Analysis of Student Demographic Information Using Data Mining classification with Decision Tree

Citra Kurniawan[1], Ifit Novita Sari[2], Diah Puji Nali Brata[3]
{airakurniawan@gmail.com[1], vita@unikama.ac.id[2], diah.puji@stkipjb.ac.id[3]}

Sekolah Tinggi Teknik Malang, Jl. Soekarno Hatta Nomor 94 Malang, East Java, Indonesia[1],
University of Kanjuruhan Malang, Jl. S. Supriadi 48 Malang, East Java, Indonesia[2],
STKIP PGRI Jombang, Jl. Pattimura III No. 20 Jombang, East Java, Indonesia[3].

Abstract. The purpose of this study was to process Student Demographic Information using data mining analysis using decision tree technique. This study retrieved data from 50 students with visual data attributes - verbal preferences, self-efficacy, gender, and interest. This study uses orange data mining to process data with decision tree technique. The study found that decision node as the best predictors are visual preferences, in which visuals had a 76% of 50 attributes. Female students had a 100% distribution as a visual preference, while male had 68.4% of the distribution as a visual preference. The results found that the attributes that predictor were visual preferences. The decision tree gets the rule from the root node to the leaf nodes as many as four rules are R1, R2, R3, and R4.

Keywords: Data Mining, Decision Tree, Student Demographic.

1 Introduction

Student demographic information can be valuable information if treated well. Student demographic information consists of domain-specific information (DSI) and domain-independent information (DII). The DSI contains specific information such as students' level of knowledge, knowledge insight, and learning habits, while DII tends toward student learning objectives, cognitive abilities, motivations, preferences and student backgrounds [1]. It agrees with Esichaikul et al. (2011) stating that the demographic information of the students contains seminars on matters relating to student information such as behavior, learning levels and other information [2]. Student information data is an attribute of gender, the result of initial ability level assessment and learning preference [3]. One method to process demographic data information of students is data mining classification with decision tree technique.

A decision tree is a data mining technique that uses supervised learning for object classification. The decision tree divides data into multiple sets based on input variables in the form of a tree structure hierarchical. A decision tree is a statistical tool for classification, interpretation [4], and the best choice [5]. The decision tree involves the collection of data/variables, classification, and analysis of data/variables that aim to predict outcomes. The classification of the data on the decision tree consists of two phases: 1) Grouping of data/variables in the form of classification patterns/rules as a classification model; 2) Use of classification models and test data to estimate the accuracy of classification patterns [6].

2 Methodology

Demographic information of students processed by decision tree technique in this study is visual-verbal preferences, self-efficacy, gender, and interest. The decision tree technique establishes a classification in the form of a tree based on the information already collected. The decision tree technique breaks the data continuously into smaller sets until it obtains the final result of decision nodes and leaf nodes. The data in this study consists of 50 student, where each data has different demographic information, shown in Figure 1. The algorithm for constructing the decision tree is ID3 developed by J. R. Quinlan (1986) [7]. ID3 consists of two processes: entropy and information gain. Entropy is the top-down decision tree building process by calculating the homogeneity of the sample which is then derived from the information gain after attributes divide the dataset. This study uses Orange software to classify data Mining classification with Decision Tree. Figure 1 shows data of the data mining classification consists of four attributes: visual-verbal preferences, self efficacy, gender and interest in image objects. Visual-verbal attribute preferences have two class: visual and verbal. Visual-verbal measurement preferences use visual-verbal questionnaire (VVQ) developed by Richarson (1988) and Kirby (1988) [3], [8], [9]. The self-efficacy attribute has two class: high and low. Measurement of self efficacy level using self-efficacy questionnaire developed by Bandura (2005) [10]. The gender attribute has two class: male and female. Attributes of interest, especially in the interest of the object image has two class of interest in image object and no interest in image object.

