Utilizing the Genetic Algorithm to Pruning the C4.5 Decision Tree Algorithm

Maad M. Mijwil[1] and Rana A. Abttan[2]

[1] Computer Techniques Engineering Department, Baghdad College of Economic Sciences University
Baghdad, Iraq
Email: mr.maad.alnaimiy [AT] baghdadcollege.edu.iq

[2] Computer Techniques Engineering Department, Baghdad College of Economic Sciences University
Baghdad, Iraq
Email: rana.ali.abttan [AT] baghdadcollege.edu.iq

ABSTRACT— *A decision tree (DTs) is one of the most popular machine learning algorithms that divide data repeatedly to form groups or classes. It is a supervised learning algorithm that can be used on discrete or continuous data for classification or regression. The most traditional classifier in this algorithm is the C4.5 decision tree, which is the point of this research. This classifier has the advantage of building a vast data set and does not stop until it reaches the desired goal. The problem with this classifier is that there are unnecessary nodes and branches leading to overfitting. This overfitting can negatively affect the classification process. In this context, the authors suggest utilizing a genetic algorithm to prune the effect of overfitting. This dataset study consists of four datasets: IRIS, Car Evaluation, GLASS, and WINE collected from UC Irvine (UCI) machine learning repository. The experimental results have confirmed the effectiveness of the genetic algorithm in pruning the effect of overfitting on the four datasets and optimizing confidence factor (CF) of the C4.5 decision tree. The proposed method has reached about 92% accuracy in this work.*

Keywords— Genetic algorithm, C4.5 Decision tree, Optimizing, Pruning, Machine learning.

1. INTRODUCTION

In the past few years, machine learning has changed the world around us by creating miraculous tasks that have helped implement many applications and systems that serve humanity [1]. Machine learning helps to analyse data excellently and provides high accuracy to achieve the expected goal of data analysis [2][3]. In addition, deep learning takes machine learning to a whole new level by distinguishing tasks, and its ideas are inspired by human brain neural networks [4]. For example, a telephone is equipped with a voice control system and remote control devices for the TV. In short, machine learning is a data modelling technique. The most famous machine learning techniques are Linear regression, Logistical Regression, Random Forest, Support Vector Machines, Decision Trees, and Naive Bayes [5]. These techniques can help the system learn from experience [6]. It has the ability to enable the system to acquire knowledge and integrate it through extensive observations, and to improve itself by learning new knowledge instead of programming using that knowledge [7].

Machine learning techniques can well organize existing knowledge and gain entirely new knowledge through smart logging and logical thinking about data [8]. Machine learning systems have achieved various intelligent results ranging from memorizing system changes to creating whole new scientific theories. Moreover, it has the capacity for continuous self-improvement to enable its systems to become more intelligent and useful [9]. Figure 1 explains how machine learning techniques work. This study focuses on the decision tree technique using a genetic algorithm to update it. Decision trees are one of the tree-based techniques used in classification and regression problems. It can be used in complex datasets (see section 3).

The contribution of this paper lies in the application of a genetic algorithm to reduce the overfitting that occurs as a result of the large number of unnecessary branches that arise in the C4.5 decision tree, pruning these branches, improving the confidence factor, and reaching a satisfactory result. As well as, the making of the decision tree is straightforward in the process of tracking it. The genetic algorithm is applied to four methods; each method contains a set of different data and includes the number of attributes and the number of instances.

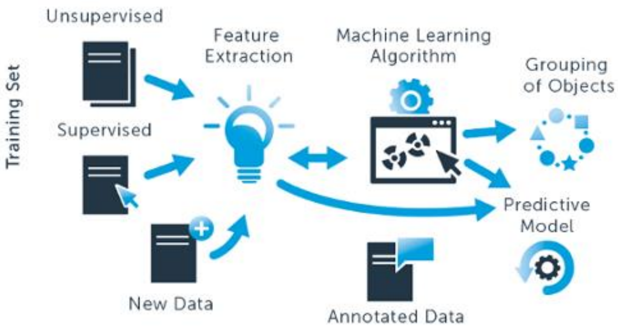


Figure 1: Machine learning techniques work [7].

The rest of the paper is organized as follows: The second section has a literature review of different state-of-the-art techniques. The third section has a description of all about datasets. The fourth is all about the algorithms (C4.5 decision trees and genetic algorithm). The fifth section has a result and evaluation of the proposed method, and at last, the sixth section has the conclusion and future work.

