Vol. 01, Iss. 3 (2024) 95-102, DOI: 10.61552/JSI.2024.03.001

Journal of Sustainable Development Innovations

www.jsi.aspur.rs

Using a Tree-Based Algorithm to Categorize Students Who Have Dropped Out Evidence from Universities in Vietnam

Nguyen Huu Tan[a], Nguyen Duc Huu[b]*

[a]*Phenikaa University, Ha Noi, Vietnam,*
[b]*Trade Union University, Ha Noi, Vietnam.*

Keywords:

Machine learning, algorithms, data mining, Tree-based, student, university, Vietnam.

* Corresponding author:

Nguyen Duc Huu
E-mail: huund@dhcd.edu.vn

Received: 14 April 2024
Revised: 21 May 2024
Accepted: 16 July 2024

A B S T R A C T

Purpose: The purpose of this article is to identify and comprehend the mechanism of operation of the Decision Tree and Random Forest algorithms. comprehend precise implementation steps, better comprehend nature, and synthesize experience in the practice process.

Methodology: The study was conducted based on an analysis of dropout data from Trade Union University, University of Social Labour, and Phenikaa University in Vietnam.

Result: The article investigates and develops a machine learning model for classifying students as dropouts or not. Develop scales to measure the accuracy and dependability of the model.

Value: The article assesses the model's outcomes, performance, and accuracy. Give your suggestions for future development and improvement of the topic. Propose several solutions to the challenge of identifying students as dropouts, based on the real scenario employed by the algorithm during the analysis.

1. INTRODUCTION

Machine learning is an area of artificial intelligence that involves the study and development of techniques that allow systems to "learn" automatically from data to solve specific problems. Currently, along with the explosion of internet information sharing and the advancement in the construction of highly capable computers, Machine learning is being applied more and more and improved, especially predictive (predicting material prices, estimating demand for use,...) and recognition (handwriting recognition, sign recognition, object recognition,...) In agriculture today, Machine Learning is being applied quite specifically to the problem of recognizing plant varieties in the wild and identifying to classify fruits based on color.

The current directions for most image recognition are available based on the characterization data warehouse There are huge cumulative losses due to such diseases that reduce productivity and increase economic losses in the agricultural sector. The agricultural sector needs to sustain and thrive from such obstacles in order to be highly profitable In this paper, we propose to separately identify and locate strawberries - a highly economical plant in the image based on Machine Learning applications and Python programming language. Aimed at the self-improving characteristics of Machine Learning and the relevance of python programming language to image objects. Machines are used instead of the human eye for measurement and evaluation With the strength of self-improvement, Machine Learning is perfectly suited for recognizing and building databases to help the system be optimized [1-5].

In the field of Machine Learning, tree-based algorithms have always been a widely used group of methods. Possesses good interpretation, accompanied by excellent scalability. It is not difficult to understand that this group of algorithms is currently being applied in countless practical problems, bringing very high performance and the ability to adapt to a variety of situations. In this study, the authors want to learn about how Decision Tree and Random Forest algorithms work. Object detection is a computer technology that deals with computer vision and image processing, which involves the detection of instances of semantic objects of a certain class (such as people, buildings, or cars) in digital images and videos Well-studied areas addressing object detection include: face detection and pedestrian detection. Object detection has had applications in many fields where computers are used, including image retrieval and video surveillance. Besides, tree-based algorithms have a dialectical attachment to the machine learning process. Machine learning is very close to statistical inference, although terminology differs Some practical applications show that machines can "learn" how to classify, for example, whether email can be considered spam and automatically sort messages into corresponding folders. Machine learning is also considered part of artificial intelligence. Machine Learning algorithms build a model based on sample data, called training data, to make predictions or decisions that have not been explicitly programmed to do so If the analysis is based on experience and a sufficient number of honest background examples are available, then a machine learning method is shown Machine learning algorithms used in a variety of applications where it is difficult or impossible to develop conventional algorithms to perform the necessary tasks. There are many studies on mathematical optimization, which have provided methods, theories and applications to the field of Machine Learning. With the advancement of social science, artificial intelligence technology has also developed rapidly, and humans have made breakthrough advances in the study of machine learning By the application of the tree-based algorithm, the authors want to research and build a machine learning model that classifies students as dropouts or not. Develop scales to assess the accuracy and reliability of the model, thereby evaluating the results, performance as well as accuracy of the model and proposing some suggestions for classifying students who are dropouts from the practice of some universities in Vietnam today [6-8].