	Visual-Verbal Preferences	Self Efficacy	Gender	Interest in Image Object
1	Visual	High	Male	Interest with Image
2	Visual	High	Male	No interest with image
3	Visual	High	Male	Interest with Image
4	Visual	High	Male	Interest with Image
5	Visual	High	Male	Interest with Image
6	Verbal	High	Male	Interest with Image
7	Verbal	High	Male	No interest with image
8	Verbal	Low	Male	Interest with Image
9	Visual	Low	Male	No interest with image
10	Visual	Low	Male	No interest with image
11	Visual	High	Male	Interest with Image
12	Visual	Low	Male	No interest with image
13	Visual	Low	Male	Interest with Image
14	Verbal	Low	Male	No interest with image
15	Visual	Low	Male	Interest with Image
16	Visual	High	Male	Interest with Image
17	Visual	Low	Male	Interest with Image
18	Visual	Low	Female	Interest with Image
19	Visual	Low	Female	Interest with Image
20	Visual	High	Male	Interest with Image
21	Verbal	Low	Male	No interest with image
22	Visual	Low	Male	No interest with image
23	Visual	Low	Male	Interest with Image
24	Visual	Low	Male	No interest with image
25	Visual	High	Female	No interest with image
26	Visual	High	Male	Interest with Image
27	Visual	Low	Male	Interest with Image
28	Visual	Low	Male	No interest with image
29	Visual	Low	Male	No interest with image
30	Visual	High	Female	No interest with image
31	Verbal	High	Male	Interest with Image
32	Visual	Low	Female	No interest with image
33	Visual	Low	Female	Interest with Image
34	Visual	Low	Male	No interest with image
35	Visual	Low	Female	Interest with Image
36	Visual	High	Female	Interest with Image
37	Visual	High	Female	Interest with Image
38	Verbal	Low	Male	No interest with image
39	Verbal	Low	Male	No interest with image
40	Visual	Low	Male	Interest with Image
41	Visual	High	Female	Interest with Image
42	Verbal	High	Male	Interest with Image
43	Visual	Low	Female	Interest with Image
44	Verbal	Low	Male	No interest with image
45	Verbal	Low	Male	Interest with Image
46	Visual	Low	Male	No interest with image

Fig. 1. Student Demographic Information

Data in data mining is an instance that has attributes and classes. The attribute is a description that belongs to the data, while the class is the status of the instance or the

conclusion of each data. Classification analyzes a set of training data and builds a model on each class. The classification phase in data mining consists of Classification model development, model implementation, and evaluation. Model development can be seen in the following models:

```
If (preferences=visual) then Interest in image=Yes
Else If (preferences=verbal)
If (self efficacy=high) then Interest in image=No
Else If (self efficacy=low) Interest in image=Yes
```

The classification model if written in the flowchart as follows:

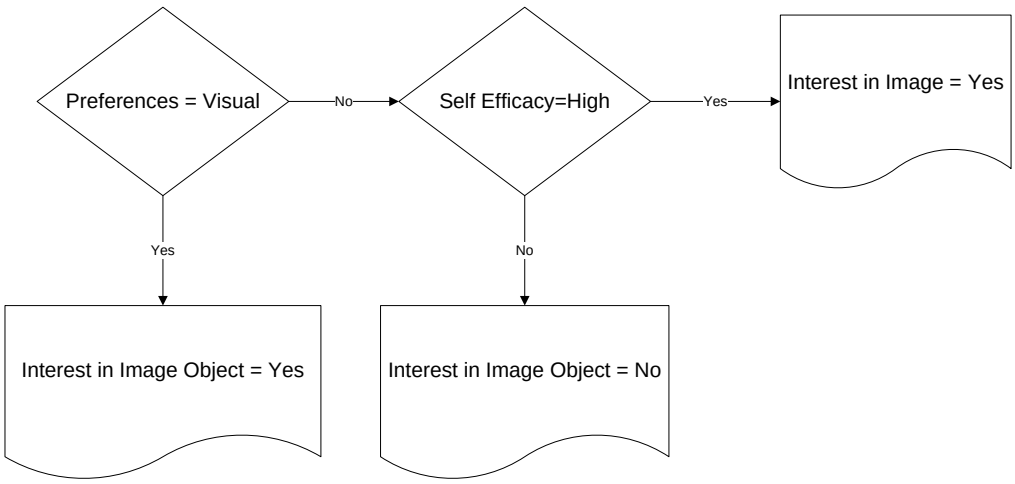

Fig. 2. Flowchart of Classification Model

Figure 2 shows the flowchart of the classification model used. The state of the classification rule indicates that if a person has a visual preference, then he/she has an interest in the image object. If someone has a non-visual preference (verbal preferences), then he/she will get a further test in the form of self-efficacy level validation. If a person has a high degree of self-efficacy, then he/she has an interest in the image object. If a person has a low self-efficacy, then he/she has no interest in the image object.