2. LITERATURE REVIEW

This section covers some previous studies and their findings using the genetic algorithm or other techniques in the C4.5 decision tree pruning process. In fact, there are not many studies on the internet that look like the current study. A number of studies have been investigated in terms of their closeness and achievement with regards to the work of the current paper.

The first is a study conducted by Fu et al. [11], they suggest to build a C 4.5 decision tree for a set of realistic marketing data using a genetic algorithm. The data for this study is divided into a training, recording and testing groups. In this study, the researchers find that their proposed method has proven effective in creating high-quality decision trees. In a study by Chen et al. [12], they employ a genetic algorithm to reduce the overfitting that occurs in C4.5 decision trees, delete useless branches, and make the C4.5 decision tree easy to track. The data for this study includes four datasets from the UCI website. The researchers achieve better results in the decision tree pruning process with the genetic algorithm. In another study, Jankowski and Jackowski in 2016 [13], apply evolutionary algorithms to decision tree induction. The purpose of this research is to reduce tree size and misclassification rate. Seven data sets are used from the UCI website: abalone, ecoli, winequality-red, page-blocks, breast tissue, seeds, winequality-white. It has achieved remarkable results in reducing the size of the decision tree. Khanbabaei and Alborzi [14], discuss the application of the C4.5 decision tree to bank data to provide credit facilities for each category of customers. A genetic algorithm is used to track the decision tree better and easier, remove unnecessary branches in the C4.5 decision tree, and reduce this tree's size. This study achieves good results in the classification of data banks with 90% accuracy. A study conducted by Muslim et al. from Nigeria [15] uses data mining techniques to extract data from C4.5 decision trees. This research includes a chronic kidney disease dataset in which pessimistic pruning is applied to identify and delete unnecessary branches from this tree. The accuracy of this study in diagnosing patients with chronic kidney disease reaches more than 96%.

3. DATASETS DESCRIPTION

This section presents the datasets that are obtained from the UCI website. This website includes more than 590 datasets to assist researchers in the machine learning community. This website is created in 2007 by A. Asuncion and D. Newman. The datasets in this study are IRIS [13], WINE [14], Car Evaluation [15], and GLASS [16].

3.1 IRIS Dataset

IRIS is the most prominent and famous classification dataset type for C4.5 decision tree. It contains three classes (Setosa, Versicolor, and Virginica), each class has 50 cases, each class symbolizes an iris plant. First class is linearly separable, while the other two classes are not linearly separable. Each row of the table represents an iris flower, including its type and the size (in centimetres) of its plant parts (sepals and petals). The rows are the samples, and the columns are the sepal length, sepal width, petal length and petal width. Figure 2 shows three classes of IRIS dataset. The purpose of this dataset is to train a C4.5 decision tree classifier to classify any type of data based on given attributes that are the sepal and petal size.

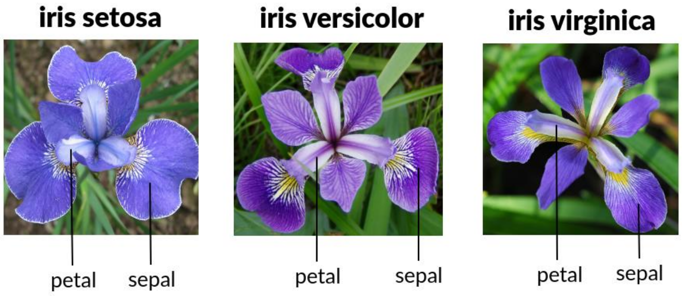


Figure 2: Three classes of IRIS dataset for classification [17]

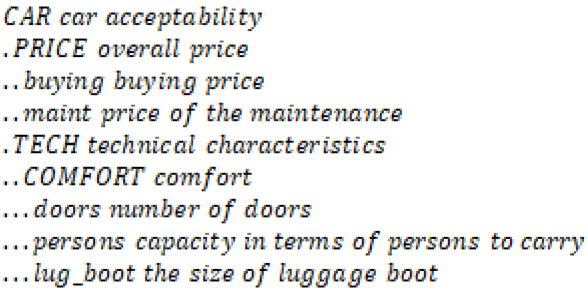
3.2 WINE Dataset

Italy is one of the oldest wine-producing countries in the world. This country produces about 20% of the wine in the world. Nowadays, there are more than 20 different wine regions in Italy that have a wide variety of wines and is also home to many vineyards. Figure 3 shows the most famous wines in Italy. The idea of this dataset is the chemical analysis of the wines produced in a region in Italy, there are three different types of wine. The analysis consists of 13 elements found in each of the three types. These elements are Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Colour intensity, Hue, OD280/OD315 of diluted wines, and Proline.

