.Coli," Prosiding SENAMIKA, Vol 2, No 1 (2021)

[15] Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath " Loan default prediction using decision trees and random forest: A comparative study " 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India

[16] Simon Heqelich, " Decision Trees and Random Forests : Machine Learning Techniques to Classify Rare Events" Vol 2, Issue 1 Spring 2016

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1985, Copyright ©2024
Submitted : October 6, 2023, Revised : January 4, 2024, Accepted : January 11, 2024, Published : February 15, 2024