2. METHODOLOGY AND DATA BASE

A computer program is considered to learn how to perform a class of tasks through experience, for a competency scale if using competency we measure the program's performance capacity to improve after experience" (machine learned). One of the other focuses of machine learning is to achieve generalization, the property of a program that can work well with data that it has never encountered before (unseen data) in order to gradually gain some judgment and be able to update in real time with the data it encounters. For the study, the authors used statistical data on the number of Vietnamese dropouts at Trade Union University, University of Social Labor and Phenikaa University. In addition, the article uses data sources from UC Irvine, a reputable website specializing in providing data for common problems in Machine Learning. This is a data file that has been made public, which can be retrieved directly. This data is generated from a higher education institution (collected through several separate databases) regarding students pursuing different university degrees. The dataset includes information known at the time students enroll – learning pathways, demographics, and socioeconomic factors. The problem is formulated in the form of a three-

category classification task (Dropout, Enrolled and Graduated) at the end of the normal time of the course. This is the data supported by the SATDAP - Portugal program [7-10].

A decision tree is a structured hierarchical tree, used to classify objects based on sequences of rules. This is one of the supervised machine learning models that works efficiently and powerfully for classification, prediction, or regression problems [11].

Decision Tree is actually an algorithm that simulates how people make predictions and think. We will create a decision tree where [12]:

- Each node represents a characteristic (properties, properties of data).
- Each branch represents a rule.
- Each leaf represents a result (a specific value or a continuation branch, or a layering result) [12].

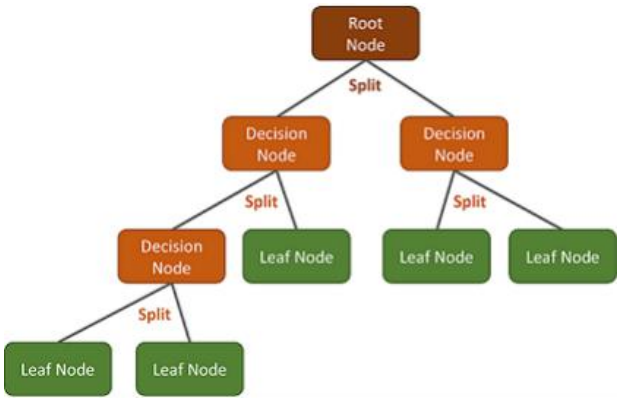

Fig. 1. Illustration of the Decision Tree.

- The study also applied the ID3 (Iterative Dichotomiser 3) algorithm commonly used in the construction of decision trees. ID3 uses divide and conquer to divide the dataset into smaller pieces. Each division, the algorithm selects the optimal attribute to create the greatest uniformity in each section. The ID3 algorithm performs the steps in turn as follows [13]:
  - Calculate the entropy of the dataset.
  - For each attribute/feature: ID3 calculates entropy for all taxonomies and IG levels of that feature.
  - Look for features to get maximum IG.
  - Repeat the steps performed until the desired tree is reached

The formula for calculating Entropy can be represented as follows [14]:

$$Entropy(S) = \sum_{i=1}^{C} - \frac{S_i}{S} \times log_2(\frac{S_i}{S})$$

Entropy (S): The impurity of the dataset

S: Original dataset (sample set)