Figure 3: The most famous wines in Italy.

3.3 Car Evaluation Dataset

This database is based on a simple hierarchical decision model, as this model helps to evaluate cars according to a specific structure, which is as follows:



CAR car acceptability
.PRICE overall price
..buying buying price
..maint price of the maintenance
.TECH technical characteristics
..COMFORT comfort
...doors number of doors
...persons capacity in terms of persons to carry
...lug_boot the size of luggage boot
. . safety estimated safety of the car

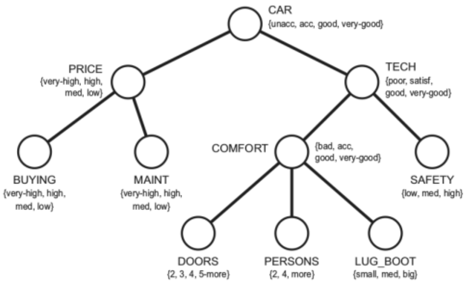

CAR
{unacc, acc, good, very-good}
PRICE
{very-high, high, med, low}
TECH
{poor, satisf, good, very-good}
BUYING
{very-high, high, med, low}
MAINT
{very-high, high, med, low}
COMFORT
{bad, acc, good, very-good}
SAFETY
{low, med, high}
DOORS
{2, 3, 4, 5-more}
PERSONS
{2, 4, more}
LUG_BOOT
{small, med, big}

Figure 4: Case of the car evaluation tree [18]

This model relies massively on three concepts: PRICE, COMFORT, TECH. Every concept is in the original model linked to its lower-level descendants. Figure 4 shows a simple case of a car evaluation decision tree.

3.4 GLASS Dataset

This dataset includes testing glass products by Vina company, using three algorithms, namely BEAGLE, the nearest-neighbour algorithm, and discriminant analysis. This test aims to challenge whether the type of glass is float or not. Figure 5 shows Vine products.

Figure 5: Vina Products (download from Google).

4. THE ALGORITHMS

4.1 C4.5 Decision Trees

C4.5 algorithm [19][20] is developed by American computer scientist J.R. Quinlan. It is one of the most popular classification algorithms in the community of researchers and practitioners in the area of machine learning. This algorithm is widely used for comparing the performance of classifiers on unbalanced datasets. Also, decision trees are common to use because they are easy to understand and interpret. A decision tree structure consists of root, node, branch and leaf. The lowest part in the tree structure is called the leaf and the uppermost part is called the root. Each attribute in the data set represents the nodes. The part that provides the connection between nodes is named a branch (see Figure 6). The disadvantages of the C4.5 algorithm are that it builds null branch with zero values as well overfitting occurs when the algorithm model collects data with unusual characteristics, especially when the data is unnecessary. That is why, the defects that are mentioned by using a genetic algorithm to obtain a decision tree free of defects are eliminated in this paper. This work is accomplished in two steps: the first is the process of creating a decision tree by four datasets, and the other step is the implementation of the pruning algorithm on these datasets. Deciding according to which attribute value the branching will take place can be considered as the essential process step in creating decision trees. In addition, there are three criteria that are widely applied as decision criteria, namely Gini index, Information gain, and Towing rule. In this paper, Information gain is applied as a decision-making criterion. Moreover, the effect of an entropy-based value on the outcome is calculated for each feature in this algorithm.

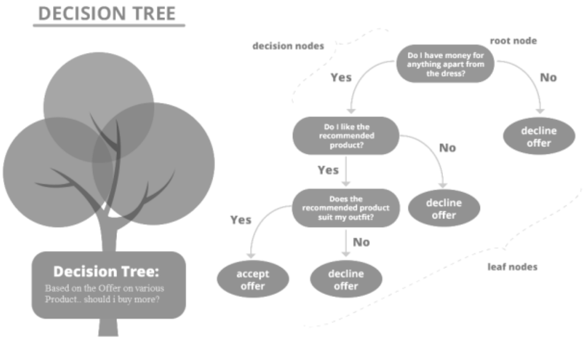

DECISION TREE
decision nodes
root node
Do I have money for anything apart from the dress?
Yes
No
Do I like the recommended product?
decline offer
No
Yes
Does the recommended product suit my outfit?
decline offer
Yes
No
accept offer
decline offer
leaf nodes
Decision Tree:
Based on the Offer on various Product.. should i buy more?

