Science & Technology Development Journal 2023, 26(2):2828-2837

Open Access Full Text Article

Engineering and Technology: Research Article

A deep learning model of major consulting support

Kha Tu Huynh[1,2,*], **Nga Tu Ly**[1,2]

Use your smartphone to scan this QR code and download this article

ABSTRACT

Introduction: Major selection is always a matter of concern for students who have just graduated from high school and parents who have children to go to universities. Currently, there are many students who selected the wrong major, leading to unexpected learning results and wasting time and money. In fact, many students do not know which majors they are suitable for. The paper proposes a model of decision-making support in choosing majors for students immediately after graduating from high school using deep learning. **Methods**: The model applied the XGBoost algorithm to build a decision tree for classification, mining educational data from which the student's ability and learning propensity are predicted and the appropriate majors are suggested. **Results**: The data used for the system are collected from 1709 students' results at the high school, the survey results on personal interests and personality, the teacher's comments and the results on major selection after graduation. From these data, the authors have built a model to advise students choosing the right major to continue their higher education. **Conclusion**: The model is evaluated and verified through actual experiments with a high accuracy of 86% and proves the contribution of deep learning models to education.

Key words: Deep learning, convolutional neural network (CNN), decision making, majors selection

INTRODUCTION

Majors consulting is always a necessary topic and one of the top activities prioritized by high schools annually. This helps students make the right decisions to choose a major for the higher education level. Career orientation for high school students is considered an important first step in the process of training and developing national human resources. However, it has not been synchronized, and many schools have not implemented it effectively. That is because there are few tools to support this activity yet. Currently, parents and students themselves do not truly appreciate their own abilities, interests and aspirations. In Vietnam, many high school seniors who have not yet assessed the appropriateness of their abilities, conditions, and forte have applied for admissions in majors due to trends or feelings. These have led to many cases of dropping out, changing majors or not being able to continue studying because of failing to meet academic requirements, which have caused a huge waste of time and resources for themselves and their families and society.

Currently, with the continuous development of information technology, the computerization of management in many different fields and activities has created a huge data library that is ready to serve anyone interested in. It is truly one of the most valuable information resources if we use these datasets as a basis for decision support in management to bring significant efficiency. However, the problem here is how we classify the resources to make the most effective use in each specific field. In fact, all high schools have a system to store and manage all students' score data for each school year. At the same time, universities and colleges are now almost using the results of the three-year high school education to consider enrolling students in the fields of study offered by the school. Therefore, we can take advantage of these data sources in combination with data mining techniques to build supporting tools for helping learners make decisions when choosing an appropriate field of study. The main contribution of the paper is to suggest an alternative model to solve the problems for low-end machines by:

- Analyzing steps of the major consulting process, combined with data mining techniques from which XGBoost is applied to classify the educational data.
- Proposing a model to predict students' ability to suggest suitable majors based on academic performance in the high school years.

Following this introduction, the paper presents six contents: problem statements, proposed model, dataset, experimental results, discussion and conclusion.

[1]*International University, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City*

Correspondence

Kha Tu Huynh, International University, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City

Email: hktu@hcmiu.edu.vn

History

- Received: 2023-05-09
- Accepted: 2023-07-20
- Published: 2023-07-23

DOI :

https://doi.org/10.32508/stdj.v26i2.4087

Check for updates

VNU-HCM Press

Cite this article : Huynh K T, Ly N T. **A deep learning model of major consulting support**. *Sci. Tech. Dev. J.* 2023; 26(2):2828-2837.

PROBLEM STATEMENTS

Data mining is increasingly being researched and developed and has achieved remarkable achievements, especially in education. Choosing a suitable career is very important for students as well as their parents. This starts with choosing the right major for the desired career, which is determined by many factors, such as personal ability and interests. Therefore, many parents and students find it difficult to orient their majors without specific analysis or suggestions. The problem is stated as follows: For a student who has just graduated from high school and is preparing for higher education, how to advise them on appropriate majors.

The authors have researched and built a model to suggest suitable majors for learners based on the study results in the last 3 years of high school and other factors such as the student's personality which of the six majors according to John Holland's theory, the teacher's comment to help high school students can self-direct and choose the right majors suitable to their abilities, personality and interests, and help the teachers advise students more accurately in determining their majors. The authors solved the problem by the following steps:

- We researched and built a dataset including academic results, exam results and students' choice of majors when entering university in previous years.
- Data mining and analysis to determine the relationship between course groups and majors and the possibility of success.
- Building a model of academic counseling by combining the trained model above with other factors such as student personality and teachers' comments.
- The XGBoost algorithm is used to build a decision tree to predict the best solutions.

PROPOSED MODEL

Decision Support System

In the 1970s, Scott Morton introduced the first concept of a decision support system (DSS). He defines DSSs as interactive computer systems that help decision makers use data and models to solve unstructured, complex, fuzzy problems with incomplete solutions [1]. Although there is no uniform definition of DSS, all studies have confirmed that the most basic purpose of DSS is to support and improve decision making. A DSS consists of three main components: database, software system and user interface.

A fundamental feature of a DSS is that there must be at least one decision support model. The selection and construction of the model is in the second phase (design phase) of the decision-making process. Models are generalizations or abstractions of real problems into qualitative or quantitative models. It is a process that combines both science (correctness, logic) and art (creativity). A decision-making model typically consists of three basic components:

- *Decision Variables*: These are choices determined by the decision maker.
- *Uncontrollable Variables*: These are variables that are not controlled by the decision maker (affected by external factors).
- *Result Variables*: This is the result variable of the model.

When choosing the final decision, the decision maker may want an optimal decision or a satisfactory, partial optimal decision. Therefore, two types of decision support models can be applied: normative models and descriptive models. The normative model considers all the alternatives and chooses the optimal one, while the descriptive model considers a set of conditions at the user's discretion and considers the alternatives under these conditions and gives a satisfactory result. Since this model does not consider all alternatives, the final result is only close to optimal.

Factually, normative models are often used in single-objective optimization problems, and descriptive models are often used in multi-objective optimization problems when these goals may conflict.

DSS is often classified into communication-driven, data-driven, document-driven, knowledge-driven and module-driven [2,3] and has abilities such as supporting decision makers in unstructured and semistructured situations that cannot be solved by other calculation systems; supporting different levels of management from executors to managers; and for the diversity of decision-making processes and decision types in which there is a match between the DSS and the personality of the individual decision maker, such as vocabulary and decision-making style, improving the efficiency of the decision-making process, such as correctness, accuracy, time and quality. The decision maker controls all steps of the decision-making process in solving problems. DSS is intended to assist, not replace, decision makers. The decision maker can ignore the computer's advice at any stage in the processing. DSS often uses models for the analysis of decision-making situations. The modeling capabilities allow experimentation with

different strategies and with different configurations. High-level DSS is equipped with an intellectual component, so it allows for potential and effective solutions to difficult problems [4].

Data Classification

Data classification assigns new samples to classes with the highest accuracy to predict newly labeled datasets and samples. Classification also predicts the class type of the label. The input is defined as a set of examples (training data) and a classification label for each data sample, while the output is a predictive model or data classifier.

The data classification process is conducted in two steps:

Step 1: Build a model

This step describes a set of predefined classes in which:

- Each assigned set or pattern belongs to a predefined class as identified by the class label attribute.
- The training dataset is the set of sets used to build the model.
- Classification models are represented by mathematical formulas, decision trees or classification rules [5].

Step 2: Use the model

When new objects are put in, the system classifies these new objects.

Evaluate the effectiveness and accuracy of the model:

- The tested label of the sample, which is defined and known in advance, is compared with the classification results obtained from the model.
- The effectiveness and accuracy of the model is determined by two common methods of using training data and K-fold cross validation.

Using training data. This is the simplest method to evaluate a machine learning model. From the original dataset, we divide it into 2 groups, called training data and testing data, in a certain proportion (usually 7: 3, 8: 2 or even 9: 1 depending on the size and characteristics of datasets). Then, the model is run on the training data, and the training model is used to predict the testing data. The quality of the model is evaluated based on the predicted results.

Using K-Fold Cross-Validation. Cross-validation is an extension of the above method to avoid the problem of high variance. The steps are as follows:

- Shuffle data randomly.
- Divide the original dataset into k parts (k = 5,10...), and each part is called a fold. Train the model on the k-1 folds and evaluate on the remaining fold.
- Repeat the above steps k times so that every fold in the dataset has a chance to become a test dataset.

After completing the process, we have k different evaluation results, and the final result is aggregated based on the mean and standard deviation of those k results [5].

Decision Tree

A decision tree is a kind of structure that describes knowledge in the form of a tree to classify data objects into certain classes [6]. The decision tree structure shows the hierarchy of nodes and branches. The branch represents the possible value of the attribute (arrows), while network nodes have different attributes depending on their kind. There are three kinds of nodes, including root nodes, leaf nodes and internal nodes.

- The root node is the top vertex of the tree.
- The leaf node (leaf node) is the outermost node, carrying the classifier attribute (rectangle).
- The internal node is the remaining node, which has the categorical attribute (circle).

The description of the decision tree is simple and user friendly. If the decision tree has a reasonable number of leaves, it can be captured by nonprofessional users. Additionally, the decision tree can be converted to a rule set and is capable of handling datasets with missing values.

XGBoost algorithm

The eXtreme Gradient Boosting algorithm, abbreviated as XGBoost, is a new algorithm introduced by Tianqi Chen and Carlos Guestrin in 2016 [7]. XGBoost emerged as a powerful tool in many classification fields [8,9] and won many classification competitions organized by Kaggle [10]. XGBoost [11] is upgraded from gradient tree boosting (GTB) [12]. The basic principle used in the GTB algorithm is the combination of high-error learning models into a more robust learning model tree with a sequential rule. The training process of the GTB algorithm is illustrated in Figure 4. Assume that we have training data with N samples $X=\{X_1,X_2,\ldots. X_N\}$ with defined output parameters $y=\{y_1,y_2,\ldots.y_N\}$. In the first loop, an arbitrary tree is

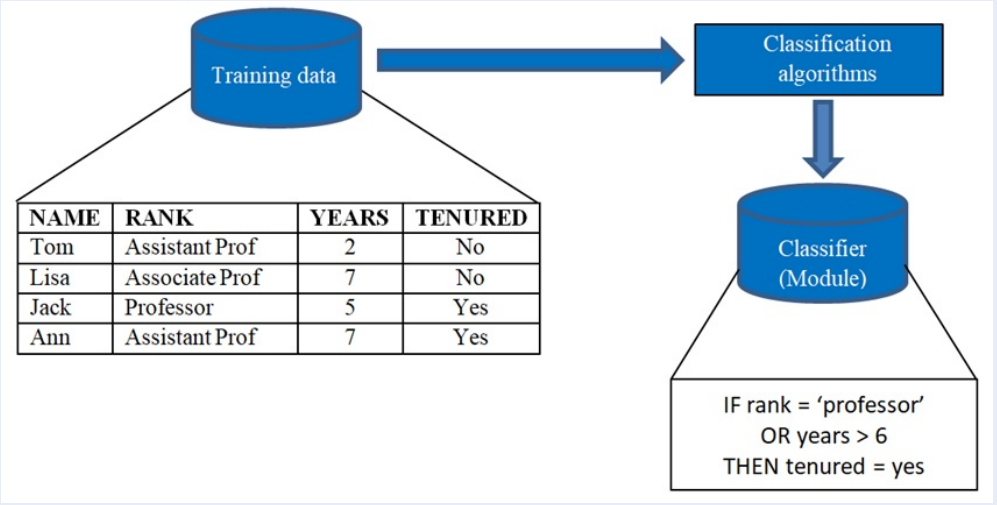

Figure 1: **Building the classification model.**

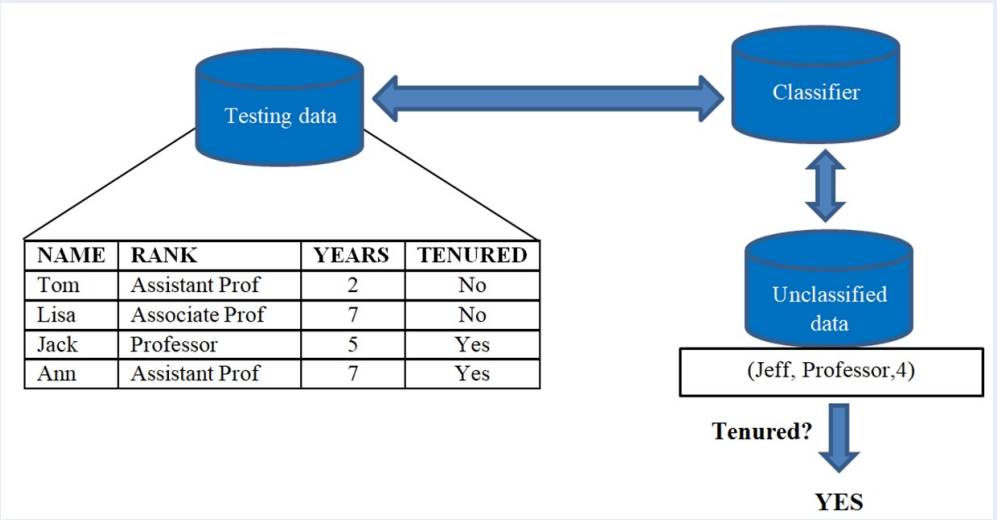

Figure 2: **Using the classification model.**

created, and the output values are estimated as $f_1(X)$. These estimated values will differ from the exact value of y by the residual $g_1(X)$. The residual can be considered the error of the model. To obtain a better training model, this residual value should be reduced. Therefore, the second tree will then be built for the purpose of estimating the values of that residual (not the value y). Similarly, when estimating residual $g_1(X)$, the second tree will estimate the value $f_2(X)$, which produces residual $g_2(X)$. To estimate the residual $g_2(X)$, a third tree is applied. The iterative process continues. The final estimated value will be $\sum f_n(X)$.

To improve the performance of the GTB model, in the XGBoost model, a component is added in the loss function. Then, the loss function of the XGBoost model is defined as in (1).

$$J(\theta) = L(\theta) + \Omega(\theta) \quad (1)$$

where the parameters of the trained model are denoted Θ; L is a loss function, and Ω is an added component called regularization to evaluate the complexity of the model. Adding the regularization component helps to improve the obtained parameters of the trained model and avoid overfitting. Using the normalized objective function as in (1) helps the model

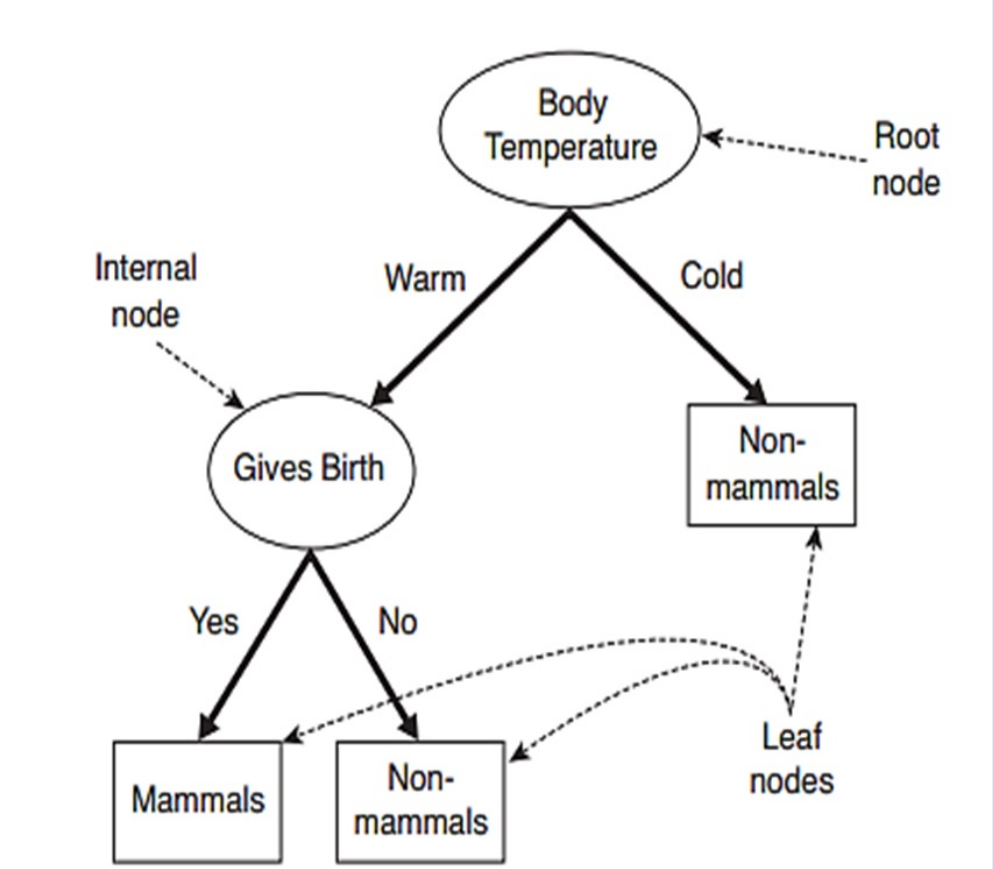

Figure 3: **An example of a decision tree.**

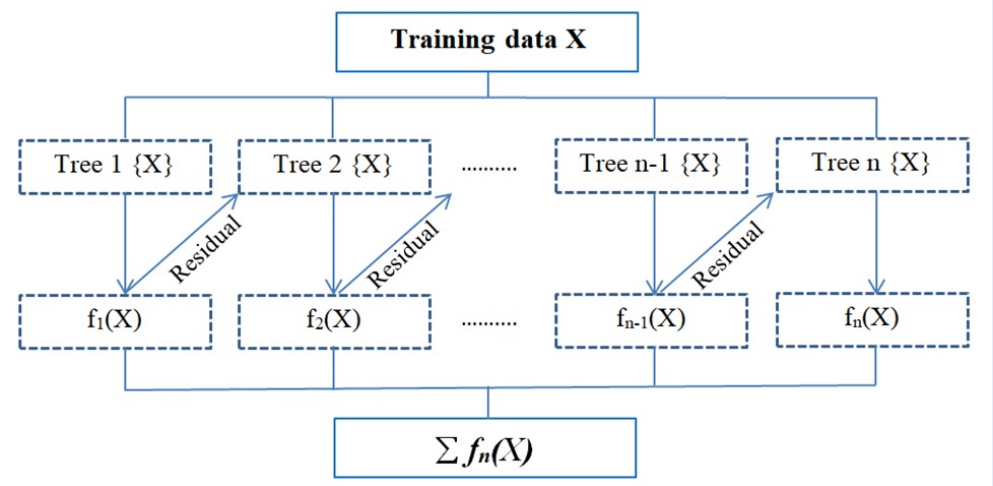

Figure 4: **Training process of the GTB algorithm.**

use simple and predictable functions. The simpler the model, the better it is to avoid overfitting. Due to the tree-based trained model, the final estimated values are defined in (2)

$$y_i^{'} = \sum_{i=1}^{n} f(n)(X_i) \quad (2)$$

The loss function at the t^{th} loop has the form as in (3)

$$J^{(t)} = \sum_{i=1}^{N} L\left(y_i, y_i^{'}\right) + \sum_{i=1}^{t} \Omega(f_n) \quad (3)$$

The estimated value for the output y_i in the t^{th} loop, $y_i^{'(t)}$ is calculated in (4)

$$y_i^{'(t)} = \sum_{n=1}^{n} f_n(X_i) = y_i^{'(t-1)} + f_t(X_i) \quad (4)$$

The regularization Ω (f_n) is formulated in (5)

$$\Omega(f_n) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

where γ is the complexity parameter and represents the complexity of the leaves in the decision tree; T is the number of leaves in the decision tree; λ is the penalty parameter; and w_j^2 is the output of leaf node j. Generally, the XGBoost algorithm is described as follows:

Input: A set of training data. These data include attributes that describe an object or situation and a categorized value.

Output: The decision tree is capable of classifying the training data, and the tree will be able to correctly classify even the unseen data.

The XGBoost algorithm has many advantages, including regularization, parallel processing, handling missing values, cross validation and effective tree pruning.

The Proposed Model

The majors counseling support system is built based on the academic results in the 3 years of high school and other factors, such as the student personality survey that fits into which of the 6 major groups according to the John Holland theory, teachers' comments recommending the majors to provide counseling results in a group of majors, suggesting that high school students self-orient and choose the right majors in line with their abilities, their own personality and interests, and helping counselors in high schools to better understand and advise students more accurately in choosing majors.

Therefore, based on available data such as academic results, exam results and university major selection of students in previous school years, the authors use data mining and analysis techniques with the XGBoost algorithm to build a decision tree with the desired results and predict the best solutions to suggest to the students.

Building the decision tree using the XGBoost algorithm. From the collected dataset, the authors build an algorithm for the career counseling problem based on the Python programming language and XGBoost model. Data are represented with 17 attributes including 12 attributes of average score of 12 subjects in 3 years at high schools; 1 attribute of survey result of student personality test questionnaire, 1 attribute of the homeroom teacher's comments and 1 attribute on the results of statistics of majors after graduating from high school. The dataset is divided into 2 groups: 70% for training and 30% for testing.

The processes of data segmentation and classification are shown in Figure 5 and Figure 6, respectively.

The main function of the model is major suggestion. This means that when the user enters information such as the results of the subjects of the 3 years (10th grade, 11^{th} grade, 12^{th} grade), doing the personal survey, providing comments of the homeroom teacher, then the system will predict the suitable majors from six groups of John Holland's majors and suggest to the students. From the prediction, the system will also connect to the websites of a number of universities that have corresponding academic curricula for students to have information and convenient choices.

DATASET

The data used to train the model consist of the average score of all subjects in 3 years at high schools, student personality test questionnaire, comments of the teacher for grade 12 students, and statistics of majors after graduating from high school.

- Average score of all subjects in 3 years at high schools of 1709 students: The authors collect the data from 2 cohorts. The extracted data table is an Excel file containing personal information and the average score of subjects for every student: Math, Physics, Chemistry, Biology, Informatics, Literature, History, Geography, Foreign Languages, Technology, National Defense - Security Education and Citizenship Education.
- Student personality test questionnaire: Student personality survey according to John Holland code theory is conducted for 12th grade students, and students will take the test on the Office form application. The test content consists of 6 questionnaires corresponding to 6 occupational groups of John Holland theory (Realistic, Investigative, Artistic, Social, Enterprise and Conventional). Each worksheet will have 9 multiple choice questions about the student's interests and personality. Students choose the answers to the questions according to the score

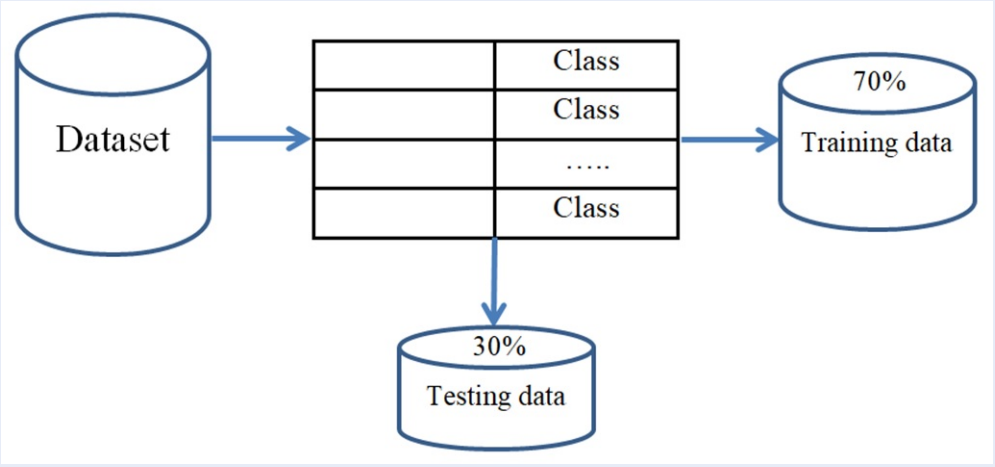

Figure 5: **The process of data segmentation**.

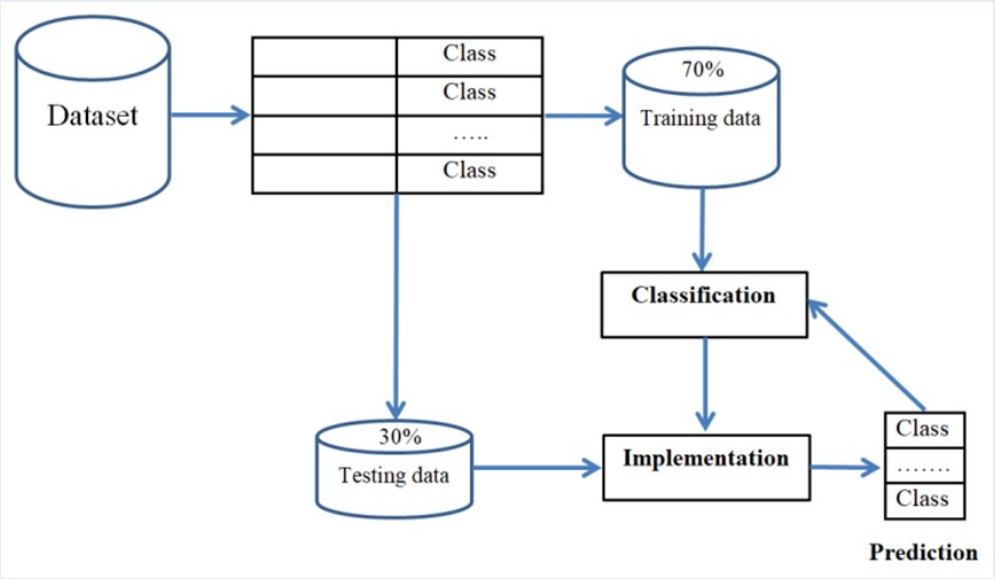

Figure 6: **The process of data classification**.

from 1 to 5. After completing 6 tables, the table with the highest total score corresponding to that majors group will be selected as the result of the survey process. The steps to conduct the student majors survey by John Holland code theory are as follows:

Step 1: Reading the instructions to take the survey
The contents listed in each table correspond to individual qualities and abilities. For every piece of content, there will be many levels. For every level, a corresponding score will be assigned. This corresponding score is assessed by the survey takers and self-scoring according to the following rules: You see that the content has never been true for you – equivalent to 0 point; You only found that content true in one case – equivalent to 1 point; You just found that the content is only true in a few cases – equivalent to 2 points; You see that the content is only half true for you - equivalent to 3 points; You find that almost true for you in most cases, only a few cases are not quite right - equivalent to 4 points; You see that the content is absolutely true for you, it cannot be otherwise - equivalent to 5.

Step 2: Adding points to each content in every table and adding the total score of that table, determine the

table with the highest score. For every piece of content, there will be many levels. For each level, a corresponding score will be assigned. Each checked box will be counted from 0 to 5 points. A high score is not the best, but the survey takers must choose the option to ensure that it is the most suitable for their thinking. Each majors group has 9 items, corresponding from 0 to 45 points.

Step 3: Finding the table with the highest score. The table with the highest total score is the one with the majors group that best suits your personality.

After the students took a survey about their personal interests and personality, to confirm the accuracy of the data, the authors carried out cross-validation by talking back to the teacher, who always monitors the students in learning as well as understanding the personality of each student in their class. Teachers' comments are compared with the results of the survey to complete the data for the problem.

- Comments of the teacher for the grade 12 students: For high school students, the homeroom teacher is the one who knows the personality and learning ability of each student in the class that she/he is assigned to the homeroom. Therefore, the homeroom teacher's advice about the majors to study after graduation is very important for students. Teachers give the comments by Office form for students based on their interests and abilities to advise that a student should choose which major out of 6 major groups of John Holland's theory.
- Statistics of majors after graduating from high school: After the end of the school year, many high schools have statistical reports on the results of the school year, as well as statistics on students' majors after graduation. This report helps the school to know the percentage of students enrolled in the majors to have an orientation for the next school year plan. The authors used the data of the statistics table of the majors extracted from the report file of the statistics of the students' majors after graduation and analyzed the number of students enrolled in the major groups according to John Holland's theory to enrich the data for the majors consulting model.

EXPERIMENTAL RESULTS

To evaluate the effectiveness of the model, the authors combined the accuracy with the confusion matrix and conducted a survey of students/parents using the system.

For the assessment of accuracy, the authors compared XGBoost with algorithms such as ID3 and random forest to prove that XGBoost is the most suitable choice in building a decision tree. The implementation results are shown in Table 1 and Figure 7.

DISCUSSION

The confusion matrix of the proposed model is shown in Table 2. The authors test the model for 523 students of testing data. From the results in Table 2, it is clear that the values in the diagonal are the correctly predicted records. Therefore, the larger the number on this diagonal, the better the model is.

In the confusion matrix above, the values 1 to 6 correspond to 6 occupational groups of John Holland theory (1= Realistic, 2 = Investigative, 3 = Artistic, 4 = Social, 5 = Enterprise and 6 = Conventional). The values in columns are actual majors, and those in rows are predicted majors from the proposed model. From Table 2, we can see that:

- At the first row: Among 77 students whose actual label is 1, there are 62 students correctly predicted with label 1, 10 students incorrectly predicted with label 2 and 05 students incorrectly predicted with label 6.
- Similarly, in rows 2, 3, 4, 5 and 6, 75, 62, 135, 48 and 71 students are correctly predicted, respectively.
- The average accuracy from the confusion matrix is 86%.

In addition, to evaluate the efficiency of the model, the authors collected a survey of 908 students and 250 parents with good feedback, as shown in Table 3. The students evaluating the system are students who have entered university in the last 03 years, mainly good students and students who previously attended high schools for the gifted. The authors selected these students because they possess good perspectives and sense of learning to provide objective evaluations of the system. The author also selected 250 parents randomly, from close friends and acquaintances or introduced through acquaintances. These people have children who have graduated from college or are in college and are quite successful. These students and parents are suggested to test the system and evaluate the level of satisfaction with the system.

CONCLUSION

In this paper, the authors build a model of major consulting support by analyzing the steps of the major

Table 1: Comparison of algorithms in building decision trees

Algorithms	ID3	Random forest	XGBoost
Accuracy (%)	58%	74%	86%

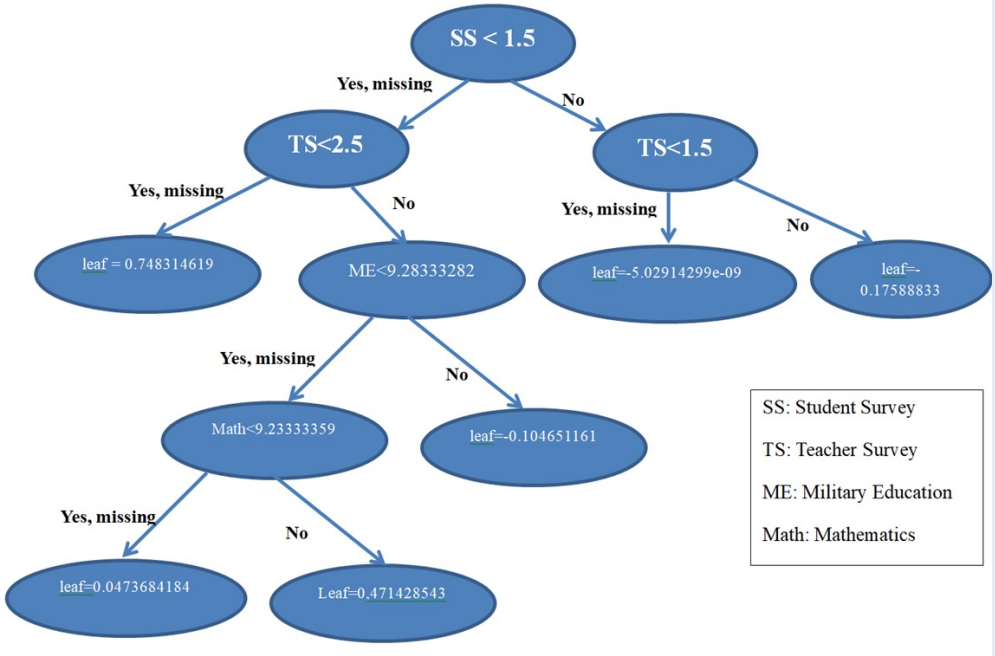

Figure 7: The implementation results of XGBoost.

Table 2: The confusion matrix of the model of major consulting support

		Actual majors					
		1	2	3	4	5	6
Predicted majors	1	62	10	0	0	0	5
	2	0	75	0	0	0	0
	3	0	5	62	3	1	5
	4	5	0	0	135	0	10
	5	5	0	0	5	48	0
	6	0	2	0	14	0	71

Table 3: The results of satisfaction levels on the model

Levels of satisfaction	Students		Parents	
	Numbers	Percent (%)	Numbers	Percent (%)
Very useful	756	83.26	215	86
Useful	135	14.87	35	14
None	17	1.87	0	0
Total	908		250	

consulting process, combined with data mining techniques from which XGBoost is applied to classify educational data. The proposed model predicts students' ability to suggest suitable majors based on academic performance in the high school years. The model saves time for students and their parents when looking for a major for further study and is evaluated to be effective with high accuracy. The authors proved that XGBoost classifies the data with an accuracy higher than that of ID3 and random forest. In addition, the efficiency and reliability of the proposed models are also shown in the accuracy, confusion matrix and user survey. In the coming research direction, the authors collect more data related to family traditions and economic conditions, labor needs, talents... to improve the model.

LIST OF ABBREVIATIONS

DSS: Decision Support System
GTB: Gradient Tree Boosting
CNN: Convolutional Neural Network

ACKNOWLEDGEMENT

We thank the AIoT group of the School of Computer Science and Engineering for giving me the motivation to complete the model with the expected effectiveness.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Kha Tu Huynh proposed the idea, supervied the implementation, evaluated the efficiency of the model built and implemented the model, and analyzed the results. Nga Tu Ly wrote the paper and reviewed and revised the paper. All authors read and approved the final manuscript.

REFERENCES

1. McCosh AM, Morton MSS. Management decision support systems. Springer; 1978;Available from: https://doi.org/10.1007/978-1-349-02764-4.
2. Burstein F, Holsapple W, C, Power DJ. Decision support systems: a historical overview. Handbook on decision support systems 1: basic themes; 2008. p. 121-40;Available from: https://doi.org/10.1007/978-3-540-48713-5_7.
3. Le Van Duc. Decision support system. Ho Chi Minh City: Vietnam National University; 2006;.
4. Ha TTT. Textbook of decision support information systems, Institute of Information Technology and Digital Economy. National Economics University; 2020;.
5. P Đỗ. Textbook of data mining. 2012;.
6. Kiem H, Phuc D. Textbook of data mining. Center for Research and Ho Chi Minh City: Development of Information Technology, Vietnam National University; 2005;.
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016, August. p. 785-94;Available from: https://doi.org/10.1145/2939672.2939785.
8. Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. Inf Sci. 2021;572:522-42;Available from: https://doi.org/10.1016/j.ins.2021.05.055.
9. Kadiyala A, Kumar A. Applications of python to evaluate the performance of decision tree-based boosting algorithms. Environ Prog Sustainable Energy. 2018;37(2):618-23;Available from: https://doi.org/10.1002/ep.12888.
10. Abou Omar KB. XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. Preprint semester project; 2018;.
11. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al. Xgboost: extreme gradient boosting. R package version 0.4-2. 2015;1(4):1-4;.
12. Nielsen D. Tree boosting with XGBoost. Norwegian University of Science and Technology; 2016;.