Sieve Diagram For Data Exploration of Instagram Usage Habit Obtained From Indonesia Questioner's Sample

Reko Syarif Hidayatullah¹, Wahyu Nur Cholifah¹, Erlin Windia Ambarsari¹*, Nunu Kustian¹, Siti Julaeha¹

¹Informatics Department, Universitas Indraprasta PGRI, Jakarta, Indonesia

*erlinunindra@gmail.com

Abstract. Exploration data using a scatter plot made it more accessible when the datasets correlated. However, the case of Instagram Usage Habit in the previous study was hard to follow. The problem was that many datasets were not specific. Therefore, it difficult to classification for a Decision Tree. The other option of exploration data used the sieve diagram. The sieve diagram summarizes the relationship between the categorical variables using frequencies—the purpose of this study to understand the data and found out what wrong in the datasets. Based on the result of the sieve diagram in the study, the main problem found out in the age of the respondent on attributes. We deduce that several attributes had not characteristic unique for the habit of using Instagram because the attributes based on age have a similar pattern. We suggest that the questions for respondents need to be improved, such as Tiered questions. Therefore, The classification of decision trees would become more precise in the subsequent studies.

1. Introduction

Decision Tree Construction is serving as a dataset visualization in association statement IF-THEN—two kinds that use for Decision Tree visualization; Hierarchy Tree¹ and Pythagoras Tree². However, construct Decision Tree depends on the classification algorithm, such as ID3 [3], C4.5[4], C5.0[5], when the attribute classified as a category. Besides, when it has a numeric continuous data type, it as a known regression algorithm.

Based on the previous study by [6], construct Pythagoras Tree using the ID3 algorithm based on Deviation Standard. The datasets obtained from questioners that needed observation of the case. The case in this study purpose to found out the habit of netizen using Instagram, especially the young generation. However, the result to be confused because the datasets seem uncomplete. Likewise, it becomes ambiguous. Therefore, we need to exploration the data first to understanding the data and find out what wrong in the datasets.

In the study of [7], exploration data using a scatter plot to know distribution data presented by a dot. A scatter plot that makes split data easier to the classification from data dissemination. In general, the dataset needs to correlate for information entropy, which based on the probability distribution shape [8]. Therefore, the dataset in random split dependent on two attributes until datasets have a higher possibility for classification. However, Instagram Usage Habit's case has a problem. Many datasets are not specific; therefore, few datasets cannot correlate by two attributes in a scatter plot.
The other option for exploration data is the sieve diagram. This diagram represents contingency tables in which the row and column variables are independent. The sieve diagram summarizes the relationship between the categorical variables using frequencies[9]. Therefore, we analyze based on the pattern of association to discover dataset relation of Instagram Usage Habit.

2. Methodology
Datasets shall be classified based on each of the two separate criteria, and each of these criteria may contain only a finite number of distinct values[10]. It as defined a two-way contingency table. The sieve diagram is categorical data visualization to present a two-way contingency table. Categorical data means the dataset of Instagram Usage Habit had able grouping. For example, the target data "Age" represented in the previous study as a continuous data criterion[6]. "Age" in Instagram Usage Habit be able to play a role as category data. For more details, the following are raw data from Instagram Usage Habit:

| Age | Initials | The Propose to use Instagram | Download Content | Frequency of using Instagram | The duration of using Instagram | Cumulative of Using Instagram |
|-----|----------|------------------------------|------------------|-----------------------------|-------------------------------|------------------------------|
| 1   | A1       | Communicate Interactively    | Another person's photo | 1 to 3 times                | < 5 minutes                   | <0.5 hour                    |
| 2   | A2       | Uploading edited video       | Video cinemtic     | 1 to 3 times                | < 5 minutes                   | <0.5 hour                    |
| 3   | A3       | Communicate Interactively    | Group Photo        | 1 to 3 times                | 5 to 15 minutes               | 0.5 to 1 hour                |
| 4   | A4       | Uploading Content            | Selfie            | 1 to 3 times                | 30 to 60 minutes              | 1 to 2 hours                 |
| 5   | A5       | Uploading Content            | Selfie            | 1 to 3 times                | 30 to 60 minutes              | 1 to 2 hours                 |
| 6   | A6       | Communicate Interactively    | Selfie            | >9 times                    | 30 to 60 minutes              | 1 to 2 hours                 |
| 7   | A7       | Uploading Content            | Objects Photo     | 4 to 6 times                | 5 to 15 minutes               | 0.5 to 1 hour                |
| 8   | A8       | Communicate Interactively    | Selfie            | >9 times                    | 15 to 30 minutes              | 1 to 2 hours                 |
| 9   | A9       | Posting and Information      | Objects Photo     | 4 to 6 times                | 5 to 15 minutes               | 2 to 4 hours                 |
| 10  | A10      | Surfing                      | Objects Photo     | 1 to 3 times                | 15 to 30 minutes              | 2 to 4 hours                 |
| 11  | A11      | Surfing                      | Selfie            | 1 to 3 times                | < 5 minutes                   | <0.5 hour                    |
| 12  | A12      | Communicate Interactively    | Selfie            | 7 to 9 times                | 30 to 60 minutes              | 1 to 2 hours                 |
| 13  | A13      | Surfing                      | Selfie            | 1 to 3 times                | 15 to 30 minutes              | 0.5 to 1 hour                |
| 14  | A14      | Uploading Content            | Selfie            | <60 times                   | 4 to 6 hours                  |
| 15  | A15      | Surfing                      | Selfie            | >9 times                    | 5 to 15 minutes               | 4 to 6 hours                 |
| 16  | A16      | Posting and Information      | Objects Photo     | 4 to 6 times                | 15 to 30 minutes              | 1 to 2 hours                 |
| 17  | A17      | Communicate Interactively    | Selfie            | 1 to 3 times                | 15 to 30 minutes              | <0.5 hour                    |