Figure 6: Decision tree structure with simple example [21]

Mathematical calculations are started at this level. If it is assumed that there are M classes and these class values are repeated by T, the probability value of a class is calculated by equation 1. Where C_i represents the number of class values belonging to a class. The entropy value of these classes can be determined by equation 2. Considering that the T class values are divided into subsets as $T_1, T_2, T_3, \dots\dots, T_M$ in the dataset according to the Y attribute values, the information gain to be obtained by dividing the T class values by using the Y attribute values is $IG\ (Y, T)$. The equation 3 is the result of $IG\ (Y, T)$. While equation 4 indicates departure information in determining the value of its attribute (Y) for the dataset.

$$P_i = \frac{c_i}{|T|} \qquad (1)$$

$$H\ (T) = -\sum_{i=1}^{M} P_i\,log2P_i\,(2) \qquad (2)$$

$$IG(Y,T) = H(T) - \sum_{i=1}^{M} \frac{|T_i|}{|T|} H(T_i) \qquad (3)$$

$$DI(Y) = -\sum_{i=1}^{M} \frac{|T_i|}{|T|} Log_2\left(\frac{|T_i|}{T}\right) \qquad (4)$$

Information gain is provided by separating related attributes. Through this mechanism, the gain information for each parameter is calculated and the tree structure is separated according to quality with the highest gain information. After the decision tree generation is complete, the pruning process is required to be made. This process is done in two methods. The first method is pre-pruning, which is stopping the division so that the tree does not grow more, it means stopping the tree at a specific growth rate. The second method, called final pruning, is by removing the split points that are created after the tree is fully formed. In other words, the genetic algorithm does these methods.

4.2 Genetic Algorithm

Genetic algorithm [22-24] is an algorithm explained by American psychologist and computer scientist John Holland to solve optimization problems. The algorithm is based on the evolution of nature as a model. It is widely used as a tool for exploration, optimization and machine learning. The genetic principle uses an evolutionary principle, in which priority is given to individuals who obtain better and adaptable genetic information from the genes grown by their parents and are highly flexible to the environment for several generations.

Genetic algorithms represent possible solutions to problems in specific data structures, and then transform them to find better solutions. The possible solution to the problem to be solved is called a single organism or entity, and their collection is called a population. A person is usually composed of one or several chromosomes, and the operons that modify the chromosomes are called genetic operons. Basically, there are three operators: selection, crossover and mutation. Figure 7 shows the structure of genetic algorithm. This figure is applied to achieve adequate or acceptable results in the decision-tree pruning process and to optimize the confidence factor.

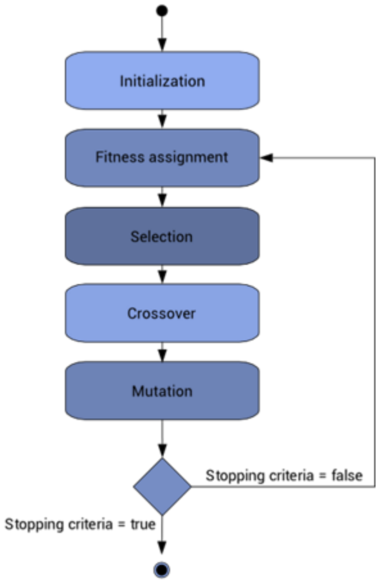


Figure 7: Genetic algorithm structure [25]

5. THE EXPERIMENTAL RESULTS

This section discusses the steps for improving the proposed method to reduce overfitting by using the genetic algorithm. Table 1 lists all the datasets used in this paper, with the number of attributes (NA), the number of instances (NI) of each dataset, and the coefficient factor. A coefficient factor (CF) is calculated using equation 5 in Weka software.

$$CF = NI \div NA \quad (5)$$

In the event that all the attributes are included in the tree structure, false results will occur for some samples in the testing phase and this condition is called overfitting. In this study, the coefficient factor (CF) is optimized in order to prevent overfitting more effectively by pruning process. In addition, the value of pruning of the tree is determined after the tree has been developed, and its value is between 0 and 1. After pruning process, the accuracy rates of the classification process are calculated with a 10-fold cross-validation process. Table 2 presents the effects of the decision tree classification success rate with the proposed method that includes CF and pruning decision without the genetic algorithm. Table 3 exhibits a comparison between the results obtained from the proposed method without using the genetic algorithm and with using the genetic algorithm.