Student data information as input data is analyzed based on decision tree analysis. The decision tree analysis incorporates data on the classification model. Then the result of the selected classification model as the subset data is grouped into several sets, as shown in Figure 3.

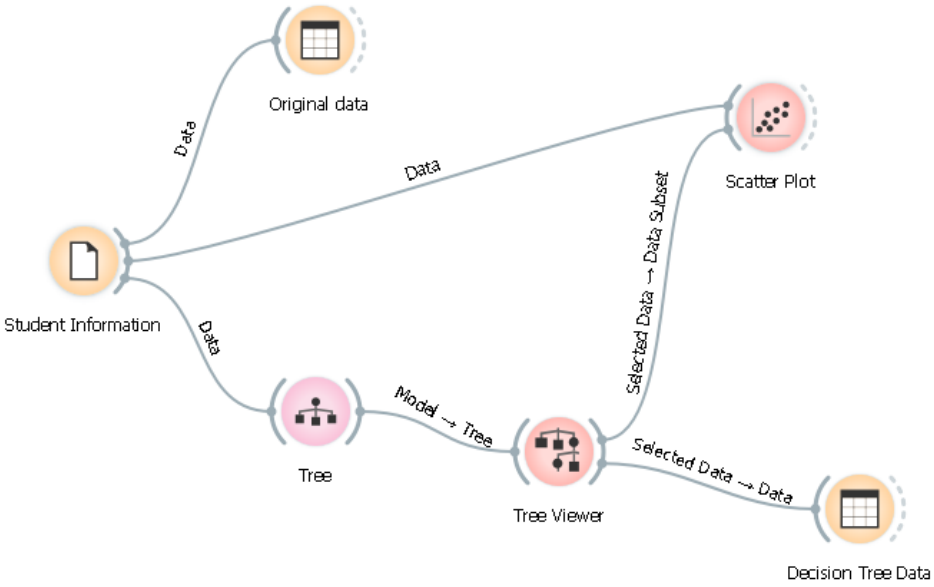

Fig. 3. The framework of Decision Tree with Orange Data Mining

Figure 3 shows student demographic student acting as input data. Decision tree analysis takes input data and forms a classification model in the decision tree. The classification model selects attribute data and inserts the data in pre-established models. The decision tree output shows a hierarchy containing attribute data and one of the attributes that act as a predictor.

3 Results and Discussion

The study found that decision node as the best predictors are visual preferences where visuals had 76% distribution or 38 data of 50 data. Female students had 100% distribution as visual preferences, while male had 68.4% distribution as visual preferences (26 of 38 data, 12 data as verbal preferences), as shown in Figure 4.

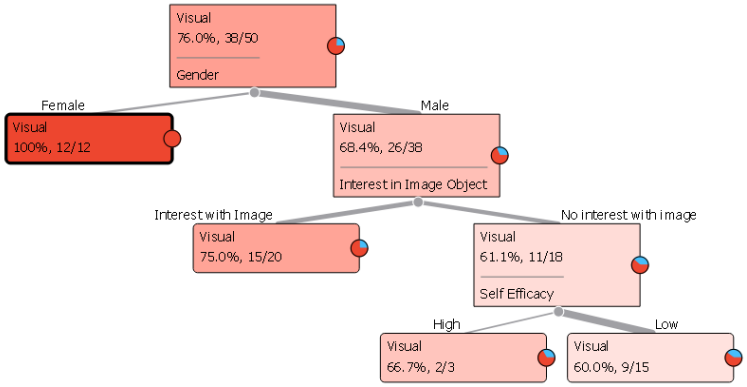

Fig. 4. The Decision Tree of study

Figure 4 shows the representation that maps the set of attributes to a predefined class of visual preferences as predictors. Visual preferences are divided into two categories: visual preferences in male and visual preferences in the female, and then related with interest attributes (interest in image object and no interest in image object) and self-efficacy attributes (high and low).

Data classification has the distribution of dominant data on visual preferences compared to verbal preferences. The scatter plot shows the tendency of attribute data preference, self-efficacy, and interest in the image, as shown in Figure 5.

