Manuscript Title: Application of Decision Tree in Classifying Secondary School Students’ Tendencies to Choose TVET in Malaysia

Chia Ming Hong1, Chee Keong Ch’ng2*, TehRaihana Nazirah Roslan3

1,2 Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
3 Othman Yeop Abdullah Graduate School of Business, Universiti Utara Malaysia, 50300 Kuala Lumpur, Malaysia
1 carmenhongcm@gmail.com, 2*chee@uum.edu.my, 3 raihana@uum.edu.my

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract
The wave of Industry Revolution (IR 4.0) highlights the importance of technology in our life. The demand for technologist and skilled workers in Technical and Vocational Education and Training (TVET) are increasing day by day due to their expertise. TVET provides a platform for formal and non-formal learning to equip the youngsters in contributing to the development of a prosperous and inclusive nation. Moreover, TVET promises bright job prospects especially in fulfilling the manpower demand of IR 4.0. However, students in Malaysia currently are not fully aware of the existence of TVET, since the number of students who join TVET are still below expectation. Therefore, the main objective in this study is to develop the best TVET model to classify the students’ tendency in choosing TVET after completing secondary school. From the literature, five main factors that hinder students’ interest in joining TVET are recognized, namely students’ interest, parents, society, TVET instructors and employers. In this study, 428 secondary school students from Kedah (Malaysia) are involved as respondents. Different types of decision tree models are developed based on the algorithms and the splitting criteria. Altogether, there are 15 variables derived from 5 main affecting factors mentioned above to determine the tendency of joining TVET. Consequently, the best TVET classifier with the misclassification rate of 0.2919 is selected, to predict the tendency of students who will be joining TVET in future. Our findings revealed that the variable of “Stream” plays as the primary and trifling roles. This classifier is beneficial in assisting the government to achieve the aim of upholding TVET in Malaysia.

Keywords: TVET, Data mining, Classification, Decision tree

1. Introduction

To keep abreast with the era of Industry Revolution (IR 4.0), it is undeniable that technology nowadays plays an essential role in our daily life. Today, the presence of mobile phones and artificial intelligence-based appliances are not deemed as desires, but act as one’s necessities. Thus, the demand for technologist who can handle various machines is very crucial because they are the one who possess expertise in that field.

In Malaysia, Technical and Vocational Education and Training (TVET) is a channel to prepare students to get into IR 4.0 by providing them useful skills in workplace [1]. TVET is a dual teaching process that involves general education, technology knowledge and training which are closely linked with industry. It is a tool to produce skillful workers in the future [2]. The graduates from TVET can work under many fields such as electrical, construction, maintenance etc. In Malaysia, the government has been focusing on TVET by implementing many strategies such as allocating RM5.9 billion in Budget 2020, developing online database to concentrate on important sectors, organizing various exhibition tours, forming a committee known as TVET Empowerment Cabinet (JKKPTVET), implementation of various policies etc [3],[4],[5]. Despite the efforts made, it is frustrating that students nowadays are not really concern with the TVET’s existence. This can be observed from the number of students who registered in TVET was still below expectation [6]. Not only that, it was found that only 9% students chose polytechnics when selecting tertiary education channels [7].

Overall, there were many factors that affected students’ tendency in choosing TVET. The first factor was students’ personality and interest [8]. Interest was an element to encourage the students to join TVET if they were passionate in doing vocational activities such as repairing broken things, cutting hair etc. Next, social perception was another factor that caused the students’ repudiation towards TVET. Nowadays, the society usually had a pessimistic perception towards the TVET students because they were labelled as being weak in academic [8], [9]. The negativity had also affected the parents’ mindset. Parents’ perception was a crux to determine their children’s future [10]. Parents were usually reluctant to let their children join TVET since they perceived that TVET was not good for their child’s future. In addition, underestimation on TVET’s certification was also an issue which increased TVET’s unemployment rates. Cheong and Lee [11] discussed that employers had negative perceptions towards TVET graduates’ ability by assuming that TVET graduates were not intelligent.
in academic. Other than that, the salary of TVET graduates were lower than those who graduated from non-vocational and technical courses [12]. Finally, the last factor was TVET instructors who were teaching in the TVET institutions. Due to low salary payments, TVET institutions faced the issue of insufficient teaching staffs [13]. The quality of teacher was also a crucial part in the institution. A recent study revealed that the TVET instructors were found lacking in ICT skills, knowledge, and English skills [14].

The existence of the mentioned factors will affect a student’s tendency to enter TVET. In dealing with this problem, Decision Tree is chosen among the classification models in this research. Decision Tree is a popular approach in data classification since it is not affected by the missing values. Also, due to its transparency, Decision Tree can be easily interpreted by decision maker. It also allows pruning.

Decision tree was developed by J. Ross Quinlan [15]. It is a tree-based diagram which contains root node, internal node, and leaf nodes. Root node is on the top of the tree, representing the whole population and will be divided into two or more sets. Internal node is in the middle of the tree. The leaf node is the node on the most bottom which does not split further [16]. The structure of decision tree is displayed in Figure 1.

Figure 1. Structure of decision tree

Decision tree will split the variables with the highest homogeneity. In order to deal with overfitting problem, pruning is a crucial step in partitioning the data into smaller subsets. It is a process to remove the useless parts in the tree so that can reduce the size of the tree. There are four algorithms that are commonly used in decision tree namely, ID3, C4.5, CART and CHAID [16].

ID 3 was developed by Quinlan [15]. It can create the shortest and fastest tree but unable to handle the missing values and continuous data [17]. The splitting criteria in ID3 are Entropy and Information gain [18]. Entropy measures the unpredictability in the dataset. The entropy will be zero if the sample has a high homogeneity. The formula of entropy is,

\[ H(x) = -\sum_{x} p(x) \log_2 p(x) \]  

where 
\( x \) = Random variable, and 
\( p(x) \) = The possibility of result \( x \) of variable \( x \).

The measurement for entropy before and after splitting is known as information gain. The formula of information gain (IG) is given by,

\[ IG = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \]  

where 
\( n \) = number of attributes \( A \), 
\( |S_i| \) = number of cases in the partition \( S_i \), 
\( |S| \) = total number of cases in \( S \).

On the other hand, C4.5 was developed by Quinlan [15]. It is the extension of earlier ID3 algorithms, and can handle the missing value and continuous data [17]. The splitting criteria of C4.5 is information gain ratio. If the
data is under one branch, the gain ratio value will be smaller compared to data that are evenly distributed[19]. The formula of gain ratio is presented as,

\[
GAINRatio = \frac{GAIN(X,T)}{INFO(X,T)}
\]

where

\[
GAIN(X,T) = Info(T) - Info(X,T),
\]
\[
INFO(X,T) = - \sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot Info(T_i),
\]
\[
T_i = \text{Number of records in partition } i.
\]

Next, the Classification and Regression Tree (CART) was developed by Breiman et al. [20]. Besides handling the outliers, it can also handle the categorical or numerical variables. Meanwhile, Gini index will be used as splitting criteria to develop the binary tree classifier [17]. Gini index is a measurement of the impurity in the dataset [18]. The formula of Gini index is displayed as,

\[
Gini(S) = 1 - \sum_{c=1}^{k} p_j^2
\]

where

\[
p_j = \text{the proportion of the sample in class } c.
\]

Finally, the algorithm Chi-square Automatic Interaction Detector (CHAID) was developed by Kass [21]. It is used to develop non-binary tree and deals with large dataset[22]. Furthermore, Díaz-Pérez and Bethencourt-Cejas[23] claimed that CHAID has many advantages since it is a non-parametric statistic and can handle continuous and categorical attributes. For the continuous attribute, CHAID uses F-statistic to do the splitting for the dataset. The formula of F-statistic is written as,

\[
F_\alpha(k-1, N-k) = \frac{BSS/(k-1)}{WSS(N-k)}
\]

Where

\[
\alpha = \text{Critical value}
\]
\[
BSS = \text{Sum of squares between groups},
\]
\[
WSS = \text{Sum of squares within groups},
\]
\[
k = \text{Number of groups},
\]
\[
N = \text{Sum of sample size for the groups}.
\]

For the categorical attribute, CHAID uses Chi-square to split the criteria. The formula of chi-square is presented as,

\[
\chi^2 = \sum_{i=1}^{k} \frac{(Observed-Expected)^2}{Expected}
\]

where

\[
Observed = \text{Observed frequency for each category},
\]
\[
Expected = \text{Expected frequency for each category},
\]
\[
k = \text{Number of categories}.
\]

The splitting rules will affect the performance of Decision Tree [24]. Therefore, all the models created will be compared by using misclassification rate. A lower misclassification rate means it is a better model. The formula of misclassification rate is displayed as,

\[
\text{Misclassification rate} = \frac{FP+FN}{FP+FN+TP+TN}
\]

Where

\[
FN = \text{False negative, the actual class is positive but wrongly predicted as negative},
\]
\[
FP = \text{False positive, the actual class is negative class but wrongly predicted as positive},
\]
\[
TN = \text{True negative, the actual class is negative and correctly predicted as negative},
\]
\[
TP = \text{True positive, the actual class is positive and correctly predicted as positive}.
\]
Previous researchers often implemented decision tree in the education field. MaerMatei, Mocanu and Zamfir[25] had applied decision tree in determining students’ choice between general education or vocational education in Romania. They investigated four factors which are age, education, gender, and income among the students by using CART algorithm. As a result, the researchers found that the variable of age had the highest importance among all factors. Next, a study done by Gonzales, Lagman and De Angel [26] used Decision Tree algorithm to create a model to predict TVET students’ performance so that they could determine the students’ success in the future. The researchers had applied J48 method in their research. Furthermore, Hassan and Awan[27] used the data mining techniques in classifying the behavior of students enrolled in the TVET courses. In their research, the researchers had applied methods such as association, clustering, and classification methods in discovering the enrollment behavior to produce the skilled workers. For the classification model, the researchers applied Decision Tree, Random Forest, Neural Network and Naïve Bayes. The results showed that all the models had the same accuracy rate.

2. Materials and Methods

In this study, 428 respondents of Form 4 and Form 5 students were selected in Kedah, a Northern state in Malaysia. To ensure non-biasness, stratified random sampling and simple random sampling were used to determine the samples. Stratified random sampling method assisted in organizing the Kedah state based on District Education Offices (DEO) and regions. By applying stratified random sampling, 36 regions were classified into 9 DEOs. After that, simple random sampling method was applied in each DEO to find out the respective regions. Next, the questionnaires were distributed to the students in the listed schools. The questionnaire focused on five main factors namely students, TVET instructors, social perception, parents, and employers in the previous section.

The objective of this study is to develop a TVET classifier to classify the secondary school students’ tendency in choosing TVET. There are three outputs consisting of “yes”, “no”, and “maybe” respectively. “Yes” stands for “students will join TVET”, “No” stands for “students will not join TVET”, and “Maybe” stands for “students might be joining TVET”. Figure 2 displays the process of tree development in SAS E-miner.

![Figure 2. Tree development using SAS E-miner](image)

a) File import

In this study, 15 variables were imported from Microsoft Excel. The details of each variable are discussed as in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Level</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FamilyInvolved</td>
<td>Input</td>
<td>Nominal</td>
<td>The family members of student join the TVET.</td>
</tr>
<tr>
<td>FatherEducation</td>
<td>Input</td>
<td>Nominal</td>
<td>The highest education of the student’s father.</td>
</tr>
<tr>
<td>FatherJob</td>
<td>Input</td>
<td>Nominal</td>
<td>The occupation of the student’s father</td>
</tr>
<tr>
<td>Gender</td>
<td>Input</td>
<td>Binary</td>
<td>Gender of student</td>
</tr>
</tbody>
</table>
b) Impute

Decision tree required less data cleaning compared to some other methods as it is not affected by missing values and outliers. Data removal is preferable for small number of missing data values whereas data replacement is more appropriate for large number of missing data values [28]. Moreover, the best surrogate predictor can be used when the value of the optimal split predictor for an observation is missing. Decision tree is normally robust to the outliers. Most likely outliers will have a negligible effect because the nodes are determined based on the sample proportions in each split region [29]. However, there are some approaches that researchers used to handle the anomalies such as Robust Decision Tree [30] and Winsorized tree [31]. In short, Decision Tree is insensitive towards the presence of missing values and outliers. Hence, it produces a comprehensive analysis of the consequences along each branch.

c) Data Partition

In this node, the dataset was allocated into 70% training sets and 30% testing set. From the 428 samples, 299 data will be set as training data, while 129 data will be set as testing data.

d) Decision Tree models

Four types of decision trees were created based on the splitting criteria and maximum branches. Since the variables consist of nominal and ordinal, therefore the splitting criteria were based on Gini Index and Entropy. Gini for the nominal attributes and Entropy for the ordinal attributes were selected in Decision Tree 1, whereas Gini for both nominal and ordinal attributes were selected for Decision Tree 2. Decision Tree 3 and Decision Tree 4 used Entropy for the nominal attributes and Gini for the ordinal attributes, and Entropy for both nominal and ordinal attributes, respectively. As we discussed before, Entropy is used to split the multiple branches while Gini is used to split the binary branches. Hence, Decision Tree 1 and Decision Tree 4 had the maximum branch of 3, whereas Decision Tree 2 and Decision Tree 3 had the maximum branch of 2. Table 2 shows the details for Decision Tree models.

<table>
<thead>
<tr>
<th>Decision Tree models</th>
<th>Splitting criteria</th>
<th>Maximum branches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree 1</td>
<td>Nominal target: Gini, Ordinal target: Entropy</td>
<td>3</td>
</tr>
<tr>
<td>Decision Tree 2</td>
<td>Nominal target: Gini, Ordinal target: Gini</td>
<td>2</td>
</tr>
<tr>
<td>Decision Tree 3</td>
<td>Nominal target: Entropy, Ordinal target: Gini</td>
<td>2</td>
</tr>
<tr>
<td>Decision Tree 4</td>
<td>Nominal target: Entropy, Ordinal target: Entropy</td>
<td>3</td>
</tr>
</tbody>
</table>

e) Model comparison

| Table 2. Decision Tree models |
In the model comparison node, the decision tree models were compared using the misclassification rate. The model with the lowest misclassification rate is the best model to predict the tendency of students in selecting TVET in future.

3. Results and Discussion

The Model comparison node shows the misclassification rate for each model. From the results obtained, the best Decision Tree model is Decision Tree 3, which has the lowest misclassification rate (0.2919) by simpler classifier. Table 3 shows the misclassification rate for the Decision Tree.

<table>
<thead>
<tr>
<th>Decision tree models</th>
<th>Misclassification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree 1</td>
<td>0.3154</td>
</tr>
<tr>
<td>Decision Tree 2</td>
<td>0.3221</td>
</tr>
<tr>
<td>Decision Tree 3</td>
<td>0.2919</td>
</tr>
<tr>
<td>Decision Tree 4</td>
<td>0.3054</td>
</tr>
</tbody>
</table>

Decision Tree 3 has binary branches with a depth of 7. Hence, the variables with higher importance is assigned at higher levels. Figure 3 shows the complete Decision Tree 3 model.

Since Figure 3 is too large and cannot be seen clearly, it is divided into left side and right side in Figure 3 (a) and Figure 3 (b).
The variables with highest homogeneity are selected along the splitting process. Thus, some unimportant features might be chosen if it does not contribute to the decrease of impurities. Based on the result, variable of “Stream” had the highest importance (1.0000), followed by IMP_MotherEducation (0.8684), VocationalCourseInteresting (0.8679), TVETStudHigherJobChance (0.8166), Gender (0.7108), IMP_MotherJob (0.6269), IMP_FatherJob (0.5621), TVETBright (0.4853), VocationalSubjStatus (0.3872), IMP_FatherEducation (0.3647), FamilyInvolved (0.3339), and InterestedField (0.3177). Two variables had the importance of “0.0000”, namely TVETStudHighSalary and HeardTVET. The summary of importance for each variable is displayed in Table 4.

Table 4. Importance of variable

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream</td>
<td>1.0000</td>
</tr>
<tr>
<td>IMP_MotherEducation</td>
<td>0.8684</td>
</tr>
<tr>
<td>VocationalCourseInteresting</td>
<td>0.8679</td>
</tr>
<tr>
<td>TVETStudHigherJobChance</td>
<td>0.8166</td>
</tr>
<tr>
<td>Gender</td>
<td>0.7108</td>
</tr>
<tr>
<td>IMP_MotherJob</td>
<td>0.6269</td>
</tr>
<tr>
<td>IMP_FatherJob</td>
<td>0.5621</td>
</tr>
<tr>
<td>TVETBright</td>
<td>0.4853</td>
</tr>
<tr>
<td>VocationalSubjStatus</td>
<td>0.3872</td>
</tr>
<tr>
<td>IMP_FatherEducation</td>
<td>0.3647</td>
</tr>
<tr>
<td>FamilyInvolved</td>
<td>0.3339</td>
</tr>
<tr>
<td>InterestedField</td>
<td>0.3177</td>
</tr>
<tr>
<td>TVETStudHighSalary</td>
<td>0.0000</td>
</tr>
<tr>
<td>HeardTVET</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Other than that, Decision Tree can be linearized into a set of decision rules that represent the view for the whole system as shown Table below.

Table 5. Details for each node

<table>
<thead>
<tr>
<th>Node</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node = 12</td>
<td>IF VocationalCourseInteresting &lt;= 3 AND Stream IS ONE OF: SCIENCE or MISSING THEN “Maybe” (67%), “No” (33%).</td>
</tr>
<tr>
<td>Node = 13</td>
<td>IF VocationalCourseInteresting &lt;= 3 AND Stream IS ONE OF: ACCOUNT THEN “Maybe” (12%), “NO” (88%).</td>
</tr>
</tbody>
</table>
Node = 20  
IF VocationalCourseInteresting >= 5 AND TVETStudHigherJobChance <= 3 AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND InterestedField IS ONE OF: MANUFACTURING, CONSTRUCTION or MISSING THEN “Maybe” (100%).

Node = 21  
IF VocationalCourseInteresting >= 5 AND TVETStudHigherJobChance >= 4 or MISSING AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND InterestedField IS ONE OF: MANUFACTURING, CONSTRUCTION or MISSING THEN “Yes” (88%), “Maybe” (12%).

Node = 25  
IF VocationalCourseInteresting >= 5 or MISSING AND Stream IS ONE OF: TECHNICAL, ACCOUNT, SCIENCE AND FamilyInvolved IS ONE OF: YES THEN “Yes” (17%), “Maybe” (83%).

Node = 27  
IF VocationalCourseInteresting <= 4 or MISSING AND TVETBright >= 3 or MISSING AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND Imputed: MotherEducation IS ONE OF: SPM, NO COMPLETE SECONDARY SCHOOL, DEGREE or MISSING AND Imputed: FatherJob IS ONE OF: CLERICAL/ CIVIL SERVICE, UNEMPLOYED, AGRICULTURE THEN “Maybe” (72%), Yes (9%), No (19%).

Node = 29  
IF VocationalCourseInteresting <= 4 or MISSING AND TVETBright >= 3 or MISSING AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND Imputed: MotherJob =Missing AND Imputed: MotherEducation IS ONE OF: DIPLOMA, STPM THEN “Yes” (100%)

Node = 32  
IF VocationalCourseInteresting >= 5 AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND InterestedField IS ONE OF: CLERICAL/ CIVIL SERVICE, SELF-EMPLOYED, PROFESSIONAL AND Imputed: FatherEducation IS ONE OF: NO COMPLETE SECONDARY SCHOOL or MISSING AND Gender IS ONE OF: MALE or MISSING THEN “Yes” (100%)

Node = 37  
IF VocationalCourseInteresting >= 4 or MISSING AND TVETStudHigherJobChance >= 4 AND Stream IS ONE OF: TECHNICAL, ACCOUNT, SCIENCE AND Imputed: MotherEducation IS ONE OF: SPM, DIPLOMA, DEGREE or MISSING AND FamilyInvolved IS ONE OF: NO or MISSING THEN “Maybe” (72%), “Yes” (16%), “No” (12%).

Node = 38  
IF VocationalCourseInteresting >= 4 or MISSING AND TVETStudHigherJobChance <= 3 AND Stream IS ONE OF: TECHNICAL, ACCOUNT, SCIENCE AND Imputed: MotherEducation IS ONE OF: NO COMPLETE SECONDARY SCHOOL, PHD AND FamilyInvolved IS ONE OF: NO or MISSING THEN “Maybe” (80%), “No” (20%).

Node = 39  
IF VocationalCourseInteresting >= 4 or MISSING AND TVETStudHigherJobChance >= 4 or MISSING AND Stream IS ONE OF: TECHNICAL, ACCOUNT, SCIENCE AND Imputed: MotherEducation IS ONE OF: NO COMPLETE SECONDARY SCHOOL, PHD AND FamilyInvolved IS ONE OF: NO or MISSING THEN “Maybe” (14%), “No” (86%).

Node = 48  
IF VocationalCourseInteresting >= 5 AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND InterestedField IS ONE OF: CLERICAL/ CIVIL SERVICE, SELF-EMPLOYED, PROFESSIONAL AND Imputed: MotherJob IS ONE OF: SELF-EMPLOYED or MISSING AND Imputed: FatherEducation IS ONE OF: SPM THEN “Yes” (38%), “No” (63%).

Node = 49  
IF VocationalCourseInteresting >= 5 AND Stream IS ONE OF: ARTS, BUSINESS, ICT, LITERATURE, VOCATIONAL or MISSING AND InterestedField IS ONE OF: CLERICAL/ CIVIL SERVICE, SELF-EMPLOYED, PROFESSIONAL AND Imputed: MotherJob IS ONE OF: OTHERS AND Imputed: FatherEducation IS ONE OF: SPM THEN “Yes” (80%), “No” (20%).
4. Conclusions

In this paper, the best TVET classifier is developed for classifying and predicting the tendency of students in choosing TVET after their secondary school. The decision tree model can be used to increase the number of TVET students in the future, by targeting potential students based on our findings. Since our results revealed that the variable of “Stream” has the highest importance, the school teachers should play vital roles in raising the awareness on students’ capabilities. In addition, different parties should take responsibilities in raising the awareness on student’s capabilities. For example, teachers should always focus on those students who are interested in joining TVET and provide them the proper information needed. Other than that, the government must organize more programs or training for students in schools to increase students’ awareness towards TVET’s attractiveness. Besides that, non-government organizations (NGO) should provide scholarships or incentives for TVET courses to attract more students to join TVET.

5. Funding Statement

The authors received funding from Ministry of Education (MOE) through Fundamental Research Grant Scheme (FRGS) with S/O code: 14868.

6. Acknowledgments

This research was supported by Ministry of Education (MOE) through Fundamental Research Grant Scheme S/O Code 14868 (FRGS/1/2020/SSI0/UUM/02/18). The authors also want to extend their gratitude to the editor of the journal and reviewers for their valuable suggestions on the paper.

References