Table 1: Properties of the datasets used.

Dataset	Number of Instances	Number of Attributes	Coefficient factor
IRIS	150	4	0.026
Car Evaluation	1728	6	0.003
GLASS	214	10	0.046
WINE	178	13	0.073

Table 2: The results of decision tree classification with the proposed method without genetic algorithm.

Dataset	No Pruning-Increased overfitting (Normal Tree)	Proposed Method without genetic algorithm (Increasing)
IRIS	94.20%	94.20%
Car Evaluation	69.10%	77.18%
GLASS	63.40%	70.70%
WINE	91.30%	93.25%

Table 3: The results of decision tree classification with the proposed method using genetic algorithm.

Dataset	Proposed Method without genetic algorithm	With genetic algorithm (Reducing)
IRIS	94.20%	92.17%
Car Evaluation	77.18%	93.11%
GLASS	70.70%	91.42%
WINE	93.25%	92.31%

Through Table 2, it is noticed that the proposed method without using the genetic algorithm has increased the percentage of overfitting of the Car Evaluation, GLASS and WINE datasets, except for IRIS. This method is incorrect because the overfitting of these trees will become very complicated. As can be seen in Table 3, the proposed method with genetic algorithm provides a better success rate in all dataset. This method has succeeded in reducing the rate of overprocessing significantly and positively by improving the coefficient factor (CF). This indicates that the genetic algorithm is working correctly in enhancing the decision tree for any datasets.

6. CONCLUSIONS AND FUTURE WORK

The purpose of the current paper is to achieve the final pruning and to have a stable and easy-to-track decision tree by executing a genetic algorithm on it. This algorithm assists in achieving excellent results in the pruning process and reducing unnecessary branches as well as optimizing confidence factor (CF) of the C4.5 decision tree. The results achieved are effective by about 92% in reducing overfitting in Weka software with PC Laptop: RAM:8GB, Intel® Core™ i5-1130G7 Processor, Hard disk:256GB SSD, and running on Ubuntu 18.04.4 LTS 64-bit. In the future, improvements can be made for datasets for which the number of features related to this study is limited. In addition, the proposed method can be applied to different data sets and the classification success rates can be compared.

7. REFERENCES

[1] Brohi S. N., Pillai T. R., Kaur S., Kaur H., Sukumaran S., and Asirvatham D., "Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education," *In Proceedings of International Conference on Emerging Technologies in Computing (iCETiC 2019)- Springer*, pp: 254-261, London, United Kingdom, 19-20 August 2019. https://doi.org/10.1007/978-3-030-23943-5_19

[2] Sejnowski T. J., "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences of the United States of America*, vol.117, no.48, pp: 30033–30038, December 2020. https://doi.org/10.1073/pnas.1907373117

[3] Zorins A. and Grabusts P., "Artificial Neural Networks and Human Brain: Survey of Improvement Possibilities of Learning," *In Proceedings of the 10th International Scientific and Practical Conference*, pp:228-231, Rēzekne, Latvia, 2015, http://dx.doi.org/10.17770/etr2015vol3.165

[4] Pranckevičius T. and Marcinkevičius V., "Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Baltic Journal of Modern Computing*, vol.5, no.2, pp:221-232, January 2017. http://dx.doi.org/10.22364/bjmc.2017.5.2.05

[5] Holzinger A., "Introduction to Machine Learning & Knowledge Extraction (MAKE)," *Machine Learning and Knowledge Extraction- MDPI*, vol.1, no.1, pp:1-20, https://doi.org/10.3390/make1010001

[6] Kersting K., "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines," *Frontiers in Big Data*, Vol.1, Article 6, pp:1-4, November 2018, https://doi.org/10.3389/fdata.2018.00006

[7] Tanuka M., "A Beginners Approach to Machine Learning Algorithms," August 2018, Article link: https://tanukamandal.com/2018/08/16/beginners-approach-to-machine-learning-algorithms/

