https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis

Edward E. Ogheneovo[1,*], Promise A. Nlerum[2]

[1]Dept. of Computer Science, University of Port Harcourt, Nigeria.
[2]Dept. of Computer Science and Informatics, Federal University, Otuoke, Nigeria.

***Abstract*— *Decision trees are very important machine learning algorithms used for the classification and predictive analytic purposes in computer science and related disciplines. ID3 decision tree algorithm was designed by Quinlan in 1986.The algorithm is based on Hunt's algorithm and was serially implemented. ID3 tree is constructed in two phases: tree building and tree pruning. Data is sorted at every node during the tree building phase to choose the best splitting single attribute. The main ideas behind the ID3 algorithm are: 1) each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. In this paper, ID3, a machine learning algorithm is used to predict weather condition for an out tennis match. The paper demonstrates the use of ID3 decision tree to predict weather conditions with outlooks such as sunny, overcast, and rain; temperature conditions such as hot, mild, and cool; humidity conditions such as high and normal; wind conditions such as weak and strong and the necessary conditions such as yes or no. Based on the results computed using the entropy and information gains, a decision tree is constructed thus providing information for tennis and other sports athletes who wish to play out-door games.*

***Keywords*— Decision trees, ID3, Decision trees, machine learning, entropy, information gain.**

I. INTRODUCTION

Decision Tree [1] [2] is a machine learning algorithm used for data classification and predictive analysis. It is a cross-disciplinary technique for prediction and classification, artificial intelligence, machine learning, knowledge discovery and inductive rule-builder used in data mining, knowledge discovery, machine learning and artificial intelligence problems [3]. Decision learning algorithms guarantees decision trees from the training data to solve classification and regression problem. Basically, they are used as predictive models to predict the value of a target variable by learning from simple decision rules inferred from the features of data. These rules are usually in the form of if-then-statements [4] [5]. Information gain is then used to decide which features to split on at each setup in building the tree. Several types of decision trees are available in literature. They include ID3, CART, C4.5, C5.5, etc. In this paper, we discussed ID3 decision tree algorithm for data classification and predictive analysis.

Iterative Dichotomiser 3 (ID3) decision tree algorithm was designed by Quinlan in 1986 [6] [7]. ID3 algorithm is based on Hunt's algorithm and was serially implemented. ID3 tree is constructed in two phases: tree building and tree pruning [8] [9] [10]. Data is sorted at every node during the tree building phase to choose the best splitting single attribute. The main ideas behind the ID3 algorithm are: 1) each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root to that leaf node, 2) in a "good" decision tree, each non – leaf node should correspond to the input attribute which is the most informative about the attribute amongst all the input attributes not yet considered in the path from the root to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average, and 3) entropy is used to determine how informative a particular input attribute is about the output attribute for a subset of the

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

training data. Entropy is a measure of uncertainty in communication systems. Thus ID3 uses Entropy function and Information gain as metrics [11] [12].

Decision tree algorithms transform raw data to rule-based decision-making trees. Dchomisation is the process of dividing something or object into two completely opposite things. Therefore, the ID3 algorithm is an iterative algorithm that iteratively divide attributes into two groups the dominant attributes and others to construct a tree. It then calculates the entropy and information gains of each attribute. This way, the most dominant attribute can be determined and put on the tree as decision node. After this the entropy and gain scores are calculated again among the other attributes. Then the next most dominant attribute is determined and used. This process continues iteratively, until a decision is reached for that brace. The example of decision-making factors to play tennis in the open (outside) for the previous 14 days is considered and used [13] [14] [15].

In this paper, ID3, a machine learning algorithm is used to predict weather conditions for an out-door tennis match. Section 2 discuss related work and section 3 discussed an example of the use of ID3 decision tree for predicting weather condition. That is, it demonstrates the use of ID3 decision tree to predict weather conditions with outlooks such as sunny, overcast, and rain; temperature conditions such as hot, mild, and cool; humidity conditions such as high and normal; wind conditions such as weak and strong and the necessary conditions such as yes or no. Section 4 discussed the results based on the implementation of the ID# algorithm for weather forecast computed in our example and the output of the program is displayed. Based on the results computed using the entropy and information gains, a decision tree is constructed thus providing information for tennis and other sports athletes who wish to play out-door gam. Finally, section 5 draws the conclusion.

II. RELATED WORK

Kumar and Kiruthika [16] provide a review of the classification algorithms in data mining. Basically, their work considered various decision tree classification algorithms such as CHAD, ID3, C4.5 and C5.5 algorithms by classifying data into disjoint groups. They further discussed the major differences between these decision tree algorithms. Krishna et al. [17] proposed a technique for predicting students' performance in examinations using classification and Regression Trees (CART) decision tree classifier to classify students and predict those at risk based on the impact of four online activities: message exchanging, group wiki content creation, course file opening, and online quiz taking. The correct classification results show that the CART model is very good as a classification algorithm. Their work further tests the ability of CART analysis to predict success in web-based blended learning environment using online interactions stored in the system log file. The number of messages exchanged by team members with their colleagues and instructors as well as the number of contributions made by individual to the team content creation activities were noted. These were used to determine the performance of each student.

Lakshimi et al. [18] proposed an empirical study of decision tree classification algorithms such as CART, ID3, C4.5, CHAD and MARS by providing a brief description of the basic concepts of each of the decision tree algorithms. They pointed out the basic features of each of the algorithms and the advantages and disadvantages of each. They also suggested the various situations when it is best to use any of the algorithms. Patel and Rana [19] provides a survey of decision tree algorithm by making a comparative study of some of the decision tree algorithms such as ID3, C4.5, C5.0 and CART. The authors also discussed the basic features of each algorithm, their advantages and disadvantages and the challenges in each of the algorithms. They concluded that the performance of these algorithms strongly depends on the entropy, information gain and the nature of data used in the classification.

Mianye et al. [20] proposed a technique for predicting the performance of decision tree algorithms using in-depth review of the various techniques used. Comparative study of various decision tree algorithms are discussed. Yang et al. [21] proposed an improved ID3 algorithm for medical classification or the prediction of diseases. Using a heuristic strategy, the improved ID3 algorithm solves the problem of multi-value biasness when selectin test/split attributes. The technique also solves the problem of numeric attribute discretization by storing the classifier model in the form of rule-based for easier model in the form of rule-based for easy understanding and memory usage. The result obtained shows that the improved ID3 performs better using criteria obtained shows that the improved ID3 performs better using criteria such as accuracy, stability, and error corrections.

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

Oshoiribor et al. [22] proposed the use of ID3 decision tree for tax fraud control and prevention. They used ID3 decision tree classifier to classify tax payers into tears for proper monitoring, control and reduction of fraudulent activities in the collection of taxes. First, they gathered data, they then provide a profitability model for determining the profit earned. The profitability model is then used to compute the amount of tax due to each worker or business thus providing an automated system for tax collection so as to ensure issues of cash suppression and diversion and controlled. Adhatrao et al. [23] proposed a technique for predicting students' performance using ID3 and C4.5 decision tree classifiers. They analyzed dataset containing information about students using criteria such as gender, marks scored and rank in entrance examinations and their previous year's results to predict the general and individual performance of the students in future examinations.

III. METHODOLOGY

In this work, we demonstrate the use of ID3 decision tree to predict weather conditions with outlooks such as sunny, overcast, and rain; temperature conditions such as hot, mild, and cool; humidity conditions such as high and normal; wind conditions such as weak and strong and the necessary conditions such as yes or no. Table 1 provides a summary of decision-making factors or necessary conditions that can lead to whether tennis players can play the game of tennis or not.

Table 1: Decision making factors to play tennis outside

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Week	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Using Table 1, we can summarize the ID3 algorithm as follows:

Entropy

The first thing to do is to calculate the entropy. Decision column consists of 14 instances with two possible outcomes "yes" or "no." There are 9 decisions labelled **yes** while the remaining 5 are labelled **no**. To calculate the entropy, we have:

$$\text{Entropy (S)} = \sum - P(1) \, . \log_2 P(1)$$

$$\text{Entropy (Decision)} = - P(Yes). \operatorname{Log}_2 P(Yes) - P(No) \, . \operatorname{Log}_2 (No)$$

$$\text{Entropy (Decision)} = - (9/14) \, . \operatorname{Log}_2 (9/14) - (5/14) \, . \operatorname{Log}_2 (5/14) = 0.940$$

Next we determine the most dominant factor for decisioning.

Wind Factor on Decision

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

The wind factor are either strong or weak. There are 6 strong and 8 weak wind factors. To determine the gain, we compute the wind factor using the formula:

Gain (Decision, Wind) = Entropy (Decision) - ∑[P(Decision | Wind). Entropy (Decision | Wind)]

Strong Wind Factor on Decision

For the strong wind attribute, we have:

Table 2: Strong wind factor on decision

Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	Hot	High	Strong	No
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

As stated earlier, there are 6 instances of strong wind. Decision is divided into two equal parts of 3 each for **yes** and **no**.

Entropy (Decision | Wind = Strong) = - P(No). Log_2 P(No) – P(Yes). Log_2 P(Yes)

= - (3/6) . Log_2 (3/6) – (3/6) . Log_2 (3/6)

= 1

Weak Wind Factor on Decision

For the weak wind attribute, we have:

Table 3: Weak wind factor on decision

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Week	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

As seen in Table 3, there are 8 instances for weak wind out of which 2 has **no** as decision and 6 has **yes** as decision. We now calculate the entropy.

Entropy (Decision | Wind = Weak) = - P(No). Log_2 P(No) – P(Yes). Log_2 P(Yes)

Entropy (Decision | Wind = Weak) = - (2/8) . Log_2 (2/8) – (6/8) . Log_2 (6/8)

= 0.811

Therefore, using the gain equation:

Gain (Decision, Wind) = Entropy (Decision) – [P(Decision | Wind = Weak). Entropy (Decision | Wind = Weak)] – [P(Decision | Wind = Strong). Entropy (Decision | Wind = Strong)].

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

∴ Gain (Decision, Wind) = 0.940 – [(8/14) . 811] – [(6/14) . 1] = 0.048

At this stage, calculation for wind column has been completed. Now we need to apply same calculations for other columns to determine the most dominant factor on decision.

Information Gain

Information gain is a property that measures how well a given attribute separates the training examples based on their target classification.

Information Gain (S, A) = Entropy (S) - ∑ [P(S/A). Entropy (S/A)]

To determine the outlook factor, there are three labels for outlook attribute: Sunny, Overcast, and Rain. Using the information gain, we have:

Gain (Decision, Outlook) = Entropy (Decision) - ∑[P(Decision | Outlook) . Entropy (Decision | Outlook])

Since there are 3 labels, we have

Therefore, Gain (Decision, Outlook) = Entropy (Decision) – [P(Decision | Outlook = Sunny . Entropy (Decision | Outlook = Sunny)] – [P(Decision | Outlook = Overcast) . Entropy (Decision | Outlook = Overcast)] – [P(Decision | Outlook = Rain) . Entropy (Decision | Outlook = Rain)]

As this state, we need to calcite (Decision | Outlook = Sunny), (Decision | Outlook = Overcast), and (Decision | Outlook = Rain) respectively.

Sunny Outlook Factor on Decision

Table 4: Sunny Outlook Factor on Decision

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Table 3.O shows that there are 5 instances of Sunny Outlook with decision of 3 items **No** and 2 items **Yes**. These are used to compute the entropy for Sunny Outlook as shown below.

Entropy (Decision | Outlook = Sunny) = - P(No) . Log_2 P(No) – P(Yes) . Log_2 P(Yes)

∴ Entropy (Decision | Outlook = Sunny) = - (3/5) . Log_2(3/5) – (2/5) . Log_2 (2/5)

1. Gain (Outlook = Sunny | Temperature) = 0.570
2. Gain (Outlook = Sunny | Wind) = 0.019
3. Gain (Outlook = Sunny | Humidity) = 0.970

Here, humidity is the decision because it produces the highest score if outlook were sunny. That is, decision will always be **No** if humidity were high. This is shown in Table 5.

Table 5: High Humidity

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

On the other hand, decision will always be **Yes** if humidity were normal. This is shown in Table 6.

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

Table 6: Normal Humidity

Day	Outlook	Temperature	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Finally, it means we need to check the humidity and decide if outlook were sunny.

Overcast Outlook Factor on Decision

Table 7: Overcast Outlook Factor on Decision

Day	Outlook	Temperature	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

As seen in Table 7 decision will always be **Yes** if outlook were overcast

Other Factors on Decision

Using similar process to calculate the other columns (i.e., Outlook, Temperature, and Humidity), we have:

1. Gain (Decision, Outlook) = 0.246
2. Gain (Decision, Temperature) = 0.029
3. Gain (Decision, Humidity) = 0.151

It can be seen that the outlook factor on decision produces the highest score. Thus outlook decision will appeared at the root node of the tree.

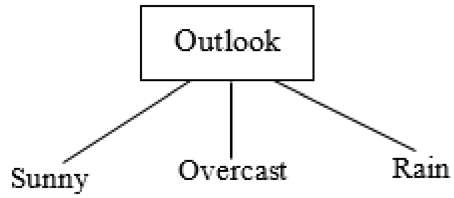

Fig.1: Root decision on the tree

Rain Outlook on Decision

Table 8: Rain Outlook on Decision

Day	Outlook	Temperature	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Week	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

https://dx.doi.org/10.22161/ijaers.74.60 ISSN: 2349-6495(P) | 2456-1908(O)

1. Gain (Outlook = Rain | Temperature)
2. Gain (Outlook = Rain | Humidity)
3. Gain (Outlook = Rain | Wind)

In this case, wind produces the highest score if outlook were rain. Thus we need to check wind attribute in second level if outlook were rain. Therefore, the decision will always be **Yes** if wind were weak and outlook were rain.

Table 9: Decision when Wind are weak and Outlook were rain

Day	Outlook	Temperature	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Week	Yes
10	Rain	Mild	Normal	Weak	Yes

Table 10: Decision is no when wind are strong and outlook were rain

Day	Outlook	Temperature	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

IV. RESULTS AND DISCUSSION

Based on the various computations done using entropy and information gain, we constructed the decision tree as show in figure 3. This decision tree provides a summary of all the computation done so far in this work.

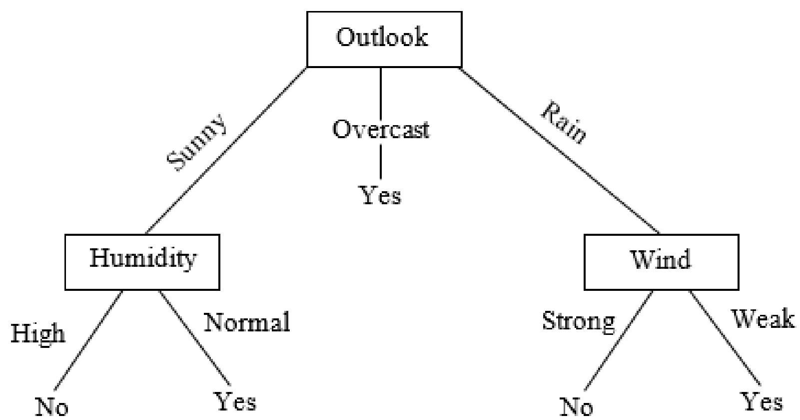

Fig.3: Decision tree for playing tennis outside

As seen in the decision trees in figure 3, weather outlook can be sunny, overcast or raining. Whenever the weather is sunny, there is humidity which can be high or normal and whenever there is rainfall there is wind, which of course could be strong or weak.

V. CONCLUSION

ID3 decision tree algorithm was designed by Quinlan in 1986. The algorithm is based on Hunt's algorithm and was serially implemented. ID3 tree is constructed in two phases: tree building and tree pruning. Data is sorted at every node during the tree building phase to choose the best splitting single attribute. The main ideas behind the ID3 algorithm are: 1) each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. In this paper, ID3, a machine learning algorithm is used to predict weather conditions for an out tennis match. The paper demonstrates the use of ID3 decision tree to predict weather conditions with outlooks such as sunny, overcast, and rain;

temperature conditions such as hot, mild, and cool; humidity conditions such as high and normal; wind conditions such as weak and strong and the necessary conditions such as yes or no. Based on the results computed using the entropy and information gains, a decision tree is constructed thus providing information for tennis and other sports athletes who wish to play out-door gam.

REFERENCES

[1] Rokach, L., Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers: A Survey. IEEE Transaction on Systems, Man, and Cybernetics – Part C: Applications and Reviews, Vol. 35, 4, pp. 476-487.

[2] Barros, R. C., M. P. Basgalupp, A. C. P. L. F. Carvalho, A. A. Freitas (2012). A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews, Vol. 42, No. 3, pp. 291–312.

[3] Painsky, Amichai; Rosset, Saharon (2017). Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 39, No. 11, pp. 2142–2153.

[4] Barros, R. C., R. Cerri, P. A. Jaskowiak, A. C. P. L. F. Carvalho (2011). A Bottom-Up Oblique Decision Tree Induction Algorithm. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011). pp. 450–456.

[5] Letham, B., C. Rudin, T. McCormick, D. Madigan (2015). Interpretable Classifiers Using Rules and Bayesian Analysis: Building A Better Stroke Prediction Model. Annals of Applied Statistics. Vol. 9, No. 3, pp. 1350–1371.

[6] Quinlan, J. R. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research, 4: 77 – 90.

[7] Witton, I. H., F. Eibe, M. A. Hall (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, San Francisco, USA.

[8] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques, Infomatica 31 (2007) 249 – 268.

[9] Pandya, R. and J Pandya (2015). C5.0 Algorithm to Improve Decision Tree with Feature Selection and Reduced Error Pruning. Int'l Journal of Computer Applications, Vol. 117, No 16, pp. 18 – 21.

[10] Le, Q. V., M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, A. Y. Mg (2012). Building High – Level Features Using Large Scale Unsupervised Learning. In Proceedings of the 29th Int'l Conference on Machine Learning, Edinbargh, Scotland, UK.

[11] Xiaohu, W. W. Lele, N. Nianfang (2012). An Application of Decision Tree Based on ID3. In Proceedings of 2012 Int'l Conference on Solid State Devices and Material Science. Physics Procedia, Vol. 25, pp. 1017-1021.

[12] Ben-Gal, I. and C. Trister (2014). Parallel Construction od Decision Trees with Consistently Non-Increasing Expected Number of Tests. Applied Stochastic Models in Business and Industry.

[13] Bhatt, H., S. Mahta and L. R. Dinelo (2015). Use of ID3 Decision Tree Algorithm for Placement Prediction. Int'l Journal of Computer Science and Information Technologies, Vol. 6, No. 5, pp. 4785-4788.

[14] Prasanthi, L. S. and R K. Kumar (2015). ID3 and Its Applications in Generation of Decision Trees Across Various Domains.

[15] Prasanthi, L. S., R K. Kumar and K. Srinivas (2016). An Improved ID3 Decision Tee Algorithm on Imbalance Datasets Using Strategic Oversampling. Int'l Journal of Database Theory and Application, Vol. 9, No. 5, pp. 241-250.

[16] Kumar, S. V. K. and P. Krathika (2015). An Overview of Classification Algorithm in Data Mining. Int'l Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 12, pp. 255-257.

[17] Krishua, Mk. K., B. S. B. P. Rani, G. K. Chakravarthi, B. Madhavrao and S. M. B. Choudary (2020). Int'l Journal of Innovative Technology and Exploring Engineering, Vol 9, Issue 3, pp. 3349-3356.

[18] Lakshmi, B. N., T. S. Indumathi and N. Ravi (2015). An Empirical Study of Decision Tree Classification Algorithms. Int'l Journal of Science, Engineering and Technology, Vol. 4, Issue 1, pp. 3705-3709.

[19] Patel, S., M. Agrawal and V. R. Baviskar (2015). Efficient Processing of Decision Tree Using ID3 & Improved C4.5 Algorithms. Int'l Journal of Computer Science and Communication Technologies, Vol. 6, No. 2, pp. 1956-1961.

[20] Mienge, I. D., Y. Sun and Z. Wang (2019). Predicting Performance of Improved Decision Tree-Based Algorithms: A Review. In Proceedings of the 2nd Int'l Conference on Sustainable Materials Processing and Manufacturing (SMFM'19). Procedia Manufacturing, pp. 698-703.

[21] Yang, S. J.-W. Jin and J/-Z. Guo (2018). An Improved ID# Algorithm for Medical Data Classification. Computers and Electrical Engineering, Vol. 65, pp. 474-487.

[22] Oshoiribbor, E. O., M. John-OtumuAdetokunbo and C. E. Ojieabu (2016). Journal of Computer Science an Information Technology, Vol. 4, No. 2, pp. 73-93.

[23] Adhatrao, K, A. Guykar, A. Dhawan, R. Jha and V. Honrao (2013). Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms. Int'l Journal of Data Mining & Knowledge Management Process, Vol. 3, No. 5, PP. 39-52.