Evaluation of Classification Algorithms in Tracing Malicious Telephone Numbers

Van Vuong Ngo*

Hanoi, Vietnam

*Corresponding author: vanvuong.ngo.vt@gmail.com

(Received:7 August 2024; Accepted: 27 August 2024)

Abstract— Mobile phones and telecommunications networks have recently played an important role in modern society. They are dispensable parts of our lives as they facilitate the way we communicate. However, apart from their benefit, their proliferation has some drawbacks as telephone networks can be exploited. For example, commercial calls can be made repeatedly to advertise companies' products. These calls annoy customers because they promote products without considering customers' interests. These unexpected calls not only cause a negative impact on the networks but also disturb mobile phone users. To confront this problem, the network administrators need some methods to detect the phone numbers that are used to make the harassment. Therefore, we proposed a solution based on machine learning classification models. Then the performance of some models, namely K-Nearest Neighbors, Decision Tree, and Logistic Regression, is compared. By applying the machine learning models, network administrators can identify and restrict malicious telephone numbers.

Keywords: *Telecommunications, Classification, Machine learning, Network.*

1. INTRODUCTION

This section introduces telecommunications networks and the issue of telephone harassment. It also shows some information about machine learning and its algorithms.

1.1 Telephone Harassment

In recent years, mobile phones have become ubiquitous. With the advancement of technology, mobile phones help us in many aspects of our lives. For example, they enhance the ability to communicate frequently. Besides, telecommunications systems are expanding rapidly. Apart from 4G networks [1], the telecommunications vendors are also researching on 5G or 6G networks [2][3]. Fig. 1 depicts a simple telecommunications network with many subsystems such as Evolved Packet Core (EPC) [4], Public Switching Telephony Network (PSTN) [5], or 5G New Radio (5G NR) [6].

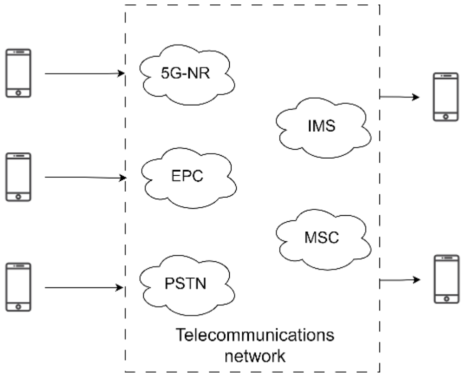


Fig. 1. A telecommunications system.

DOI: https://doi.org/10.69955/ajoeee.2024.v4i2.64

Nevertheless, the expansion of telecommunications systems and personal devices brings some things that could be improved. One of the issues is telephone harassment. Telephone harassment can come in many forms. They can be advertising calls to introduce some products which the customers do not care about. Another form is disruptive calls that clog the hotlines of companies. This type of attack is often performed automatically by pre-programmed software. These malicious calls affect telecommunications networks and other mobile phone users.

1.2 Machine learning algorithms

There are some categories of machine learning algorithms: supervised algorithms, unsupervised algorithms... [7]. Supervised algorithms are divided into two types: classification and regression. While regression predicts the output values based on the input data, classification categorizes outputs into predefined values or classes. For example, predicting the price of a house is an example of regression, while determining whether an animal in a photo is a cat or a dog is an example of classification. Some classification algorithms are shown in Fig. 2.

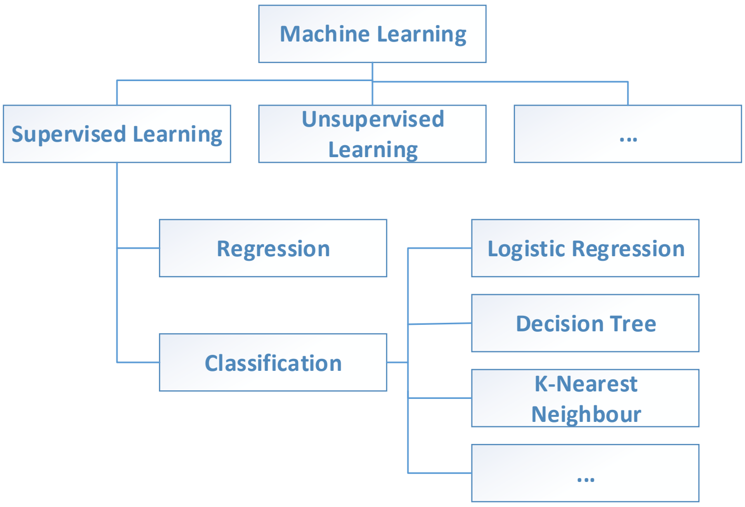


Fig. 2. Machine learning categories.

2. MOTIVATION AND RELATED WORKS

The advent of machine learning (ML) applications in telecommunications has recently attracted researchers' interest. In [8], the authors proposed a multi-layer model to determine the reasons for call failure. By analyzing the Call Detail Records (CDRs) [9], the model can classify problems into different failure categories, such as Charging Failure or Media Failure.

Another novel topic is Security using Machine Learning. An intrusion detection method has been developed for multimedia platforms [10]. This intrusion detection method protects against flooding attacks, which can cause network congestion. Sammer [11] also proposed an ML-based approach for intrusion detection in Mobile Ad hoc Networks (MANETs).

Regarding other topics, a manuscript [12] discusses the application of ML in analyzing customer behavior based on features such as age, gender, or annual income. Other ML applications were introduced in [13][14], which are about detecting fake data or predicting cryptocurrency prices.

Therefore, we came up with the idea that Machine Learning models can mitigate the telephone harassment problem. By Applying machine learning models, network administrators can identify the malicious telephone numbers that cause problems for networks and users.

3. CLASSIFICATION MODELS FOR TRACING MALICIOUS TELEPHONE NUMBERS

3.1 Logistic Regression

This classification model is based on the neural network concept [15]. The neural network consists of three layers: the input layer, the hidden layer, and the output layer. Each layer has some nodes which have their own weights. When the inputs are forwarded to a layer, the inputs are multiplied by adjusted weights to produce the outputs of that layer. Finally, the output layer applies a binary activation function that can predict whether the result is 0 or 1. In the case of our study, the activation function suggests whether a mobile phone number makes telephone harassment or not. The Sigmoid function [16] is a suitable choice for the activation function because it is monotonic and its values range between 0 and 1. The Sigmoid function is defined as formula (1) and depicted in Fig. 3.

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

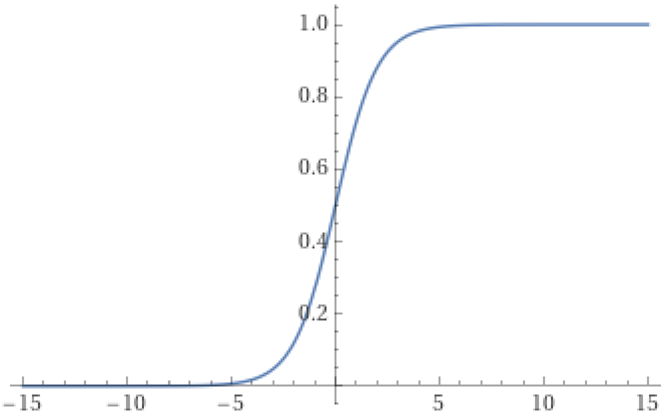


Fig. 3. The Sigmoid function

Fig. 4 depicts a neural network for the classification model. The input of the neural network is the matrix [N*M]. Each row of the input matrix represents a phone number and its attributes. While N is the number of mobile phone numbers that need to be examined, M is the number of attributes. For instance, if the number of calls per day and the average call duration are chosen as attributes, then M equals 2 in this case. It can be predicted that mobile phone numbers that make a significant number of calls with short call duration are the cause of telephone harassment. At the final layer, the output is the matrix [N*1]. Each row of the output matrix takes the binary value 0 or 1, which indicates whether the mobile phone number of this row makes telephone harassment or not. Thanks to this result, network administrators can impose restrictions on phone numbers that cause harassment. Fig. 5 shows an example of the evaluation with the neural network.

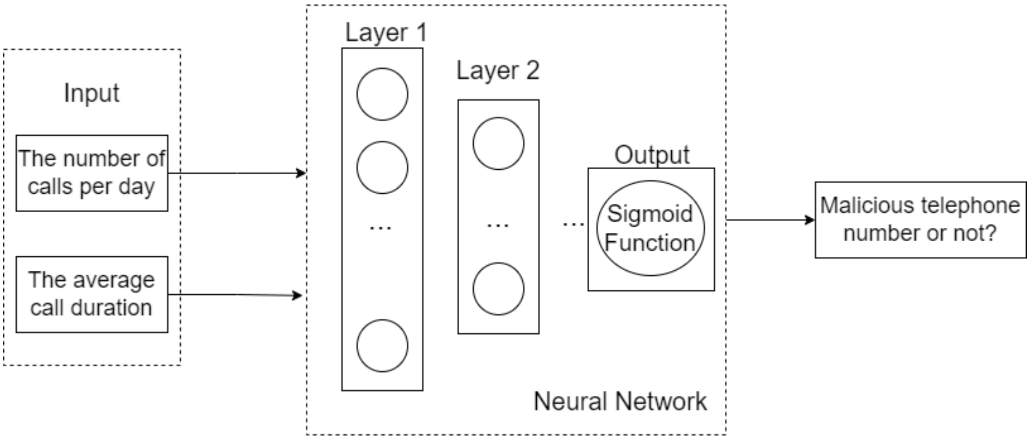
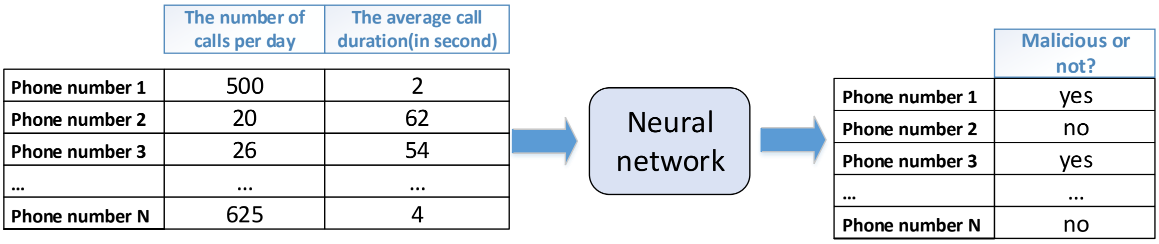

Input
The number of calls per day
The average call duration
Layer 1
Layer 2
Output
Sigmoid Function
Neural Network
Malicious telephone number or not?

Fig. 4. The neural network model



	The number of calls per day	The average call duration(in second)
Phone number 1	500	2
Phone number 2	20	62
Phone number 3	26	54
…	…	…
Phone number N	625	4

	Malicious or not?
Phone number 1	yes
Phone number 2	no
Phone number 3	yes
…	…
Phone number N	no

Fig. 5. An example of the evaluation with the neural network

3.2 *K-nearest Neighbors*

K-nearest neighbors (KNN) is a popular supervised machine learning algorithm [17]. The idea of KNN is that the characteristic of a data point is similar to its closest neighboring points. K is the number of nearest neighbors to use. The distance between the given point and other points is calculated to determine which points are closest to a given data point. The distance between these points is calculated using Euclidean distance formula as follows:

$$d(x,y) = \sqrt{(y-x)^2} \qquad (2)$$

Fig. 6 illustrates the KNN algorithm. Fig. 6a shows the case where K = 1, and the data point is predicted as class 1 because its closest point belongs to class 1. In Fig. 6b, for K =3, among the three closest points of the given point, there are two class 2 points and one class 1 point. Therefore, the given point is predicted to be class 2. Applying the KNN algorithm to the problem of malicious telephone numbers, the vertical axis can represent the number of calls per day, whereas the horizontal axis can represent the average call duration, as can be seen in Fig. 7.

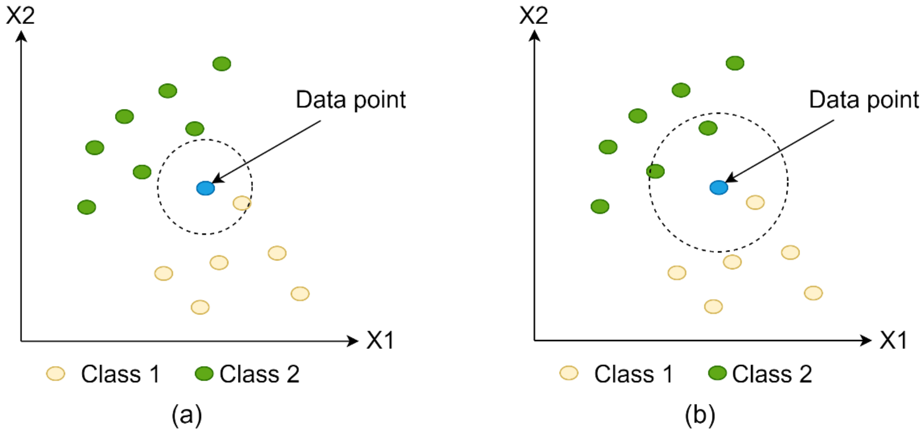


Fig. 6. The KNN algorithm

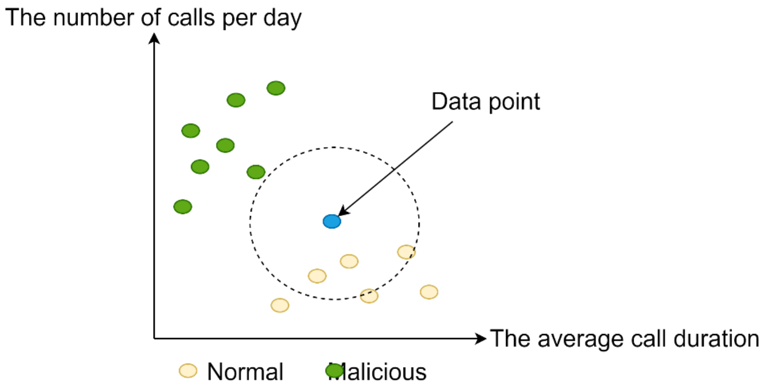


Fig. 7. The KNN diagram

3.3 Decision Tree

Decision Tree is a tree-structured classifier that is preferred for solving classification problems [18]. The initial node is the root of the tree, branches represent the decision rules, and leaves are the outcomes. Fig. 8 shows a model of a decision tree.

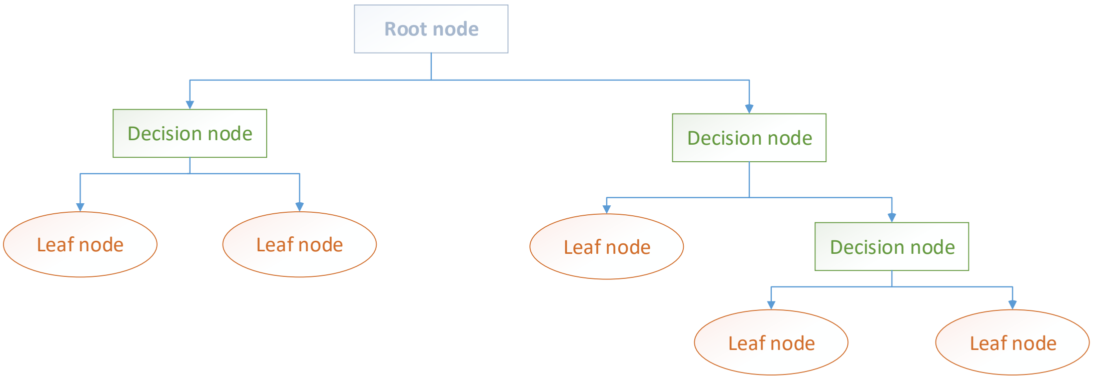


Fig. 8. The decision tree model

In a decision tree, the algorithm commences from the root of the tree. It evaluates the value of the data point using the condition in the root node. Based on this evaluation, the data point will follow a specific branch to the next decision node. In this decision node, the algorithm compares the data again and moves the point further. This process continues until the point reaches a leaf node of the tree.

An example of the decision tree for harassment problems can be shown in Fig. 9. By analyzing the number of calls and the average call duration that a phone number makes per day; network administrators can predict whether this phone number is malicious or not.

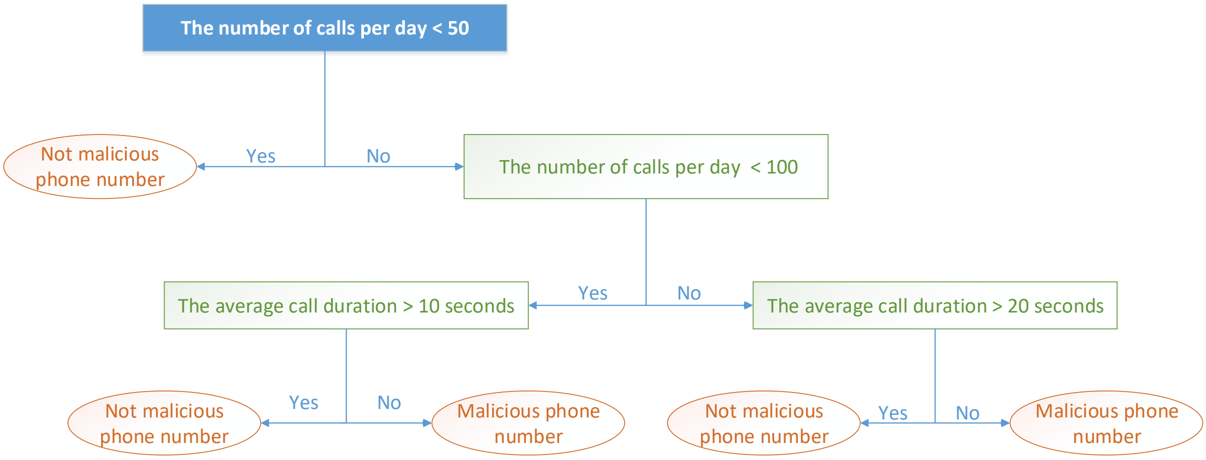

The number of calls per day < 50
Not malicious phone number
Yes
No
The number of calls per day < 100
The average call duration > 10 seconds
Yes
No
The average call duration > 20 seconds
Not malicious phone number
Yes
No
Malicious phone number
Not malicious phone number
Yes
No
Malicious phone number

Fig. 9. An example of the decision tree

4. EVALUATION OF CLASSIFICATION MODELS

In this section, the KNN algorithm, decision tree algorithm, and the neural network with logistic regression are evaluated with the task of predicting malicious telephone numbers. A simulated dataset is prepared for this evaluation. The dataset is just some examples to compare and evaluate the performance of the classification models. The 10-fold cross validation is also applied to utilize the datasets for better results [19]. The 10-fold cross validation divides the initial dataset into 10 folds, then it uses 9 folds to train and the last fold is used for validation. This training procedure repeats 10 times, so each fold becomes a test fold once.

It can be seen that the KNN algorithm has the best accuracy, while the decision tree algorithm's accuracy is similar to that of logistic regression. Fig. 10 shows a decision tree for the dataset with 60 instances, while Fig. 11 shows that of the dataset with 45 instances. In Fig. 10, the root node first checks whether the average call duration exceeds 15 seconds. If the average call duration of a phone number exceeds 15 seconds, it can be inferred that the user makes normal conversation calls (indicated by “no”). On the other hand, if the average call duration is below 15 seconds, the tree checks whether the number of calls per day exceeds 15. If the number of calls exceeds 15 calls, it can be predicted that the phone number is malicious (indicated by “yes”). If the number of calls is 15 or fewer, the phone number is predicted to be expected. In Fig. 11, the tree’s decision progress is similar.

Table 1: Comparison with the dataset of 60 instances

Number of instances in the dataset	Models	Accuracy
60 instances	Logistic regression	95%
	KNN (with K=3)	98.33%
	Decision tree	95%

Table 2: Comparison with the dataset of 45 instances

Number of instances in the dataset	Models	Accuracy
45 instances	Logistic regression	95.566%
	KNN (with K=3)	97.778%
	Decision tree	95.566%

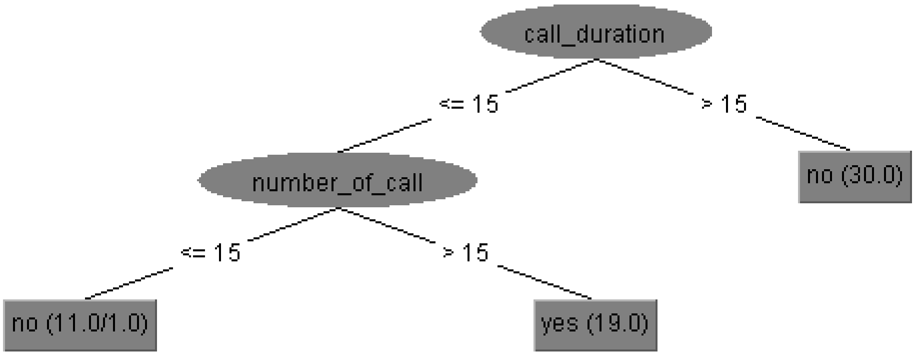


Fig. 10. The decision tree of the dataset with 60 instances

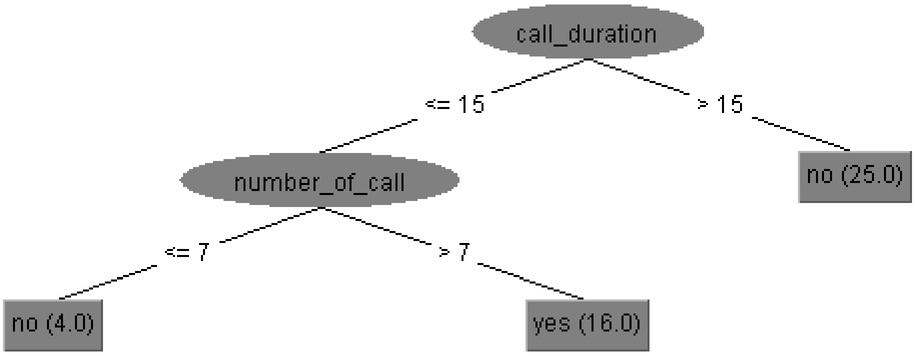


Fig. 11. The decision tree of the dataset with 45 instances

5. CONCLUSION

The malicious telephone numbers can harm the telecommunications network and disrupt the experience of mobile phone users. Machine learning classification algorithms can facilitate tracing these malicious telephone numbers. Network administrators can detect suspicious telephone numbers and impose restrictions on them thanks to these algorithms. In this manuscript, a simple comparison is conducted with the KNN algorithm, the Decision Tree algorithm, and the neural network with Logistic Regression.

REFERENCES

[1] Hicham, Magri & Abghour, Noreddine & Ouzzif, Mohammed. (2015). 4G System: Network Architecture and Performance.

[2] Zaame, I. & Mazri, Tomader & Elrhayour, A.. (2020). 5G: Architecture Overview And Deployments Scenarios. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLIV-4/W3-2020. 435-440. 10.5194/isprs-archives-XLIV-4-W3-2020-435-2020. https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-435-2020

[3] Tataria, Harsh & Shafi, Mansoor & Molisch, Andreas & Dohler, Mischa & Sjoland, Henrik & Tufvesson, Fredrik. (2021). 6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities. Proceedings of the IEEE. PP. 1-34. 10.1109/JPROC.2021.3061701. https://doi.org/10.1109/JPROC.2021.3061701

[4] Hayashi, Toshiki. (2012). Evolved Packet Core (EPC) network equipment for Long Term Evolution (LTE). Fujitsu scientific & technical journal. 48.

[5] ETSI TR 101 292, Public Switched Telephone Network (PSTN), 1999-09.

[6] Sauter, Martin. (2021). 5G New Radio (NR) and the 5G Core. https://doi.org/10.1002/9781119714712.ch6

[7] Sah, S. Machine Learning: A Review of Learning Types. Preprints 2020, 2020070230. https://doi.org/10.20944/preprints202007.0230.v1

[8] A. Bahaa, M. Shehata, S. M. Gasser, S. El-Mahallawy, "Call Failure Prediction in IP Multimedia Subsystem (IMS) Networks," in Applied Science Journal, 2022,12,8378. https://doi.org/10.3390/app12168378

[9] ETSI TR 122 115, Charging and Billing, 2000-01.

[10] C. Hsu, S. Wang, Y. Qiao, "Intrusion detection by machine learning for multimedia platform," in Multimedia Tools and Applications, 2021, pp. 29643-29656, https://doi.org/10.1007/s11042-021-11100-x

[11] A. R. Sammer, "A deep and machine learning comparative approach for networks intrusion detection", Asian Journal of Convergence in Technology, Vol 10 No.1 (2024), pp.98-103.

[12] R. P. Sinkar, "Use of Machine Learning Application for Business Perspective", Asian Journal of Convergence in Technology, Vol 10 No.1 (2024), pp.74-79.

[13] Dharmireddy, A., & Gottipalli, M. D. (2023). Social Networking Sites Fake Profiles Detection Using Machine Learning Techniques. Asian Journal For Convergence In Technology (AJCT) ISSN -2350-1146, 9(3), 09-15. https://doi.org/10.33130/AJCT.2023v09i03.002

[14] Kawli, D. P., Chaudhari, A. S., Ingale, P. D., Telange, G. A., & Banik, A. (2024). "Cryptocurrency Price Prediction Using Machine Learning", Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146, 10(1), 19-23. https://doi.org/10.33130/AJCT.2024v10i01.004

[15] Wegner, Sven. (2024). Neural Networks. 10.1007/978-3-662-69426-8_16. https://doi.org/10.1007/978-3-662-69426-8_16

[16] Geng, Yu & Li, Qin & Yang, Geng & Qiu, Wan. (2024). Logistic Regression. 10.1007/978-981-97-3954-7_4. https://doi.org/10.1007/978-981-97-3954-7_4

[17] Cunningham, Padraig & Delany, Sarah. (2007). k-Nearest neighbour classifiers. Mult Classif Syst. 54. 10.1145/3459665.

[18] Wang, Zijun & Gai, Keke. (2024). Decision Tree-Based Federated Learning: A Survey. Blockchains. 2. 40-60. 10.3390/blockchains2010003. https://doi.org/10.3390/blockchains2010003

[19] D. Anguita, Ghio A., S. Ridella, and D. Sterpi. K-fold cross validation for error rate estimate in support vector machines. In Proc. of the Int. Conf. on Data Mining, 2009.

[20] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.