[8] Fu Z., Golden B. L., Lele S., Raghavan S., Wasil E. A., "A Genetic Algorithm-Based Approach for Building Accurate Decision Trees," *INFORMS Journal on Computing*, vol.15, no.1, pp:3-22, February 2003. https://doi.org/10.1287/ijoc.15.1.3.15152

[9] Chen J., Wang X., and Zhai J., "Pruning Decision Tree Using Genetic Algorithms," *In Proceedings of International Conference on Artificial Intelligence and Computational Intelligence- IEEE,* pp:1-6, Shanghai, China, 7-8 November 2009. https://doi.org/10.1109/AICI.2009.351

[10] Jankowski D. and Jackowski K., "Evolutionary Algorithm for Decision Tree Induction," *In Proceedings of International Conference on Computer Information Systems and Industrial Management (CISIM)-Springer*, pp:23-32, Ho Chi Minh City, Vietnam, November 2011. https://doi.org/10.1007/978-3-662-45237-0_4

[11] Khanbabaei M. and Alborzi M., The Use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring, International Journal of Managing Information Technology, vol. 5, no.4, pp:13-31, November 2013. https://doi.org/10.5121/ijmit.2013.5402

[12] Muslim M. A., Herowati A. J., Sugiharti E., and Prasetiyo B., "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease," *In Proceedings of International Conference on Mathematics, Science and Education, - Journal of Physics-IOP Publishing*, pp:1-9, Sayangan, Indonesia, 18-19 September 2017. https://doi.org/10.1088/1742-6596/983/1/012062

[13] Fisher R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol.7, no.2, pp:179-188, September 1936. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[14] Forina M., Leardi R., Armanino C., and Lanteri S., "PARVUS: An extendable package of programs for data exploration, classification and correlation," *Journal of chemometrics -Elsevier, Amsterdam, ISBN: 0-444-43012-1*, March 1990. https://doi.org/10.1002/cem.1180040210

[15] Bohanec M. and Rajkovic V., "Knowledge acquisition and explanation for multi-attribute decision making*," In Proceedings of International Workshop on Expert Systems and their Applications*, Avignon, France. pages 59-78, 1988.

[16] Evett I. W. and Spiehler E. J., "Rule Induction in Forensic Science," *Book: Knowledge Based Systems-ACM Digital Library*, pp:152–160, January 1989

[17] Sporer Z., "IRIS Species Classification—Machine Learning Model," *Morioh website*, June 2020, Article link: https://morioh.com/p/eafb28ccf4e3

[18] Jazuli H., "Using Decision Tree Method for Car Selection Problem," *Medium website*, March 2013, Article link: https://medium.com/machine-learning-guy/using-decision-tree-method-for-car-selection-problem-5272675451f9

[19] Hssina B., Merbouha A., Ezzikouri H., and Erritali M., "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications,* Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, pp:13-19, July 2014. https://doi.org/10.14569/SpecialIssue.2014.040203

[20] Özsoy S., Gümüş G., and Khalilov S., "C4.5 Versus Other Decision Trees: A Review," *Computer Engineering and Applications,* vol. 4, no. 3, pp:173-181, September 2015.

[21] Tripathi M., "Understanding Decision Trees with Python," *Data science Foundation*, May 2020, Article link: https://datascience.foundation/sciencewhitepaper/understanding-decision-trees-with-python

[22] García J. M., Acosta C. A., and Mesa M. J., "Genetic algorithms for mathematical optimization," *Journal of Physics: Conference Series- IOP Publishing*, pp:1-5, 2020, https://doi.org/10.1088/1742-6596/1448/1/012020

[23] Sivanandam S., and Deepa S., "Applications of Genetic Algorithms," *Introduction to Genetic Algorithms- Springer*, pp:317-402, https://doi.org/10.1007/978-3-540-73190-0_10

[24] Mijwil, M. M. and Abttan, R. A., "Applying Genetic Algorithm to Optimization Second-Order Bandpass MGMFB Filter," *Pertanika Journal of Science and Technology*, vol.28, no.4, pp. 1413–1425, October 2020**.** https://doi.org/10.47836/pjst.28.4.15

[25] Gomez F., Quesada A., and Lopez R., "Genetic Algorithms for Feature Selection," *Neural Designer*, Article link: https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection