byAccredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 6 (2023) 1348 - 1353 ISSN Media Electronic: 2580-0760

Comparison of the Accuracy of Drug User Classification Models Using Machine Learning Methods

Nursela Salsabilla Basuni[1], Amril Mutoi Siregar[2]
[1,2]Department of Informatics, Faculty of Computer Science, Buana Perjuangan University, Karawang, Indonesia
[1]if20.nurselabasuni@mhs.ubpkarawang.ac.id, [2]amrilmutoi@ubpkarawang.ac.id

Abstract

Caseand s of drug abuse are on the rise, with many users entering the addiction phase, often resulting in overdose and death. Drugs are chemical compounds that are capable of affecting biological functions, can induce feelings of happiness and reduce pain. To address this growing problem, a proactive measure is needed. Therefore, this study aims to classify drug users and non-users, so that health workers and therapists can educate about the dangers of drugs to non-users and rehabilitate drug users. This study uses drug consumption data taken from the UCI Irvine Machine Learning Repository. The data consists of 1885 rows with 32 attributes and 2 classes, where there are 18 types of legal and illegal drugs. This research utilizes machine learning methods, specifically Artificial Neural Network (ANN), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF), in addition to evaluation methods such as Confusion Matrix and Area Under Curve (AUC). The results showed that RF outperformed the other methods, with accuracy, precision, and recall of 93%, and an f1-score of 89%, while the AUC value was still suboptimal at 0.66. DT had the worst results, with 82% accuracy, 87% precision, 82% recall, 84% f1-score, and an AUC value of 0.56. With these results, this research can be continued into an application that can classify drug users and non-users.

Keywords: drug consumption; classification; machine learning; confusion matrix; AUC curve

1. Introduction

Drugs are natural, synthetic, or semi-synthetic substances that can cause changes in consciousness, hallucinations, and stimulation. Drug addiction is generally classified into four stages, namely occasional use, recreational use, regular use, and addiction, and can be greatly affected by emotions, consciousness, and cognition [1]. In 2017, the Centers for Disease Control and Prevention (CDC) reported 70,237 drug overdose deaths and attributed the increase in drug cases to increased accessibility and promotion through social media. Currently, several online platforms offer drugs at discounted prices [2]. Drugs are chemical compounds capable of affecting biological functions, with psychoactive drugs specifically impacting a person's mental state, often producing pleasurable effects, and reducing the user's pain or discomfort. Many factors that make teenagers often fall into the world of drugs include family economic problems, lack of love from the family, and wrong associations [3].

Drug abuse is prevalent worldwide, with excessive use often leading to addiction and even fatal overdose incidents. To avoid fatal cases, it is crucial to classify individuals as drug users and non-users. This classification can provide doctors and therapists with an important tool to educate the general public about the dangers of drug consumption. Failure to address abuse will result in more users, and consequently overdose deaths. Various machine learning methods, namely Artificial Neural Network (ANN), Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF) can be used to handle classification tasks. Several previous studies have successfully used these methods in classifying drug administration routes, with [4] achieving a 97% accuracy rate via Random Forest using drug compound data retrieved from ChEMBL. Similarly, [5] classified drugs that cause QT syndrome by investigating ECG reports using SVM and KNN, resulting in 89% accuracy using ECG data from Physionet. [6] categorized cancer drugs using Logistic Regression, DT, ANN, RF, and Multi-Layer Perceptron, which specifically showed higher accuracy results. In addition, [7] found that K-NN outperformed Naive Bayes when comparing the two methods for drug molecule classification using biochemical data taken from PubChem. Based on the problems described, this

Accepted: 01-09-2023 | Received in revised: 18-11-2023 | Published: 15-12-2023

research aims to classify drug users and non-users using various machine-learning methods. This research also aims to improve and compare the accuracy results of the algorithms used in previous research using different data.

2. Research Methods

2.1 Dataset

The dataset used in this study was sourced from the UCI Irvine Machine Learning Repository https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified, comprising 1885 rows with 32 attributes and 2 distinct classes. The 32 attributes consisted of id, age, gender, education, country, Ethnicity, Nscore (NEO-FFI-R Neuroticism), Escore (NEO-FFI-R Extraversion), Oscore (NEO-FFI-R Openness to experience), Ascore (NEO-FFI-R Agreeableness), Cscore (NEO-FFI-R Conscientiousness), Impulsivity (impulsivity as measured by BIS-11), SS (sensation of seeing as measured by ImpSS), Alcohol, Amphet (amphetamine), Amyl (amyl nitrite), Benzos (benzodiazepine), Caff (caffeine), Cannabis (cannabis), Choc (chocolate), Cocaine, Crack, Ecstasy, Heroin, Ketamine, Legalh (illicit drug), LSD (alcohol), Meth (methadone), Mushroom (Magic mushroom), Nicotine, Semer (Semeron fictitious drug), and VSA (volatile substance abuse) consumption classes. The dataset further categorized individuals into users and non-users. This study used 5 machine learning methods namely, ANN, DT, KNN, SVM, and RF. Figure 1 presents the various stages.

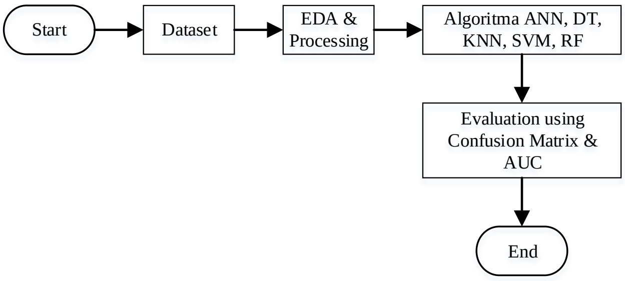
Start
Dataset
EDA & Processing
Algoritma ANN, DT, KNN, SVM, RF
Evaluation using Confusion Matrix & AUC
End

Figure 1. Study Stages

2.2 Exploratory Data Analysis (EDA)

EDA is a process of conducting preliminary investigations on data to identify patterns, and anomalies, test hypotheses, and verify assumptions using summary statistics [8]. It primarily aims to discern patterns within large datasets by reducing their dimensions and utilizing visualization techniques [9]. EDA can be grouped into graphical and non-graphical, as well as univariate and multivariate. Univariate focuses on a single variable, while multivariate includes several variables [10]. The following are various steps of EDA: observing the dataset, searching for missing values in the data, categorizing data into numerical and categorical variables, identifying relationships between variables, and detecting outliers and anomalies in data.

2.3. Artificial Neural Network (ANN)

ANN is an artificial intelligence computational network, whose design methods draw inspiration from the biological structure of the human brain [11]. It typically consists of the input, hidden, and outer layers, which are respectively interconnected with i-th, j-th, and k-th nodes [12].

2.4 Decision Tree (DT)

DT is used for deciding or analyzing relevant attribute information and corresponding classification results in a dataset [13]. It is comprised of decision nodes, which are responsible for making decisions with multiple branches, and leaf nodes in the form of output stemming from decision nodes, lacking branches. The initial node in DT, which subsequently branches out, is referred to as the root node can be seen in Formula 1 [14].

$$Entropy(S) = \sum_{i=1}^{n} - pi \, . log2 \; pi \qquad (1)$$

S represents the set of cases, n denotes the number of partitions of S, and pi signifies the proportion of Si to S. Formula 2 is used for calculating the gain:

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \ Entropy\ (Si) \qquad (2)$$

S stands for the set of cases, A denotes the features, n represents the number of partitions of attribute A, |Si| signifies the proportion of Si to S, and |S| indicates the number of cases in S. ID3, CART, and C4.5 algorithms can be used to build decision trees [15].

2.5 K-Nearest Neighbor

KNN, a method for classification tasks, operates by determining unknown values based on their proximity to neighboring data points, with the Euclidean distance being a prevalent choice. It is known for ease of implementation and effectiveness, even with a relatively large dataset. Also, it is used for calculating the k value, which can either be a specified constant or a randomized variable [16] and exhibits tolerance for noise. Each neighbor is assigned a weight using the similarity equation and d0 as shown in Formula 3.

$$Score(d0, Ci) = \sum_{dj \in KNN(d0)} Sim(d0, dj) \delta(dj, Ci) \qquad (3)$$

DOI: https://doi.org/10.29207/resti.v7i6.5401

KNN(d) is the closest K-neighbor set from document d0, δ(dj,Ci) denotes the classification for dj documents related to class Ci can be seen in Formula 4.

$$\delta(dj,Ci) = \begin{cases} 1\ dj\ \epsilon\ Ci \\ 0\ dj\ \epsilon\ Ci \end{cases} \quad (4)$$

Formula 5 is used to make the final decision with KNN:

$$C\ =\ arg\ max_{ci}\ arg\ max \quad (5)$$

2.6 Support Vector Machine (SVM)

SVM is used for solving classification and regression tasks, primarily mapping data samples into a feature space through a kernel function and subsequently classifying them using hyperplane [17]. It is divided into kernel and simple SVM, which is commonly used for solving classification problems can be seen in Formula 6.

$$Large\ f(x) = sign(sum_{i}^{n} = 1y_i\ alpha_i K(x_i,x)\ +\ b) \ (6)$$

Where f(x) stands for the prediction function, x is the input feature vector, y denotes the class label (+1 or -1), α represents the weight vector, K(xi,x) is the kernel function for calculating the distance between two vectors, and b is the bias.

2.7 Random Forest (RF)

RF is an ensemble tree-based method that uses DT and bagging as base learning. This method is primarily used for classification tasks and operates by randomly selecting N samples from the training data [18]. The decision tree Formula 7 is formulated before RF:

$$dpi\ 150\ large(x) = sum_{i}^{m} = 1w_i h_i(x) \quad (7)$$

f(x) represents the output of DT, m signifies the number of nodes, wi corresponds to the weights of each node, and hi(x) is the function yielding a value of 0 or 1. RF is formulated as Formula 8.

$$dpi\ 150\ large\ F(x) = frac1Msum_{i}^{M} = 1\ f_i(x) \quad (8)$$

f(x) is the output of RF, M denotes the number of DT, and fi(x) represents the output of the i-th DT.

2.8 Confusion Matrix

ConfA confusionrix is used for evaluating the performance of a classification method and comprises four terms, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as a representation of the classification results. TN denotes correctly identified negative values, while FP signifies negative values incorrectly identified as positive [19]. This model is used for calculating accuracy, precision, recall, and F1-score values. Precision reflects the comparison between the correctly predicted positives and sethe of positive values, while accuracy measures the model performance [20]. The confusion matrix Formula 9 - 12 (9), (10), (11), are presented as follows:

$$Accuracy\ = \frac{TP+TN}{TP+FP+FN+TN} \quad (9)$$

$$Precision\ = \frac{TP}{TP+FP} \quad (10)$$

$$Recall\ = \frac{TP}{TP+FN} \quad (11)$$

$$F1-Score\ = \frac{2\ x\ Recall\ x\ Precision}{Recall\ x\ Precision} \quad (12)$$

2.9 Area Under Curve (AUC)

AUC is an evaluation metric used in classification tasks, particularly at various threshold settings. It measures the degree of separation and evaluates the accuracy of the model in distinguishing between classes. To obtain the AUC value by summing the area under the ROC curve, if the lthe are the more accurate the classification result [21]. Higher AUC values indicate higher performance in correctly predicting 0 as 0 and 1 as 1 [22]. Meanwhile, a value close to 0 indicates a suboptimal model, an and a value close to 1 indicates a well-performing model [23]. AUC is used to calculate the difference in algorithm performance [24].

3. Results and Discussions

3.1 Results

EDA was conducted to identify potential missing values and duplicates. Visualization was subsequently carried out to draw valuable insights before proceeding with further data processing. Attributes, such as Choc, Semer, and Caff, which did not contribute to accurate answers were discarded. The visualization data presented in Figure 2 indicated 200 illegal, over 1600 depressant, 1000 stimulant, and 200 psychotropic drug users.

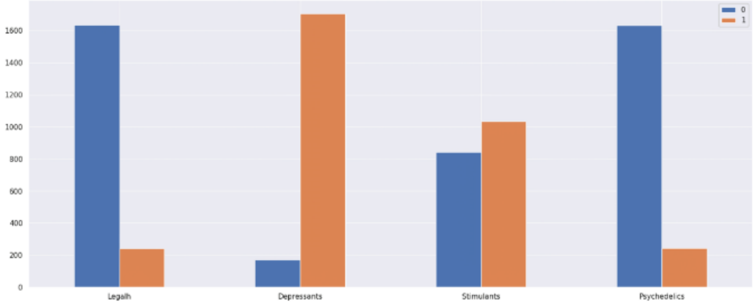

Figure 2. Number of People Using and Not Using Drugs

The subsequent process included label encoding of converting categorical data into a numerical format for enhanced comprehension and ease of processing. The data were subsequently classified into training and testing, while the experiment was repeated three times with varying data split ratios of 70 to 30, 80 to 20, and 90 to 10. The method was also evaluated using a confusion matrix and AUC. Tables 1, 2, and 3 present the comparison of accuracy, precision, recall, and f1-score results across three trials, using split data of 90 to 10, 80 to 20, and 70 to 30, respectively. Figure 3 shows a bar chart comparing the results obtained from ANN, DT, KNN, SVM, and RF. Based on this figure, it can be concluded that the RF and SVM algorithms as a whole have superior results compared to other algorithms.

Table 1. Accuracy Results with 90 to 10 Data Ratio

Methods	Accuracy	Precision	Recall	F1-Score
ANN	84%	80%	84%	82%
DT	83%	84%	83%	84%
K-NN	87%	81%	87%	84%
SVM	89%	81%	89%	85%
RF	90%	81%	90%	85%

Table 2. Accuracy Result with 80 to 20 Data Ratio

Methods	Accuracy	Precision	Recall	F1-Score
ANN	88%	85%	88%	87%
DT	83%	85%	83%	84%
K-NN	90%	86%	90%	87%
SVM	91%	84%	91%	88%
RF	92%	84%	92%	88%

Table 3. Accuracy Result with 70 to 30 Data Ratio

Methods	Accuracy	Precision	Recall	F1-Score
ANN	85%	86%	85%	86%
DT	82%	87%	82%	84%
K-NN	89%	86%	89%	88%
SVM	92%	86%	92%	89%
RF	93%	93%	93%	89%

Figure 4 shows that the ROC curve produces a large empty area above the curve away from the value of 1.0 at the true positive rate. In addition, the AUC of the ANN model produced is 0.52, this result is close to 0.5. In AUC, if the result is close to 0.5, it is called random, which means that the model can not separate positive and negative.

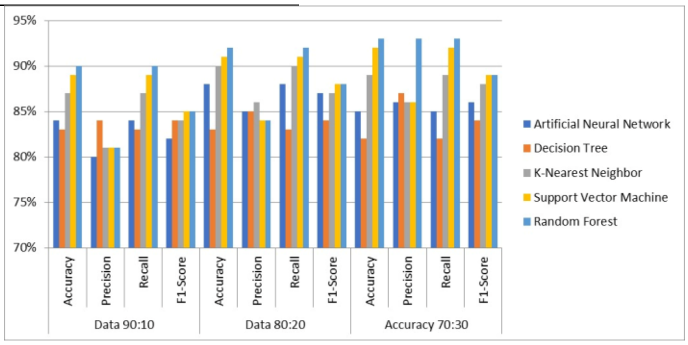

Figure 3. Comparison of Accuracy Results

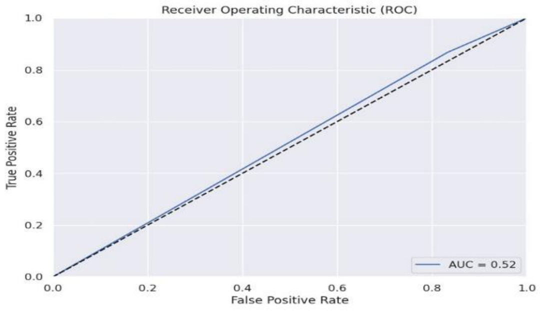

Figure 4. AUC of ANN

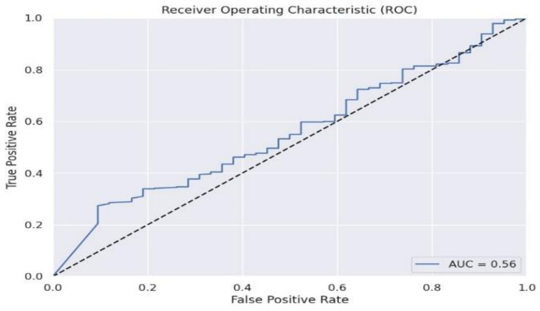

Figure 5. AUC of DT

Figures 5, 6, 7, and 8 also show that the resulting AUC curve is close to 0.5. Figures 4, 5, 6, 7, and 8 present AUC results for five methods, using a data split of 70 to 30. Moreover, a 0,52 AUC value was obtained for ANN, with 0,56 for DT, 0,50 for KNN, 0,57 for SVM, and 0,66 for RF. These values all fell within the weak category as they ranged from >0,50 to 0,60. Although Tables 1, 2, and 3 indicated satisfactory accuracy results, Figures 4, 5, 6, 7, and 8 showed that AUC results for the models used were less favorable.

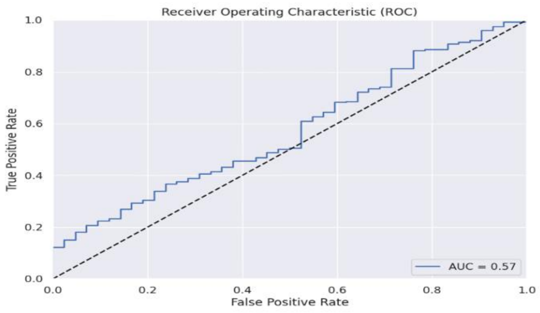

Figure 6. AUC of KNN

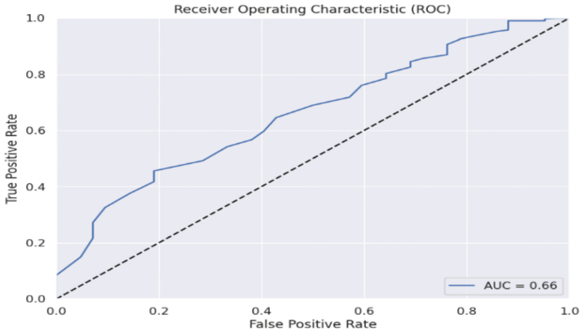

Figure 7. AUC of SVM

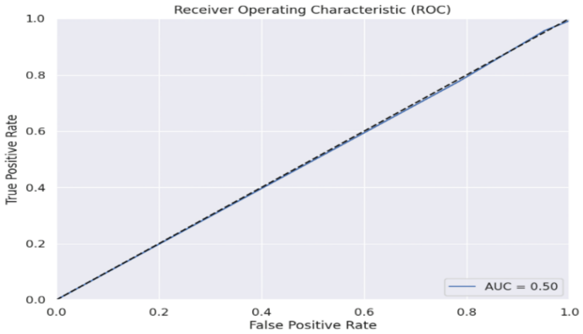

Figure 8. AUC of RF

3.2 Discussions

The performance of neural network and decision tree algorithms seems to be less suitable for handling classification cases when compared to statistical algorithms that show better accuracy results. For the decision tree algorithm, RF has superior results compared to DT. This makes statistical methods more suitable for handling categorical and numerical data classification tasks. The results showed that KNN had an accuracy of 89%, SVM reached 92%, and RF 93%. Precision, recall, and f1-score outperformed ANN, which only had 85% accuracy, and DT, with 82%. This research obtained better results than previous research because this research can improve the accuracy results of the SVM and KNN algorithms. This research also added another evaluation model besides the confusion matrix, namely the AUC Curve. However, the author has not been able to find previous research using drug consumption data, so this research cannot be compared using the same data.

4. Conclusion

In conclusion, it turns out that the classification using drug consumption data carried out by the author has good results, it can improve the accuracy results of previous studies and can add other evaluation models. Evaluation of various machine learning methods showed significant differences between accuracy and AUC values. While the accuracy metric looks promising, the AUC results show room for improvement. RF in particular emerged as the most successful, with impressive accuracy, precision and recall of 93% and f1-score of 89%. In addition, RF also had the greatest AUC value at 66%. DT had the worst results, with an accuracy of 82%, precision of 87%, recall of 82%, and f1-score of 84%. However, SVM has the smallest AUC value of 50%. With the superior results of RF, in the future, this research can be continued by implementing the RF algorithm into an application that can classify drug users and non-users. Future research is also recommended to explore alternative evaluation methods to achieve more accurate and reliable results, as well as better data processing methods.

References

[1] J. Feng Liu and J. Xu Li, “Drug addiction: a curable mental disorder?” *Acta Pharmacologica Sinica*, vol. 39, no. 12. Nature Publishing Group, pp. 1823–1829, Dec. 01, 2018. doi: 10.1038/s41401-018-0180-x.

[2] F. Zhao *et al.*, “Computational Approaches to Detect Illicit Drug Ads and Find Vendor Communities Within Social Media Platforms,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 19, no. 1, pp. 180–191, 2022, doi: 10.1109/TCBB.2020.2978476.

[3] A. Islam, M. Sadekur Rahman, M. Tarek Habib, M. Ariful Islam Arif, S. Islam Sany, and F. Sharmin, “Prediction of Addiction to Drugs and Alcohol Using Machine Learning: A Case Study on Bangladeshi Population E+ Youth (Galvanizing Energy with Experience towards Youth Empowerment) View project machine learning View project Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population,” *Article in International Journal of Electrical and Computer Engineering*, vol. 11, no. 5, pp. 4471–4480, 2021, doi: 10.11591/ijece.v11i5.

[4] G. Shobana and S. N. Bushra, “Drug administration route classification using machine learning models,” in *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 654–659. doi: 10.1109/ICISS49785.2020.9315975.

[5] F. P. Brishty and S. Akhter, “Detection of drug-induced QT Syndrome from ECG using machine learning techniques,” Dec. 2018.

[6] G. Shobana and N. Priya, “Cancer drug classification using artificial neural network with feature selection,” in *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 1250–1255. doi: 10.1109/ICICV50876.2021.9388542.

[7] L. Mandal and N. D. Jana, “A Comparative Study of Naive Bayes and k-NN Algorithm for Multi-class Drug Molecule Classification,” *A Comparative Study of Naive Bayes and k-NN Algorithm for Multi-class Drug Molecule Classification*, 2019.

[8] V. Da Poian *et al.*, “Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry,” *Frontiers in Astronomy and Space Sciences*, vol. 10, 2023, doi: 10.3389/fspas.2023.1134141.

[9] T. Milo and A. Somech, “Automating Exploratory Data Analysis via Machine Learning: An Overview,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[10] R. Indrakumari, T. Poongodi, and S. R. Jena, “Heart Disease Prediction using Exploratory Data Analysis,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 130–139. doi: 10.1016/j.procs.2020.06.017.

DOI: https://doi.org/10.29207/resti.v7i6.5401

[11] A. Shah *et al.*, “A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN),” *Clinical eHealth*, vol. 6, pp. 76–84, Dec. 2023, doi: 10.1016/j.ceh.2023.08.002.

[12] B. P. Adedeji, “Electric vehicles survey and a multifunctional artificial neural network for predicting energy consumption in all-electric vehicles,” *Results in Engineering*, vol. 19, Sep. 2023, doi: 10.1016/j.rineng.2023.101283.

[13] Q. Ren, H. Zhang, D. Zhang, X. Zhao, L. Yan, and J. Rui, “A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree,” *J Pet Sci Eng*, vol. 208, Jan. 2022, doi: 10.1016/j.petrol.2021.109681.

[14] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[15] R. Prabaswara, J. Lemantara, and J. Jusak, “Classification of Secondary School Destination for Inclusive Students using Decision Tree Algorithm,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, Aug. 2023, doi: 10.29207/resti.v7i5.5081.

[16] Z. Xu, J. Cao, G. Zhang, X. Chen, and Y. Wu, “Active learning accelerated Monte-Carlo simulation based on the modified K-nearest neighbors algorithm and its application to reliability estimations,” *Defence Technology*, 2022, doi 10.1016/j.dt.2022.09.012.

[17] D. Cheng, Y. Shi, T. Lin, B. H. Gwee, and K. A. Toh, “Hybrid K-means clustering and support vector machine method for via and metal line detections in delayered IC images,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 12, pp. 1849–1853, Dec. 2018, doi: 10.1109/TCSII.2018.2827044.

[18] R. Yao, J. Li, M. Hui, L. Bai, and Q. Wu, “Feature Selection Based on Random Forest for Partial Discharges Characteristic Set,” *IEEE Access*, vol. 8, pp. 159151–159161, 2020, doi: 10.1109/ACCESS.2020.3019377.

[19] H. Yun, “Prediction model of algal blooms using logistic regression and confusion matrix,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2407–2413, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.

[20] D. Valero-Carreras, J. Alcaraz, and M. Landete, “Comparing two SVM models through different metrics based on the confusion matrix,” *Comput Oper Res*, vol. 152, Apr. 2023, doi: 10.1016/j.cor.2022.106131.

[21] A. Nugroho, A. Bimo Gumelar, A. G. Sooai, D. Sarvasti, and P. L. Tahalele, “Perbandingan Performansi Algoritma Pengklasifikasian Terpandu Untuk Kasus Penyakit Kardiovaskular,” *masa berlaku mulai*, vol. 1, no. 3, pp. 998–1006, 2017.

[22] M. H. Z. Al Faroby, M. I. Irawan, and N. N. T. Puspaningsih, “XGBoost and Network Analysis for Prediction of Proteins Affecting Insulin based on Protein-Protein Interactions,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 253–262, Nov. 2020, doi: 10.22219/kinetik.v5i4.1076.

[23] S. Narkhede, “Understanding AUC - ROC Curve,” *Understanding AUC - ROC Curve*, Jun. 2018.

[24] A. J. Bowers and X. Zhou, “Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes,” *J Educ Stud Placed Risk*, vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

DOI: https://doi.org/10.29207/resti.v7i6.5401