Si: Subset of data, created from sample set

C: The number of layers included in the data

Information Gain When applying the division of attribute groups, we realize that the Entropy of each group after division is evenly reduced compared to the original Entropy of the data. Entropy values represent how chaotic the data is, so as Entropy decreases the data is more orderly (or so to speak, they provide more information). Therefore, the decrease in entropy is called Information Gain and has the following formula [15]:

$$IG(S, A) = Entropy(S) - \Sigma_{v \in A} \frac{|S_v|}{|S|} \times Entropy(A_v)$$

IG (S,A): Is the Information Gain value when using property A to divide data set S into subsets Sv

Entropy (S): The uncertainty or chaos of the initial data set S

S: Original dataset (sample set) before dividing

Si: Subset of data, created from sample set

A: This is a feature in the data.

v: As a value of characteristic A.

Entropy (Av): Entropy information of characteristic A when carrying the value v.

Classification and Regression Tree (CART) is an algorithm born in 1984, developed by Breiman, Friedman, Olshen, Stone This algorithm is often used in classification and numerical value prediction problems. The Gini Index is an index that demonstrates the level of misclassification when we randomly select an element from a data set. To put it simply, this is a method of measuring the "purity" of a data set. This value is determined by the formula [7,16,17]:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

D: The Dataset Needs to Compute the Gini Index.

C: The number of classes present in the dataset.

pi: The proportion of denominators belonging to class i in data set D.

The lower the Gini Index value, the more homogeneous the data set > the selected attribute as well as the more optimal the division values. The more homogeneously the subgroups are divided, the more effective the predicted results are. In addition, after each split, the GiniSplit concept is used to measure the homogeneity of the data after each split. Gini Split of Data Splits Using Attribute A, v-Value Determined by the Formula [18,19].

$$Gini_{Split} = \Sigma_{v \in Values(A)} \frac{|D_v|}{|D|} \times Gini(D_v)$$

GiniSplit: Is the Gini Index after dividing data using an A attribute.

Values(A): The set of possible values of attribute A.

| Dv |:Number of samples in set D, whose value of attribute A is v.

| D |: The total number of samples contained in dataset D.

Gini(Dv):Is the Gini Index of Subset D

3. RESULTS AND FINDINGS

Applying the knowledge just learned above, we will build a Decision Tree model to implement this student classification problem. The team decided to choose the CART (Classification and Regression Tree) algorithm to build the tree in this problem. First, we will need to screen the features in the article to include in the model (since the dataset has 36 features, this number is too large to include in the model. If all are included, overfitting is highly likely). And to sift through the features needed for the topic, the team used the concept of Feature Importance – a scale of the importance of each feature to determine whether the feature should be included in the model. The use of FI allows us to reduce less important features, reduce the number of dimensions and increase computational performance. It also helps us understand specifically how the model makes predictions, making it easier for us to optimize the model [20].

```
Feature 7: Curricular units 2nd sem (approved) - Importance: 0.5598
Feature 3: Admission grade - Importance: 0.0912
Feature 4: Tuition fees up to date - Importance: 0.0684
Feature 0: Previous qualification (grade) - Importance: 0.0672
Feature 5: Curricular units 2nd sem (enrolled) - Importance: 0.0593
Feature 8: Curricular units 2nd sem (grade) - Importance: 0.0555
Feature 2: Father's qualification - Importance: 0.035
Feature 6: Curricular units 2nd sem (evaluations) - Importance: 0.0335
Feature 1: Mother's qualification - Importance: 0.0301
```

Fig. 2. Selected features for modeling.

After selecting Features, now it's time to start building a model based on the prepared dataset. The authors initiated the training based on 80% of the prepared data, the remaining 20% for testing.

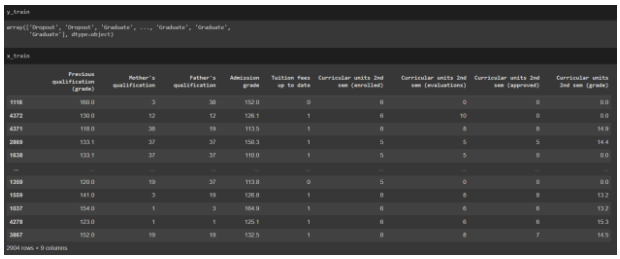

Fig. 3. Data after subdivision into training.

Initialize the 'spicy' class: The 'spicy' class is initialized with a parameter 'max_depth', which represents the maximum depth of the tree.

- Calculating Gini Impurity: The '_gini' method calculates Gini impurity for a dataset.
- Divide the dataset: The 'split_dataset' method divides the dataset into two parts based on the threshold and metrics of the characteristic.
- Finding the Best Divide Point: The 'find_best_split' Method That Finds the Characteristics and Thresholds That Produces the Best Gini Impurity When Dividing a Dataset.
- Build a tree: The '_build_ tree' method uses recursion to build a decision tree. It divides the dataset, finds the best dividing point, and builds the subbranches until the maximum depth is reached or all leaves contain only one layer or the stop condition is reached. In addition, the team also sets the property 'min_leaf' = 5 as a Pruning operation, which makes the initialization process faster and avoids overfitting
- Prediction: The 'predict' method uses the built-in tree to predict labels for new patterns. It traverses the plant from root to leaf based on the value of the characteristics of the pattern.
- Model and prediction training: The model is trained by calling the 'fit' method with training data and labels. The depth of the tree is controlled by the parameter 'max_depth'. Prediction is made by calling the 'predict' method with the test data

After running the results, the team recounted the Confusion Matrix values of the model as follows.

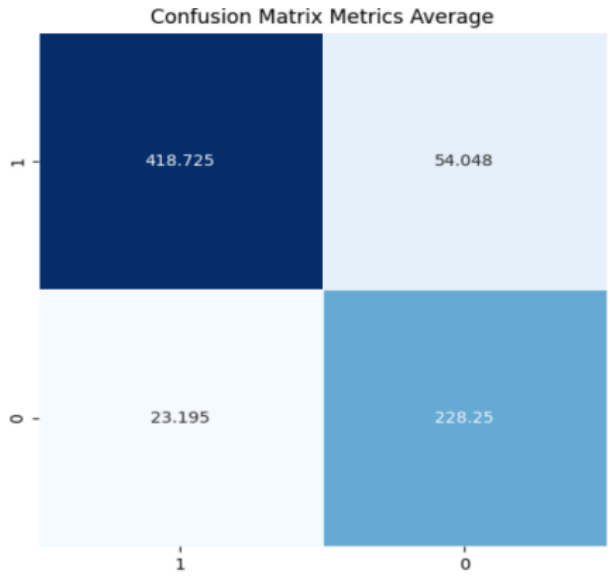

Fig. 4. Average Confusion Matrix calculation result of the Decision Tree model.

Statistics of interesting values revolving around the Confusion Matrix of 40 newly created trees are as follows:

Value	City	FP	FN	TN
Min	389,000	19,000	12,000	203.00
Max	438,000	73,000	34,000	252.00
Mean	418,725	54,048	23,195	228.25

Fig. 5. Table of Confusion Matrix values of the Decision Tree model.

The running model takes ~4 minutes and 30 seconds in total if running on Visual Code (running on Google Colab will take longer, depending on the time of day.). Based on the Confusion Matrix value, the accuracy of the model is about 89.115%, a positive value.

```
#Tính Accuracy của Decision Tree
tong_dung_decisiontree = 0
tong_decisiontree = 0
list_of_arcuracy_decisiontree = []
for i in range(40):
    tong_dung_decisiontree = tong_dung_decisiontree + list_of_cm[i][0][0] + list_of_cm[i][1][1]
    tong_decisiontree += np.sum(list_of_cm[i])
    list_of_arcuracy_decisiontree.append(tong_dung_decisiontree/tong_decisiontree)
accuracy_dt = tong_dung_decisiontree/tong_decisiontree
print(accuracy_dt)
```

```
0.8911501377410468
```

Fig. 6. Average accuracy of 40 Decision Tree model runs.

Based on the results obtained by the Decision Tree model, the authors draw some conclusions as follows:

- The model works relatively well with the actual problem this time. The average accuracy stands at >89%, a "safe" threshold in this case.
- The running time of the algorithm is about 4 minutes and 30 seconds, which is relatively fast.
- The model produces average results of ~54 False Positive (FP) and 23 False Negative (FN). That is, there were 54 cases of not dropping out of school but being filtered by the model, whereas there were 23 cases of dropping out but the model did not take into account. FN ratio ~3.2%, this value may be improved in the future.
- To prevent overfitting, the group exercise has also set the depth of the tree. In terms of improving the model through the use of hyperparameters, perhaps the team will need more research in the future to do so

4. DISCUSSION AND CONCLUSION

Based on the information collected during the study, the authors discovered some problems in the dropout situation of students. To solve this problem, recommendations can be considered:

Enhancing the position, roles and responsibilities of class teachers, subject teachers and mass organizations The roles of class teachers, subject teachers and mass organizations are of paramount importance in detecting and preventing dropouts. In particular, the role of the homeroom teacher is very important in educating students. Class teachers are the people closest to them, who directly guide and manage them throughout the learning process. Therefore, improving the position and role of class teachers in student management will have a positive impact on students' psychology, motivation and interest in learning, thereby reducing dropouts. In addition to the role of the head teacher, we cannot fail to mention the role of the subject teacher. They are the mentors of class teachers in student management. Not all the time, everywhere class teachers can care about all their students but need active support from subject teachers. Analyzing the causes of student dropout, we see that a large number of students drop out of school because they do not know how to manage their time, leading to too much free time. Therefore, mass organizations such as the Youth Union, Student Union ... Healthy playgrounds must be regularly organized for children. Those playgrounds can be organizing exchange sessions, exchanging experiences in

learning, organizing life skills education sessions, learning about careers ... In particular, in cooperation with homeroom teachers, subject teachers organize for students to participate in scientific research, make their own learning materials and models suitable for each subject [21-23].

Help students think about future careers, identify career interests and start steps to get the right job Apprentices are always looking forward to how they will become workers in the future, so raising career awareness will help them better orient. When they understand their career, they will be more confident, thereby helping them feel secure in exploring and researching scientific products. Students who understand the importance of the career will satisfy the motivation of scientific awareness stemming from the need to study, curiosity and curiosity arising in the learning process not bored of learning due to orientation from the beginning before entering school and thus will reduce the phenomenon of dropping out. In order for students to understand their careers, they can be done right in the admissions counseling, career guidance takes place at middle and high schools. Students will be more fully aware of their chosen career such as: Clearly defining their responsibilities, study and work goals after graduation, cultivating more love and passion for work for them [24].

Create conditions for students to regularly interact with businesses. Good coordination between schools and businesses will have a solid basis for analyzing and evaluating labor quality. The school needs to have investigative activities about the cooperation with enterprises, about the efficiency of using human resources recruited from the school, find out what its students after graduation have met the requirements of the business, what is missing that needs to be adjusted, supplement. From there, build a training program suitable to the requirements of the business. The school organises participatory career seminars. Here businesses can clearly indicate the strengths and weaknesses that students also face. Regularly provide opportunities for students to be exposed to special businesses near the end of the course. During the internships, enterprises will clearly present the tasks that students need to do, evaluation criteria, and notify the student's internship results to the lecturer in charge. Thus, the training link between schools and businesses is now an objective need stemming from the interests of both sides, this link is both inevitable and highly feasible [25,26].

Organize students to participate in creative experiential activities In some countries around the world, creative experiential activities have been included as educational programs that work in parallel with subject education for nearly a decade. Creative experiential activities are an element of the national basic curriculum (together with compulsory subject systems, elective activities) and are implemented throughout. . Although creative experiential activities are not a simple subject, they are still within the framework of the national general education program and play an important role in realizing educational goals. The contents of creative experience activities mentioned in the national program include: Autonomy activities, club activities, volunteer activities, career guidance activities. Creative experiential activities help students learn and live according to ethical values, acquire and practice soft skills, social integration - sharing and disseminating experiences to different children, backgrounds and ethnic groups. On the other hand, the dynamic curriculum content will give students the opportunity to experience: Experience in nature, integrated learning in a creative way; provide opportunities for them to create their own products; combined with educational and social, emotional learning and educational values; give the opportunity to be enjoyed and have fun [13,27].

Strengthen support policies for students. Universities need to exploit resources in terms of capital, facilities, technology, means and management methods of organizations and units through projects, cooperation programs, grants for people to contribute to help, encourage students to study better, limiting dropouts [28].

The situation of students dropping out of school is no longer an isolated phenomenon but is becoming more and more common, which has reduced the goal and quality of training, especially for vocational training. The proposed solutions mentioned above are all interrelated, interacting with each other. Therefore, to increase efficiency, solutions need to be implemented synchronously. With research tools, the article has focused on

learning and exploring useful knowledge about the Tree-based algorithm group. From concepts, construction ideas, basic algorithms as well as how to implement them, the authors have also successfully applied to a simple practical problem. It can be seen that both Decision Tree and Random Forest models have extremely high potential in the field of Machine Learning, especially for the group of classification and regression problems. These algorithms work with high efficiency, providing a good amount of knowledge without spending too much money. Through the topic, the group had a new perspective on the problem of classifying dropouts at the university level, a topic that is very popular and close but rarely exposed before.

REFERENCES

[1] H. Ahmed, F. Y. Mohsin Abdulazeez, A. Falah, D. Zeebaree, D. Mikaeel Ahmed, and D. Qader Zeebaree, "Predicting university's students performance based on machine learning techniques," *IEEE Xplore*, 2021, doi: 10.1109/I2CACIS52118.2021.9495862.

[2] B. Charbuty, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol.*, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[3] C. Kern and T. Klausch, "Tree-based machine learning methods for survey research," *NCBI*, 2019.

[4] Z. Zhang and Z. Zhao, "Decision tree algorithm-based model and computer simulation for evaluating the effectiveness of physical education in universities," *Complexity,* vol. 2020, pages 1-11, 2020.

[5] V. Matzavela, "Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments," *Computers and Education: Artificial Intelligence*, 2021, doi: 10.1016/j.caeai.2021.100035.

[6] A. Vultureanu-Albişi, C. Bădică, A. Vultureanu, A. Albişi, and C. Bădică, "Improving students' performance by interpretable explanations using ensemble tree-based approaches," *IEEE Xplore*, 2021, doi: 10.1109/SACI51354.2021.9465558.

[7] W. Zhang and Y. Wang, "Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China," *Education and Information Technologies*, 2022, doi: 10.1007/s10639-022-11170-w.

[8] E. Ampomah, Z. Qin, and G. Nyame, "Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models," *Informatica*, vol. 44, pp. 477, 2020, doi: 10.31449/inf.v44i4.3159.

[9] A. Sheshasaayee, "An insight into tree based machine learning techniques for big data analytics using Apache Spark," *IEEE Xplore*, pp. 1740-1743, 2017, doi: 10.1109/ICICICT1.2017.8342833.

[10] E. Fijani, K. K. R. Management, and W. R., "Hybrid iterative and tree-based machine learning algorithms for lake water level forecasting," *Water Resources Management*, vol. 37, pp. 5431-5457, 2023, doi: 10.1007/s11269-023-03613-x.

[11] A. Mansoori, "Optimization of tree-based machine learning models to predict the length of hospital stay using genetic algorithm," *J. Healthcare*, vol. 2023, doi: 10.1155/2023/9673395.

[12] D. Wolff, "Tree-based machine learning approaches for equity market predictions," *Journal of Asset Management*, vol. 20, pp. 273-288, 2019, doi: 10.1057/s41260-019-00125-5.

[13] B. Pham and K. Khosravi, "Application and comparison of decision tree-based machine learning methods in landslide susceptibility assessment at Pauri Garhwal Area, Uttarakhand, India," *Environmental Processes*, vol. 4, pp. 711-730, 2017 doi: 10.1007/s40710-017-0248-5.

[14] J. Lai, Y. Lin, H. Lin, C. Shih, Y. Wang, "Tree-based machine learning models with Optuna in predicting impedance values for circuit analysis," *Micromachines*, vol. 14, no. 2, 2023, doi: 10.3390/mi14020265.

[15] M. Jiang, J. Liu, and L. Zhang, "An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 541, 2020, doi: 10.1016/j.physa.2019.122272.

[16] K. Koc, Ö. Ekmekcioğlu, "Integrating feature engineering, genetic algorithm, and tree-based machine learning methods to predict the post-accident disability status of construction workers," *Automation in Construction*, vol. 131, 2021, doi: 10.1016/j.autcon.2021.103896.

[17] E. Parimbelli, T. Buonocore, and G. N., "Why did AI get this one wrong?—Tree-based explanations of machine learning model predictions," *Artificial Intelligence in Medicine*, vol. 135, 2023, doi: 10.1016/j.artmed.2022.102471.

[18] S. Saha et al., "Comparison between deep learning and tree-based machine learning approaches for landslide susceptibility mapping," *Water*, vol. 13, no. 19, 2021, doi: 10.3390/w13192664.

[19] D. Ernst and P. Geurts, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, 2005. [Online]. Available: https://www.jmlr.org/papers/volume6/ernst05a/ernst05a.pdf.

[20] M. Du, D. Haag, and J. Lynch, "Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: analyses based on SEER," *Cancers*, vol. 12, 2020, doi: 10.3390/cancers12102802.

[21] T. Aravind and B. Reddy, "A comparative study on machine learning algorithms for predicting the placement information of undergraduate students," *IEEE Xplore*, 2019, doi: 10.1109/I-SMAC47947.2019.9032654.

[22] M. Yasir et al., "Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance," *Antibiotics*, vol. 11, iss. 11, 2022, doi: 10.3390/antibiotics11111593.

[23] S. Papadopoulos, E. Azar, "Evaluation of tree-based ensemble learning algorithms for building energy performance estimation," *Journal of Building Performance Simulation*, vol. 11, iss. 3, 2018. [doi: 10.1080/19401493.2017.1354919.

[24] Q. Liu, A. Tang, Z. Huang, L. Sun, "Discussion on the tree-based machine learning model in the study of landslide susceptibility," *Natural Hazards*, vol. 113, no. 2, pp. 887–911, 2022, doi: 10.1007/s11069-022-05329-4.

[25] M. Gene, "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping," *Gene*, vol. 677, pp. 111-118, 2018, doi: 10.1016/j.gene.2018.07.057.

[26] B. F. Murorunkwere, J. Felicien Ihirwe, I. Kayijuka, J. Nzabanita, and D. Haughton, "Comparison of tree-based machine learning algorithms to predict reporting behavior of electronic billing machines," *Information*, vol. 14, iss. 3, 2023, doi: 10.3390/info14030140.

[27] S. A. Eslaminezhad, M. Eftekhari, A. Azma, R. Kiyanfar, and M. Akbari, "Assessment of flood susceptibility prediction based on optimized tree-based machine learning models," *Journal of Water and Climate Change*, vol. 13, pp. 2353, 2022, doi: 10.2166/wcc.2022.435.

[28] M. C. Iban, “An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants,” Habitat International, vol. 128, p. 102660, Oct. 2022, doi: 10.1016/j.habitatint.2022.102660.