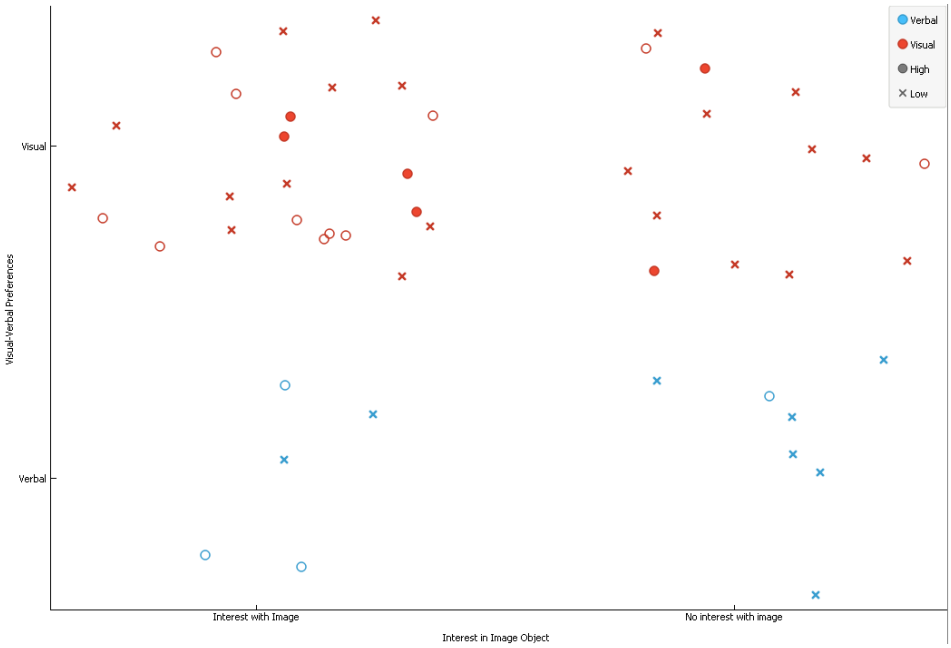

Fig. 5. Scatter plot of demographic information

Figure 5 helps identify measurable relationships between two sets of visual-verbal preferences and interest in object attributes. The visual preferences attribute shows that there is no significant difference between the attributes of interest in high and low self-efficacy. In the verbal attribute preferences have a little data distribution so that verbal not become predictor.

The decision rule on the decision tree gets the rule set from the root node to the leaf node as follows:

R_1: **IF** (Gender = Female) **THEN** (Visual)
R_2: **IF** (Gender = Male) **THEN** Interest in the image object
R_3: **IF** (Gender = Male) **AND** (No Interest in image object) **THEN** Self Efficacy=High
R_4: **IF** (Gender = Male) **AND** (No Interest in image object) **THEN** Self Efficacy=Low

4 Conclusion

Focus on this study is data processing by using data mining classification with a decision tree. The study gets the result that the attributes that become predictor are visual preferences. There are several decision rules as the data classification.

References

[1] V. Vagale and L. Niedrite, "Learner Model ' s Utilization in the e-Learning Environments," *CEUR Workshop Proc.*, vol. 924, 2012.

[2] V. Esichaikul, S. Lamnoi, and C. Bechter, "Student Modelling in Adaptive E-Learning Systems," *Knowl. Manag. E-Learning An Int. J.*, vol. 3, no. 3, 2011.

[3] C. Kurniawan, P. Setyosari, W. Kamdi, and S. Ulfa, "Electrical engineering student learning preferences modelled using k-means clustering," *Glob. J. Eng. Educ.*, vol. 20, no. 2, 2018.

[4] Yan-yan Song and Ying Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, 2015.

[5] D. L. Olson, *Descriptive Data Mining*. Gateway East, Singapore: Springer Nature, 2017.

[6] D. Singh, H. Naveen, and J. Samota, "Analysis of Data Mining Classification with Decision Tree Technique," *Glob. J. Comput. Sci. Technol.*, vol. 13, no. 13, 2013.

[7] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, 1986.

[8] A. Richardson, "Verbalizer-Visualizer: A Cognitive Style Dimension," *J. Ment. Imag.*, vol. 1, pp. 109–126, 1977.

[9] J. R. Kirby, P. J. Moore, and N. J. Schofield, "Verbal and visual learning styles," *Contemp. Educ. Psychol.*, vol. 13, no. May 2014, pp. 169–184, 1988.

[10] A. Bandura, "Guide For Constructing Self-Efficacy Scales," in *Self-Efficacy Beliefs of Adolescents*, 2005, pp. 307–337.