Pertanika J. Sci. & Technol. 31 (1): 475 - 493 (2023)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Predicting Students' Inclination to TVET Enrolment Using Various Classifiers

Chia Ming Hong[1]*, Chee Keong Ch'ng[1] and Teh Raihana Nazirah Roslan[2]

[1]Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM, Sintok, Kedah, Malaysia
[2]Othman Yeop Abdullah Graduate School of Business, Universiti Utara Malaysia, 50300 UUM, Kuala Lumpur, Malaysia

ABSTRACT

Technical and Vocational Education and Training (TVET) is an education system that delivers necessary information, skills, and attitudes related to work or self-employment. However, the TVET program is not preferred by most Malaysian students due to several factors such as students' interest, parental influence, employers' negative impression, facility in vocational institutions, inexperienced TVET instructors, and society's negative perception. Consequently, it raises the issue of skilled workers shortage. The gravest threat will be far-reaching, pushing our economy into depreciation. Therefore, it is important to identify the students' traits and interests before conducting further investigation to turn and thrive in this phenomenon. This study aims to utilise several classifiers (Decision Tree, Neural Network, Logistic Regression and Naïve Bayes) to predict students' inclination to join TVET programmes. A total of 428 secondary school students from Kedah, Malaysia, are chosen as our survey respondents. The best classifier is determined according to the lowest misclassification rate. The findings revealed that the Decision Tree-based Gini Index with three branches prevail against other classifiers with a misclassification rate of 0.1938. Therefore, the classifier could act as a steer for the Kedah Department of Education (DOE), related parties, and the TVET agency in implementing effective strategies to enliven and inspire students to join TVET programs.

Keywords: Decision tree, logistic regression, Naïve Bayes, neural network, technical and vocational education and training

ARTICLE INFO

Article history:
Received: 08 March 2022
Accepted: 27 April 2022
Published: 09 November 2022

DOI: https://doi.org/10.47836/pjst.31.1.28

E-mail addresses:
carmenhongcm@gmail.com (Chia Ming Hong)
chee@uum.edu.my (Chee Keong Ch'ng)
raihana@uum.edu.my (Teh Raihana Nazirah Roslan)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The Fourth Industrial Revolution 4.0 (IR 4.0) is rapidly transforming and changing the skills demanded by the industry. The readiness of Technical and Vocational Education and Training (TVET) institutions to produce a high-quality workforce and high-value competing manufacturers is vital for developing human capital to make Malaysia a high-income and developed nation (Sulaiman & Salleh, 2016; Yaakob, 2017). Therefore, the government should enforce the TVET program by coordinating different resources from various agencies. The education system must provide the opportunity to learn from basic to advanced levels across various institutional and work settings and focus on students' understanding of industrial processes and enormous technical skills. Previously, several efforts have been implemented by the Malaysian government to attract more students to enrol in TVET. For instance, the government had allocated RM30 million under Malaysia Budget 2019, RM5.9 billion under Malaysia Budget 2020 and RM 6 billion under Malaysia Budget 2021 to boost the quality of TVET (Sivanandam et al., 2019; Rajaendram, 2020). Currently, there are 36 polytechnics, 98 community colleges, 80 vocational colleges, 676 public-accredited vocational centres, and 642 private accredited vocational centres in Malaysia. It indicated that the government aims to produce a top-notch workforce towards becoming a fully industrialised, developed, and high-income nation.

Although the number of TVET institutions has been increasing, most school leavers are unaware of the advantages of TVET (Sabang, 2017; Azizi, 2018). One of the benefits of being a TVET graduate is that the average employability rate of TVET graduates is as high as 96% in the industry (Ismail et al., 2021). It shows that most TVET graduates can secure a job after they graduate. Other than that, TVET can equip students with specific skills to face the working environment so they will not have any difficulties in their work (Karim, 2018). However, despite our government's recognition of TVET certification, students still resist vocational education, resulting in low enrolment, which is far from the expected number (KRI, 2018; Aziz, 2019). According to the Malaysia Human Resource Minister, the government expects at least 35% of skilled workers in the industry, but currently, there are only 28% skilled workers (Aziz, 2019). As a result, the number of skilled workers is still far from reaching the market demands. Several adverse factors to the TVET program are identified, such as students' interest, parental influence, employers' negative impression, facility in vocational institutions, inexperienced TVET instructors, society's negative perception, government policy, and expensive education costs (Hong et al., 2021).

In investigating the palpable phenomenon, the mentioned factors can assist in developing models which aim to study the students' inclination using data mining approaches. It uses statistical study techniques to extract the data pattern and transform the information into an easily understandable structure for future usage (Sahu et al., 2011). There has been a surge of interest in using data mining techniques to tackle educational research issues, specifically Educational Data Mining (EDM). EDM has emerged rapidly in becoming an important field to reveal hidden yet meaningful data patterns in educational institutions. It is widely applied in many areas, such as the prediction of students' enrolment, the performance of students and teachers in school, and students' dropout rates.

Several kinds of research related to the prediction of students' enrolment in EDM have been performed using data mining techniques such as Decision Tree, Logistic Regression, Neural Network and Naïve Bayes (Khan & Choi, 2014; Raju & Schumacker, 2015; Hung et al., 2020; Messele & Addisu, 2020). However, there is a lack of a predictive model discussing students' enrolment in vocational education institutions. Most previous studies focus on the student's performance using data mining approaches. Specifically, most researchers only apply one data mining approach in their data analysis (Babu et al., 2020; Messele & Addisu, 2020; Herlambang et al., 2019). Therefore, the main objective of this study is to utilize and compare several data mining approaches, namely Decision Tree, Neural Network, Logistic Regression and Naïve Bayes, to investigate the students' inclination to TVET enrolment. This study is unique in addressing the current TVET challenges in Malaysia. So far, no specific model has been developed to help the government identify students' tendency to join TVET even though huge grants and funds have been allocated (as stated in MOHE's Malaysia Education Blueprint (Higher Education) and Malaysia Budget).

LITERATURE REVIEW

Sustainable Development Goals (SDGs)

Sustainable Development Goals (SDGs) are important to ensure that all people worldwide can enjoy endless peace and prosperity. Some goals are closely related to the TVET field. For example, Goal 4 is for quality education, and Goal 5 is for gender equality. SDGs ensure that all students receive a high standard of education and encourage them to continue studying throughout their lives (Mustapha, 2015). By 2030, all people will have affordable, high-quality technical, vocational and tertiary education. Other than the quality of education, SDGs also emphasise long-term economic growth, increased productivity, and encourage the development of new technologies (Berawi, 2019; Imaz & Sheinbaum, 2017). Consequently, there will be demand for TVET workers who act as the experts in creating technological products. In addition, sustainable industrialization is one of the goals to be achieved in SDGs. To fulfil this, TVET workers play an important role because they are trained to handle, manage and upgrade new technologies. However, the resistance issues of young generations to TVET remain unchanged.

TVET Issues

In Malaysia, most students are still reluctant to enrol in TVET programs due to some factors. Students' interests, vocational talent, demographic background, and personality are the main components in determining the students' tendency in enrolling TVET (Ismail & Hassan, 2013). However, negative social perceptions toward TVET, such as labelling TVET students as low achievers in school, cause intense impediments (Abdul-Aziz et al., 2020; Ismail & Hassan, 2013; Amedorme & Fiagbe, 2013). Consequently, this leads to poor perception from parents' perspectives (Hussin et al., 2017; Koya, 2019). Likewise, the employer is one factor that discourages students from choosing TVET because of the TVET graduates' qualification issues (Cheong & Lee, 2016; Chan, 2018).

Furthermore, TVET instructors who are inexperienced, and lacking in information, communication, and technology (ICT) and English skills can affect students' inclination to TVET (Ismail et al., 2017; Ismail & Hassan, 2013). According to Amin (2016), currently, there are two accreditation bodies (Malaysian Qualification Agency and Department for Skill Development), which confuses the graduates in choosing the right accreditation channel. Certain employers only accept accreditation from the Malaysian Qualification Agency. As a result, students may be perplexed when choosing a course with many certification agencies. Education cost is also one of the barriers for students to pursue their education in TVET institutions because technical education is costly, and the allocation from the government is not enough to cover the cost (TheStart, 2019). Lastly, poor facilities in technical institutions can sway away students' interest too—for instance, poor equipment, unfunctional air conditioners, and cramped classrooms (Bakar, 2011).

Introduction to Data Mining Approaches

This study uses descriptions, and related equations for several popular data mining approaches: Decision Tree, Neural Network, Logistic Regression, and Naïve Bayes.

Decision Tree. Researchers widely apply a Decision Tree for prediction and classification due to its simplicity and transparency (Salal et al., 2019; Yamini & Ramakrishna, 2015). It is a tree-like structure that consists of a root node, internal node, and leaf node. The root node is also the parent node, which does not have incoming edges. Branches connect the root node with the internal node, which has outgoing and incoming edges. Nodes with no outgoing edges are known as leaf nodes (Rokach & Maimon, 2015).

Several algorithms are involved in inducing a Decision Tree, such as ID3, C4.5 CART, and CHAID algorithms. The differences between those algorithms depend on the splitting criteria. Entropy and Gini index are the most commonly used splitting criteria (Breiman, 1996). Entropy is the way to measure the impurity in the sample. It ranges from 0 to 1. The Equation 1 of Entropy is displayed below:

$$\text{Entropy, } H(x) = -\sum p(x) \log_2 p(x) \qquad (1)$$

where x = random variable; p(x) = the possibility of result x in variable x.

If the entropy value equals 0, it is defined as the sample is completely homogeneous. Next, Gini Index calculates the likelihood of a randomly chosen feature being wrongly categorised. Equation 2 of the Gini Index is displayed below:

$$\text{Gini Index} = 1 - \sum [p(x)]^2 \qquad (2)$$

where p (x) = the possibility of result x of varibale x.

The Gini Index ranges from 0 to 1, with 0 as classification purity and 1 as the random distribution of elements among different classes.

Artificial Neural Network (ANN). An Artificial Neural Network (ANN) is a copy of the human brain. It consists of processing units known as neurons (Kelvin, 1997). It is a tool to discover data patterns and tackle recurring problems in certain areas. A bias value will be added along with the input, x_i (Zakaria et al., 2014). Weight, w_i is the strength of the processing ability of input. The product of weight and input is the strength of the signal. To activate the neurons, one commonly used activation function is the sigmoid function (Kukreja et al., 2016). It is important to understand how neural network solves complex problems. Equation 3 is shown below:

$$f(x) = \frac{1}{1+e^{-sum}}, \qquad (3)$$

where $\text{sum} = \sum_{i=0}^{n} x_i w_i$.

The value of the sigmoid function is ranged between 0 and 1. The structure of ANN includes the input layer, hidden layer, and output layer. The input layer begins the flow, which receives the input data to be processed in the ANN system. The hidden layer is between the input layer and an output layer which can be zero or more than one. It is used for the transformation of input to enter the network. Finally, the output layer shows the output classification that maps the input pattern (Islam et al., 2019).

Logistic Regression. Logistic regression is another approach used to model the probability of a binary or dichotomous dependent variable. For example, if the predicted value probability is less than 0.5 will be assigned to group 0; group 1 otherwise. The advantages of logistic regression are that it is relatively fast compared to other supervised classification techniques and allows the evaluation of multiple explanatory variables by extension of the basic principles. The specified form of the logistic regression model can be written as Equation 4:

$$p = \frac{1}{1+e^{-(b_o+b_1X_1+b_2X_{2+\ldots+}b_pX_p)}}, \tag{4}$$

where p = target variable; X_i = independent variable; b_i = parameter of the model.

The value will always be positive, ranging from 0 to 1.

Naïve Bayes. Naïve Bayes is a simple probability classifier applying Bayes' theorem with strong independent assumptions. Bayes' theorem was developed by a mathematician named Thomas Bayes in the 1740s. It is suitable for classification, although the output is more than two classes. For the categorical input, frequencies are used, while for the continuous input, the Gaussian density function or probability density function is used to calculate the probability (Berrar, 2018; Gandhi, 2018). For example, Equation 5 of Naïve Bayes is displayed below:

$$P(a|b) = \frac{p(b|a)\,p(a)}{p(b)}, \tag{5}$$

where $P(a|b)$ = objective; $p(a|b)$ = probability of occurance of b given a; p(a) = probablity of an in the sets of data; p(b) = probability of occurence of b.

Naïve Bayes allows you to calculate the posterior probability, $P(a|b)$ from p(a), p(b) and $p(a|b)$. As mentioned, the Gaussian density function deals with the continuous input (Gandhi, 2018). Equation 6 is shown below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \tag{6}$$

where σ = standard deviation; μ = mean.

Besides, there are three types of Naïve Bayes algorithms, namely, Multinomial Naïve Bayes, Bernoulli Naïve Bayes and Gaussian Naïve Bayes. Multinomial Naïve Bayes is used for the discrete data. Text classification is one of the problems in this algorithm. For example, counting how often the word occurs in the paper. For Bernoulli's Naïve Bayes, it is used when the vectors are binary. For example, it is "0 or 1" or "yes or no." The table of parameters for each method is displayed in Table 1.

Table 1
Parameters for each method

No	Method	Parameters
1	Decision Tree	Tree depth, number of features
2	Artificial Neural Network	Number of hidden layers, number of hidden nodes
3	Logistic Regression	Maximum likelihood estimation
4	Naïve Bayes	Prior probabilities of the class, portion of the largest variance of the features

Application of Data Mining in TVET

Data mining approaches are widely applied in the vocational education field. For example, Matei et al. (2018) used the Decision Tree model to examine the trend of Romanians to choose general or vocational education. To study the individuals' choices, factors such as age, education level, gender, and income range, were collected from the respondents from all regions in Romania through a survey form. The variables were split using the Gini Index, and the age variable was the most important factor. In addition, Babu et al. (2020) compared 16 Decision Tree models in predicting students' performance in their examinations. A Decision Tree with Gini Index was chosen as the best tree. Therefore, the model could be used as a reference for teachers to improve their quality of teaching work.

Next, ANN is useful in classifying and recognising data patterns. Previously, ANN was used to predict the academic achievement of vocational school students in their science subjects such as Physics, Chemistry, and Biology. The independent variables were the variables that influenced students' academic achievement, whereas the dependent variable was set as the mean score of students in those three subjects. Each subject had a different factor that contributed the most to affecting students' achievement. For example, for Physics and Chemistry subjects, the factor of "class size" had the highest significance, while for the Biology subject, the factor of "enthusiasm for Biology" had the highest significance (Yağcı & Ḉevik, 2017). Neural Network was also used to predict the student's acceptance of the vocational school's learning management system (LMS). A few factors act as important predictors, such as performance expectation, effort expectation, societal influence, and facilitating conditions. Of those factors, the most important factor was "performance expectation", which was defined as the trust in improving performance with the help of technology (Ozkan et al., 2020).

Next, Logistic Regression was widely used to solve education issues. For example, Karim and Maat (2019) applied Logistic Regression to develop a Student Employability Skills (SES) model for TVET students in Malaysia. Multinomial Logistic Regression was used in solving the problem because the dependent variables were more than two. The findings revealed that education level, mother's occupation, and job status were the essential variables that led to high SES. Therefore, the model was useful to perform the prediction for the students' employment after they graduated from school. Other than that, the researchers used Logistic Regression to determine if the "family background" variable would influence the students' intentions to enrol in TVET. Fathers' education, mothers' education, and family income were included as the variables in this study. According to the findings, students' enrolment into TVET was influenced by two significant factors: their fathers' education and family income (Assunção et al., 2019).

Lastly, Naïve Bayes was widely applied in real-world problems such as spamming of email, face recognition software, and text classification. Previously, the researchers applied Naïve Bayes, which were based on 10-fold-cross validation and percentage split to discover TVET students' performance. It was found that factors such as age, sex, sector, type of employment, training experience, level, the purpose of assessment, and category of the candidate were found to affect students' academic performance (Messele & Addisu, 2020). Naïve Bayes was also implemented to predict the vocational students' learning achievement. For the input, there were many variables such as age, gender, parent's education and occupation, the reason for choosing a school, duration to reach school, study duration, previous class failure, and health. In this paper, the researchers classified the students' achievement into five categories "very good," "good," "fair," "poor", and "failed." The created model had a moderate accuracy which could be a guide for reference in the future (Herlambang et al., 2019).

Despite many data mining approaches widely discussed in educational fields, little attention has been paid to incorporating various classifiers in past research, especially in the TVET issue. Therefore, this study aims to develop several classification models to predict the students' inclination to TVET enrolment in future. The models are compared based on their performances. The model that reaches the highest accuracy will be selected as our predictive model.

MATERIALS AND METHOD

Questionnaire Design

The questionnaire structure is developed based on the previous study by Trudis (2014). Some modifications have been made to that questionnaire to match our suitability. After that, the Cronbach-alpha test is used to measure the reliability of the questionnaire. From the test, the alpha value is 0.612, which means that this questionnaire can be accepted (Ursachi et al., 2015).

Data Collection Process

In the data collection process, 428 students were selected from secondary schools in Kedah, Malaysia. The reason for choosing secondary school students is because this group is eligible for tertiary education soon. The widespread recognition that tertiary education is a major driver of economic competitiveness in an increasingly knowledge-driven global economy. Therefore, the post-secondary school students' opinions are vital, particularly in the TVET course, as it promotes competency-based education and training linked to industry needs. The techniques used to select those respondents are disproportionate stratified sampling with equal allocation and simple random sampling.

Disproportionate stratified sampling is carried out according to the 9 District Education Office (PPD) and 36 State Legislative Assembly (DUN) to justify the number of samples selected in this study. Subsequently, simple random sampling is used to choose the location in each DUN. For example, three DUN are under PPD Baling: Bayu, Kuala Ketil and Kupang. Therefore, one of them is randomly selected as our targeted location.

After deciding on all the schools, an appointment will be made with the representative. Questionnaires are distributed to students after the briefing. The flowchart of the data collection process is shown in Figure 1.

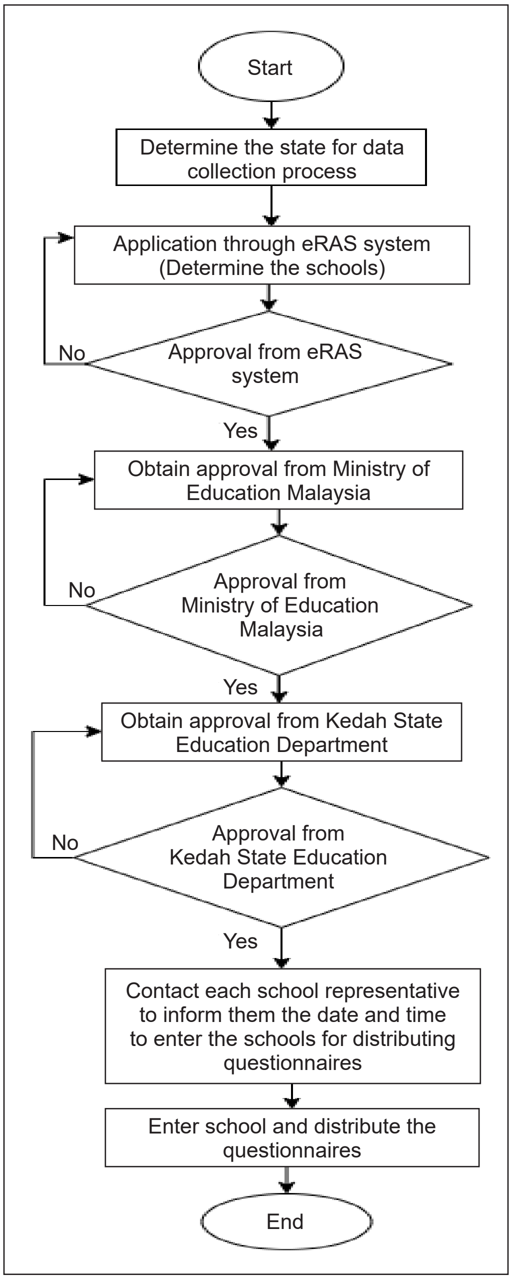
Start
Determine the state for data collection process
Application through eRAS system (Determine the schools)
No
Approval from eRAS system
Yes
Obtain approval from Ministry of Education Malaysia
No
Approval from Ministry of Education Malaysia
Yes
Obtain approval from Kedah State Education Department
No
Approval from Kedah State Education Department
Yes
Contact each school representative to inform them the date and time to enter the schools for distributing questionnaires
Enter school and distribute the questionnaires
End

Figure 1. Data collection flowchart

Process of Development of Data Mining Approaches

The process flow of this study is presented in Figure 2, and the explanations of each part are discussed.

Input Data. For the input data, the variables in the questionnaire are used to conduct the data analysis. There are 15 input variables and 1 output variable. For the ordinal type variables, the answers use the 5 Likert Scales, where 1 represents strongly disagree, 2 represents disagree, 3 represents not sure, 4 represents agree, and 5 represents strongly agree. The input values applied are based on the Likert scale values. Table 2 shows the role and description of each variable.

Data Pre-Processing. In the pre-processing data phase, the data undergoes a few steps, such as data cleaning, data integration, and data reduction, to ensure the data is clean before performing analysis. The most frequent values replace missing values in the dataset.

Data Partitioning. Data is split according to the 70:30 rule (Nguyen et al., 2021; Liu & Cocea, 2017). The training set (299 instances) is used to develop the model, whereas the test set (129 instances) is used to evaluate its performance.

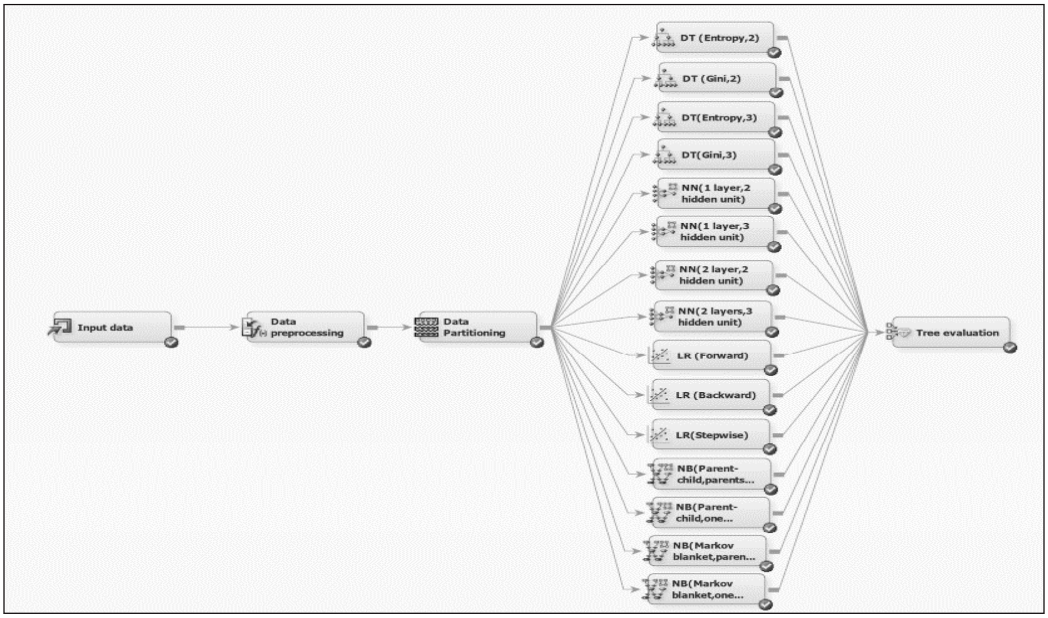

Figure 2. Process flow diagram

Table 2
Role and description for each variable

Variables	Role	Type	Description
AcademicSubjMoreImportant	Input	Ordinal	Students' agreement towards the academic subject is more important than the vocational subject.
FamilyInvolved	Input	Nominal	Students having family members who had joined TVET
FatherEducation	Input	Nominal	The highest education of the student's father.
FatherJob	Input	Nominal	The occupation of the student's father
Gender	Input	Nominal	Gender of student
HeardTVET	Input	Nominal	The students had heard the information about TVET before.
InterestedTVET	Output	Nominal	The student feels interested in joining TVET after Form 5.
MotherEducation	Input	Nominal	The highest education of the student's mother.
MotherJob	Input	Nominal	The occupation of the student's mother.
PlanAfterForm5	Input	Nominal	The plan of the student after completion of secondary school.
Stream	Input	Nominal	The stream of students in their secondary school.
TVETBright	Input	Ordinal	With the level of agreement towards students who get a TVET certification, people will think they have a bright future.
TVETStudHigherJobChance	Input	Ordinal	The level of agreement toward TVET students will have a higher chance of getting a job compared to those who do not.

Table 2 *(continue)*

Variables	Role	Type	Description
TVETStudHighSalary	Input	Ordinal	The level of agreement toward TVET students has a high salary.
VocationalSubjStatus	Input	Ordinal	The level of agreement towards the vocational subject has a high status.
VocationalCourseInteresting	Input	Ordinal	The level of agreement towards vocational courses is interesting.

Models Construction. Four types of Data Mining models were developed in this study: Decision Tree, Neural Network, Logistic Regression, and Naïve Bayes. Decision Tree mainly focuses on splitting criteria and the number of branches; Artificial Neural Network relies on the number of nodes and layers; Logistic Regression specifies different entry methods for different subsets of variables; Naïve Bayes uses conditional probabilities. Table 3 shows the type of models constructed in this study.

Table 3
Types of models constructed

No	Model	Criteria
1	Decision Tree	Entropy, 2 branches
		Gini, 2 branches
		Entropy, 3 branches
		Gini, 3 branches
2	Artificial Neural Network	1 layer, 2 hidden units
		1 layer, 3 hidden units
		2 layers, 2 hidden units
		2 layers, 3 hidden units
3	Logistic Regression	Forward
		Backwards
		Stepwise
4	Naïve Bayes	Parent-child structure, parents set
		Parent-child structure, 1 parent
		Markov-blanket structure, parents set
		Markov-blanket structure, 1 parent

Tree Evaluation. Several aspects need to be considered to measure the accuracy of the model. These aspects are in the confusion matrix, as shown in Table 4.

Table 4
Confusion matrix

Confusion matrix		Predicted	
		No	Yes
Actual	No	True Negative (TN)	False Positive (FP)
	Yes	False Negative (FN)	True Positive (TP)

True Negative (TN) is the number of correct predictions where the actual class is "no", and True Positive (TP) is the number of correct predictions where the actual class is "yes." False Negative (FN) is the number of predictions where the actual class is "yes," but it is predicted as "no." The last one is False Positive (FP), which means the actual class is "no" but wrongly predicted as "yes." The four parameters are used to calculate accuracy, which aims to evaluate the models' performance. For example, Equation 7 of the misclassification rate is given as below:

$$\text{Misclassification rate} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}}. \tag{7}$$

The model with the lowest misclassification rate is considered the best model.

RESULTS AND DISCUSSIONS

Descriptive Analysis of Input Variables

The descriptive analysis of input variables is identified and discussed in Table 5.

Table 5
Descriptive analysis of input variables

No	Variables	Outputs
1	AcademicSubjMoreImportant	Strongly disagree: 6; Disagree: 59; Not sure: 134; Agree: 151; Strongly agree: 78
2	FamilyInvolved	Yes: 64; No: 364
3	FatherEducation	Degree: 26; Diploma: 38; Foundation/ A-level/ Matriculation: 7; No Complete Secondary School: 51; Postgraduate: 13; SPM: 271; STPM: 22
4	FatherJob	Agriculture: 37; Clerical/ Civil Service: 128; Construction: 10; Manufacturing: 6; Others: 49; Professional: 3; Retail: 6; Self-employed: 167; Unemployed: 22
5	Gender	Female: 246; Male: 182
6	HeardTVET	Yes: 117; No: 311
7	MotherEducation	Degree: 26; Diploma: 37; Foundation/ A-level/ Matriculation: 11; No Complete Secondary School: 54; Postgraduate: 19; SPM: 257; STPM: 24
8	MotherJob	Agriculture: 4; Clerical/ Civil Service: 102; Manufacturing: 4; Others: 30; Professional: 4; Retail: 6; Self-employed: 120; Unemployed: 158
9	PlanAfterForm5	Form 6: 30; Matriculation: 26; Polytechnic: 91; Skill Training Centre: 74; University/ Private College: 163; Work: 44
10	Stream	Account: 65; Arts: 114; Business: 70; Catering: 5; Economy: 2; ICT: 24; Landscape: 3; Literature: 71; Science: 46; Technical: 17; Vocational: 11
11	TVETBright	Strongly disagree: 4; Disagree: 17; Not sure :142; Agree: 207; Strongly agree: 58
12	TVETStudHigherJobChance	Strongly disagree: 4; Disagree: 27; Not sure :158; Agree: 158; Strongly agree: 81

Table 5 *(continue)*

No	Variables	Outputs
13	TVETStudHighSalary	Strongly disagree: 2; Disagree: 13; Not sure :187; Agree: 162; Strongly agree: 64
14	VocationalSubjStatus	Strongly disagree: 5; Disagree: 17; Not sure :136; Agree: 220; Strongly agree: 50
15	VocationalCourseInteresting	Strongly disagree: 2; Disagree: 9; Not sure :55; Agree: 264; Strongly agree: 98

Evaluation and Description of the Model

Table 6 shows the evaluation of the models from training data (70%) and testing data (30%).

Table 6
Evaluation of the models

No	Models	Training data (70%)		Testing data (30%)		Number of depths	Number of features used
		Number of misclassified instances	Misclassification rate	Number of misclassified instances	Misclassification rate		
1	Decision Tree (Entropy,2 branches)	56	0.1873	31	0.2403	6	9
2	Decision Tree (Gini,2 branches)	52	0.1739	27	0.2093	6	12
3	Decision Tree (Entropy, 3 branches)	38	0.1271*	31	0.2403	6	11
4	Decision Tree (Gini, 3 branches)	43	0.1438	25	0.1938*	6	15
5	Neural Network (1 hidden layer, 2 hidden units)	53	0.1773	32	0.2481	Not applicable	
6	Neural Network (1 hidden layer, 3 hidden units)	50	0.1672	32	0.2481	Not applicable	
7	Neural Network (2 hidden layers, 2 hidden units)	63	0.2107	38	0.2946	Not applicable	
8	Neural Network (2 hidden layers, 3 hidden units)	53	0.1773	33	0.2558	Not applicable	
9	Logistic Regression (Forward)	58	0.1940	27	0.2093	Not applicable	
10	Logistic Regression (Backward)	58	0.1940	27	0.2093	Not applicable	
11	Logistic Regression (Stepwise)	58	0.1940	27	0.2093	Not applicable	
12	Naïve Bayes (Parent-child,Set of parents)	63	0.2107	33	0.2558	Not applicable	
13	Naïve Bayes (Parent-child,One parent)	71	0.2375	39	0.3023	Not applicable	

Table 6 *(continue)*

No	Models	Training data (70%)		Testing data (30%)		Number of depths	Number of features used
		Number of misclassified instances	Misclassification rate	Number of misclassified instances	Misclassification rate		
14	Naïve Bayes (Markov-Blanket,Set of parents)	62	0.2074	26	0.2016	Not applicable	
15	Naïve Bayes (Markov-Blanket,One parent)	95	0.3177	39	0.3023	Not applicable	

* Means lowest misclassification rate

The output shows that Decision Tree-based Entropy with three splitting branches obtained the lowest misclassification rate in training data which is 0.1271. However, the high misclassification rate (0.2403) reflects that this model is not fit enough as the best predictive model. On the contrary, the Decision Tree-based Gini Index with three branches outperformed all models by possessing the least misclassification rate (0.1938). Also, medium size but blanket all features earned a magnificent triumph over other models. The importance values of each feature and the tree diagram are displayed in Table 7 and Figure 3.

Table 7
Importance value for the variable

Variable	Importance value
VocationalCourseInteresting	1.0000
PlanAfterForm5	0.6335
AcademicSubjMoreImportant	0.4042
Stream	0.3810
MotherEducation	0.3254
FatherJob	0.2343
Gender	0.2255
TVETStudHighSalary	0.1964
FatherEducation	0.1896
MotherJob	0.1752
VocationalSubjStatus	0.1548
TVETBright	0.1413

Table 7 shows that “VocationalCourseInteresting” has the highest importance value which is 1.0000, followed by “PlanAfterForm5” (0.6335), “AcademicSubjMoreImportant” (0.4042), “Stream” (0.3810), “MotherEducation” (0.3254), “FatherJob” (0.2343), “Gender” (0.2255), “TVETStudHighSalary” (0.1964), “FatherEducation” (0.1896), “MotherJob” (0.1752), “VocationalSubjStatus” (0.1548), and “TVETBright” (0.1413).

The examples of rules for Node 13 and Node 18 are discussed in Table 8. The findings show that the variable VocationalCourseInteresting plays the most critical role in students’ desire to enrol in TVET. Students who feel a vocational course is interesting are more likely to enrol in TVET following high school. As a result, the government should work with school teachers and TVET instructors to develop practical and relevant techniques for making TVET more engaging. It can be accomplished by bringing more technical teaching

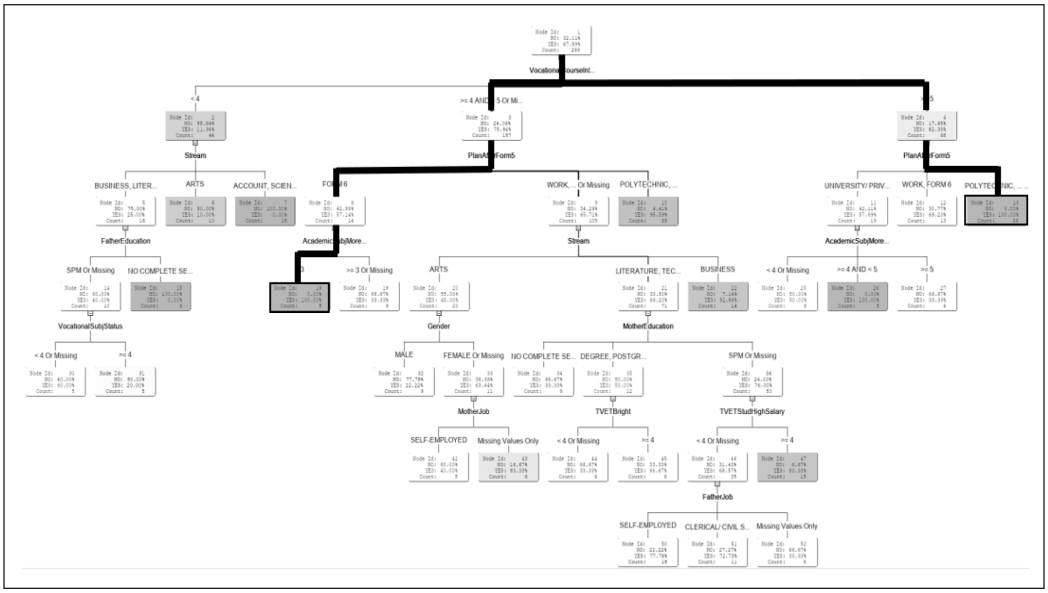

Figure 3. Decision Tree-based Gini Index with three branches

Table 8
Rules of nodes

Node	Rules
Node = 13	if VocationalCourseInteresting = 5 and PlanAfterForm5 is one of Polytechnic, Skill Training Centre or missing then Predicted: InterestedTVET=Yes = 100%
Node = 18	if VocationalCourseInteresting = 4 or missing and PlanAfterForm5 is one of: Form 6 and AcademicSubjMoreImportant =1, 2 then Predicted: InterestedTVET=Yes = 100%

tools into vocational schools, such as interactive whiteboards, 3D printing devices, mobile devices, and so on. Furthermore, the government should provide training or short courses for TVET instructors to prepare them adequately for teaching. Aside from that, students' plans after Form 5 are the second key factor influencing their desire to enrol in TVET. As a result, it is critical to understand the children's educational paths following secondary school. As a result, teachers and parents play an important role in helping students grasp their post-secondary goals.

CONCLUSION

The stigma associated with Technical and Vocational Education and Training (TVET) has led to a low number of students applying to pursue their studies in this field in Kedah Malaysia. The backgrounds, characteristics, and misperceptions of students against TVET make it dull and still unpopular even though our government has put much effort such as huge appropriation, institutions, training, campaign and facilities to engage more students to join the vocational program. Therefore, this study implemented several predictive models to discover the significant factors and predict the tendency of secondary students to join the TVET program. The Decision Tree-based Gini Index with three splitting branches has prevailed against 14 prominent models. Structural transparency, flexibility, and robustness led to enormous popularity among practitioners or researchers, especially those less proficient in data mining approaches. In conjunction with the model, Kedah DOE, related parties, and the TVET agency could have uncovered the core problem whilst implementing proper strategies to reshape the perception and policy of TVET education. Promoting TVET should not be a privilege for some to profiteer but a birth right for all Malaysians to be equitably competitive, regardless of gender, race, religion, and creed, in an increasingly challenging world economy.

ACKNOWLEDGEMENTS

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2020/SSI0/UUM/02/18).

REFERENCES

Abdul-Aziz, S. N., Zulkifli, N., Nashir, I. M., & Karim, N. A. H. (2020). Pull and push factors of students' enrolment in the TVET programme at community colleges in Malaysia. *Journal of Technical Education and Training*, *12*(1), 68-75. https://doi.org/10.30880/jtet.2020.12.01.007

Amedorme, S., & Fiagbe, Y. (2013). Challenges facing technical and vocational education in Ghana. *International Journal of Scientific & Technology Research, 2*(6), 253-255.

Amin, J. M. (2016). *Quality assurance of the qualification process in TVET: Malaysia country*. TVET Online Asia. http://tvet-online.asia/wp-content/uploads/2020/03/mohd-amin_tvet7.pdf

Assunção, M. V. D., Araújo, A. G., & Almeida, M. R. (2019). The influence of family background on the access to technical and vocational education. *Journal of Public Administration, 53*(3), 542-559. https://doi.org/10.1590/0034-761220170352x

Aziz, A. (2019, July 24). Govt struggles to overcome vocational education misconception. *The Malaysian Reserved.* https://themalaysianreserve.com/2019/07/24/govt-struggles-to-overcome-vocational-education-misconception/

Azizi, N. A. (2018, Jun 29). Ubah stigma terhadap TVET [Change the stigma against TVET]. *Berita Harian Online*. https://www.bharian.com.my/berita/wilayah/2018/06/443161/ubah-stigma-terhadap-tvet

Babu, C., Varghese, R., & Manimozhi. (2020). Predicting student's performance using educational data mining. *International Journal of Scientific Research in Computer Science Applications and Management Studies, 9*(1), 1-4.

Bakar, A. R. (2011). *Roles of technical and vocational education and training (TVET)*. UPM Press.

Berawi, M. A. (2019). The role of industry 4.0 in achieving sustainable development goals. *International Journal of Technology*, *10*(4), 644-647. https://doi.org/10.14716/ijtech.v10i4.3341

Berrar, D. (2018). Bayes' theorem and naïve bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology, 1*, 403-412.

Breiman, L. (1996). Technical note: Some properties of splitting criteria. *Machine Learning, 24*, 41-47.

Chan, Y. S. (2018, November 23). We need to change perception of TVET. *The Star Online.* https://www.thestar.com.my/opinion/letters/2018/11/23/we-need-to-change-perception-of-tvet

Cheong, K., & Lee, K. (2016). Malaysia's education crisis- Can TVET help? *Malaysian Journal of Economic Studies, 53*(1), 115-134. https://doi.org/10.1787/888933003668

Gandhi, R. (2018). *Naïve Bayes classifier*. Towards Data Science. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Herlambang, A. D., Wijoyo, S. H., & Rachmadi, A. (2019). Intelligent computing system to predict vocational high school student learning achievement using naive bayes algorithm. *Journal of Information Technology and Computer Science, 4*(1), 15-25. https://doi.org/10.25126/jitecs.20194169

Hong, C. M., Ch'ng, C. K., & Roslan, T. R. N. (2021). Students' tendencies in choosing technical and vocational education and training (TVET): Analysis of the influential factors using analytic hierarchy process. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 2608-2615. https://doi.org/10.17762/turcomat.v12i3.1262

Hung, H. C., Liu, I. F., Liang, C. T., & Su, Y. S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry*, *12*(2), Article 213. https://doi.org/10.3390/sym12020213

Hussin, A., Mohamad, M., Hassan, R., & Omar, A. (2017). Technical vocational education training branding from perspective of stakeholder (parent) in Malaysia. *Advanced Science Letters, 23*(2), 1216-1219. https://doi.org/10.1166/asl.2017.7543

Imaz, M., & Sheinbaum, C. (2017). Science and technology in the framework of the sustainable development goals. *World Journal of Science, Technology and Sustainable Development*, *14*(1), 2-17. https://doi.org/10.1108/WJSTSD-04-2016-0030

Islam, M., Chen, G. R., & Jin, S. Z. (2019). An overview of neural network. *American Journal of Neural Networks and Applications, 5*(1), 7-11. https://doi.org/10.11648/j.ajnna.20190501.12

Ismail, A., & Hassan, R. (2013). Issues and challenges of technical and vocational education & training in Malaysia for knowledge worker driven. In *National Conference on Engineering Technology* (pp. 1-11). ResearchGate. https://doi.org/10.13140/2.1.4555.2961

Ismail, J., Chik, C. T., & Hemdi, M. A. (2021). TVET graduate employability: Mismatching traits between supply and demand. *International Journal of Academic Research in Business and Social Sciences, 11*(13), 223-243. https://doi.org/10.6007/IJARBSS/v11-i13/8522

Ismail, K., Nopiah, Z. M., Rasul, M. S., & Leong, P. C. (2017). Malaysian teachers' competency in technical vocational education and training: A review. In A. G. Abdullah, T. Aryanti, A. Setiawan & M. Alias (Eds.), *Regionalization and Harmonization in TVET* (pp. 59-64). Taylor & Francis Group.

Karim, M. A. (2018, September 5). TVET, a relevant choice. *New Straits Times.* https://www.nst.com.my/education/2018/09/408470/tvet-relevant-choice

Karim, Z. I. A., & Maat, S. M. (2019). Employability skills model for engineering technology students. *Journal of Technical Education and Training, 11*(2), 79-87. https://doi.org/10.30880/jtet.2019.11.02.008

Kelvin, G. (1997). *An introduction to neural networks*. UCL Press

Khan, I. A., & Choi, J. T. (2014). An application of educational data mining (EDM) technique for scholarship prediction. *International Journal of Software Engineering and Its Applications*, *8*(12), 31-42.

Koya, Z. (2019, July 4). TVET courses are not for those who are academically weak, Kula tells parents*. The Star Online*. https://www.thestar.com.my/news/nation/2019/07/04/tvet-courses-are-not-for-those-who-are-academically-weak-kula-tells-parents

KRI. (2018). *The school-to-work transition survey of young Malaysians.* Khazanah Research Institute. http://www.krinstitute.org/assets/contentMS/img/template/editor/20181205_SWTS_Main%20Book.pdf

Kukreja, H., Bharath, N., Siddesh, C. S., & Kuldeep, S. (2016). An introduction to artificial neural network. *International Journal of Advance Research And Innovative Ideas In Education, 1*(5), 27-30.

Liu, H., & Cocea, M. (2017). Semi-random partitioning of data into training and test sets in granular computing context. *Granula Computing, 2*, 357-386. https://doi.org/10.1007/s41066-017-0049-2

Matei, M. M. M., Mocanu, C., & Zamfir, A. M. (2018). Educational paths in Romania: Choosing general or vocational education. *HOLISTICA, 9*(2), 127-136. https://doi.org/10.2478/hjbpa-2018-0016

Messele, A., & Addisu, M. (2020). A model to determine factors affecting students academic performance: The case of Amhara region agency of competency, Ethiopia. *International Research Journal of Science and Technology, 1*(2), 75-87. https://doi.org/10.46378/irjst.2020.010202

Mustapha, R. B. (2015). Green and sustainable development for TVET in Asia. *The International Journal of Technical and Vocational Education*, *11*(2), 133-142.

Nguyen, Q, H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering, 2021*, Article 4832864. https://doi.org/10.1155/2021/4832864

Ozkan, U. B., Cigdem, H., & Erdogan, T. (2020). Artificial neural network approach to predict lms acceptance of vocational school students. *Turkish Online Journal of Distance Education, 21*(3), 156-169. https://doi.org/10.17718/tojde.762045

Rajaendram, R. (2020, November 6). Budget 2021: Association welcomes bigger allocation for TVET sector. *The Star Online*. https://www.thestar.com.my/news/nation/2020/11/06/budget-2021-association-welcomes-bigger-allocation-for-tvet-sector

Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563-591. https://doi.org/10.2190/CS.16.4.e

Rokach, L., & Maimon, O. (2015). *Data mining with decision trees theory and applications.* World Scientific Publishing Co. Pte. Ltd.

Sabang, A, (2017, September 2). Kesedaran TVET perlu dipertingkat [Awareness of TVET needs to be improved]. *Utusan Borneo Online.* https://www.utusanborneo.com.my/2017/09/02/kesedaran-tvet-perlu-dipertingkat

Sahu, H., Shrma, S., & Gondhalakar, S. (2011). A brief overview on data mining survey. *International Journal of Computer Technology and Electronics Engineering, 1*(3), 114-121.

Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Technology*, *8*(4C), 54-59.

Sivanandam, H., Rahim, R., Carvalho, M., & Tan, T. (2019, October 11). Budget 2020: Every single sen for education will be used properly, says Maszlee. *The Star Online.* https://www.thestar.com.my/news/nation/2019/10/11/budget-2020-every-single-sen-for-education-will-be-used-properly-says-maszlee

Sulaiman, N. L., & Salleh, K. M. (2016). The development of technical and vocational education and training (tvet) profiling for workforce management in Malaysia: Ensuring the validity and reliability of TVET data. *Man In India,* 96, 2825-2835.

TheStart. (2019, October 12). Thanks, but RM5.9bil not enough for TVET. *The Star Online*. https://www.thestar.com.my/news/nation/2019/10/12/thanks-but-rm59bil-not-enough-for-tvet

Trudis, H. (2014). *Secondary school students' perceptions of vocational education in Barbados*. https://docslib.org/doc/7204149/secondary-school-students-perceptions-of-vocational-education-in-barbados

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance, 20*, 679-686. https://doi.org/10.1016/S2212-5671(15)00123-9

Yaakob, H. (2017). Technical and vocational education & training (TVET) institutions towards statutory body: Case study of Malaysian polytechnic. *Advanced Journal of Technical and Vocational Education, 1*(2), 07-13. https://doi.org/10.26666/rmp.ajtve.2017.2.2

Yağcı, A., & Ḉevik, M. (2017). Predictions of academic achievements of vocational and technical high school students with artificial neural networks in science courses (physics, chemistry and biology) in Turkey and measures to be taken for their failures. *SHS Web Conference, 37*, Article 1057. https://doi.org/10.1051/shsconf/20173701057

Yamini, O., & Ramakrishna, S. (2015). A study on advantages of data mining classification techniques. *International Journal of Engineering Research & Technology*, *4*(9), 969-972.

Zakaria, M., AL-Shebany, M., & Sarhan, S. (2014). Artificial neural network: A brief overview. *International Journal of Engineering Research and Applications, 4*(2), 7-12.