Figure 1. Instagram Usage Habit Sample Data

Instagram Usage Habit data obtain from 33 Respondents who fill the question most of is Indonesian young adults, which is between the ages of 18 to 22. Based on Figure 1, it was challenging to identify which age have difference habit. Therefore, when it classification to Decision Tree. The data become ambiguous. The sieve diagram has a part in exploration data to identify the pattern of association to discover the sample's problem.
Therefore, we analyze a sieve diagram by considering the requirements as below:

a. One unit square in the contingency table divided into rectangles, one for each cell. The meaning is that it represents two-way tables (for example, I x J table), the expected frequencies of cells under independence. It described as a total of IJ rectangle to construct rectangular form, which is each of the blocks has height and width propositional to the appropriate marginal frequencies for respectively of the row and column [11].

b. For instance, the two-way table presented based on Figure 1, as follows:

|                | 18  | 19  | 20  | >=21 | Total |
|----------------|-----|-----|-----|------|-------|
| Uploading edited video | 0 (0%) | 0 (0%) | 0 (0%) | 1 (3%) | 1 (3%) |
| Uploading Content      | 2 (6%) | 3 (9%) | 2 (6%) | 2 (6%) | 9 (28%) |
| Surfing                | 1 (3%) | 3 (9%) | 1 (3%) | 2 (6%) | 7 (22%) |
| Posting and Information| 2 (6%) | 1 (3%) | 0 (0%) | 0 (0%) | 3 (9%) |
| Communicate Interactively | 3 (9%) | 6 (19%) | 3 (9%) | 0 (0%) | 12 (38%) |
| Total                   | 8 (25%) | 13 (41%) | 6 (19%) | 5 (16%) | 32 (100%) |

c. The hypothesis evidence that the variable is independent or not independent, giving it a test with chi-square [12]:

\[ x^2 = \sum \frac{(O-E)^2}{E} \]  

(1)

O is the observed cell frequency, which measures the data of events. Observed frequency showed in Table 1. E is the expected cell frequency, in which the estimated predicted frequency derived from an experiment assumed to be authentic unless otherwise suggested by observational proof in a hypothesis test. The expected frequency calculated showed in equation 2 [12]. Furthermore, \( \Sigma \) is the sum of all cells in the table.

The expected frequency = \( \frac{\text{row total} \times \text{column total}}{\text{overall total}} \)  

(2)

d. Determine p-value in the chi-square test for the null hypothesis evidence.

e. The conclusion on hypothesis influenced by the degree of freedom (df) which formula is [12]:

\[ df = (\text{number of rows} - 1) \times (\text{number of columns} - 1) \]  

(3)

f. The discrepancy between the frequency of observed and expected occurs as the shading density, using the color to showed that the variance from independence is positive (blue) or negative (red). The density of shading influenced to the standardized Pearson residual for independence is [13]:

\[ \text{the standardized Pearson residual} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij} \times (1-p_i) \times (1-p_j)}} \]  

(4)

\( p_i \) and \( p_j \) are the relative marginal proportions equal.

3. Result and Discussion

Instagram Usage Habit's Sieve Diagrams represent as below:
Diagram for Age and The Propose To Use Instagram in Table 1, as an example, which it showed in Figure 3. The categorical dataset in a relation between two attributes is independence, which is $X^2 = 12.54790395 < 12.026$ or P-value = 0.402807 < 0.9, with twelve degree of freedom.

**Table 2.** The Standardized Pearson Residual For Age And The Propose To Use Instagram

|                  | 18     | 19     | 20     | >=21   |
|------------------|--------|--------|--------|--------|
| Uploading edited video | -0.59  | -0.84  | -0.49  | 2.36   |
| Uploading Content  | -0.23  | -0.53  | 0.31   | 0.64   |
| Surfing           | -0.74  | 0.14   | -0.34  | 1.07   |
| Posting and Information | 1.75  | -0.27  | -0.87  | -0.78  |
| Communicate Interactively | 0.00  | 0.84   | 0.70   | -1.89  |

The respondents in Figure 3 aim to use Instagram as follows: respondents aged over 20 years upload an edited video. Content uploading was done by ages 20 to 22. Age 19, and older than 20, surfaced. Age 18 did posting and information searching. They interactively interacted between the ages of 19 and 20.
Nevertheless, the Standardized Pearson Residual giving a calculation of the intensity of the difference between observed and expected values centered on Table 2. For examples, Observed frequency of Uploading content for 19 years old is 3 (9%), and expected frequency is \((9 \times 13)/32 = 3.65625\). Therefore, the Standardized Residual is -0.53. It means the actual data respondents with 19 years old who pick uploading content become less from what predicted. Meanwhile, the two attributes have high positive residue; then, the density of shading is high. Therefore, in Figure 3 that the high frequency of aged over 20 years upload an edited video (2.36), and age 18 had a chance to posting and information searching (1.75).

![Figure 4. Decision Tree of Instagram Usage Habit](image)

![Figure 5. Sieve Diagram For The Duration Of Using Instagram](image)

We were exploring the Decision Tree in Figure 4, especially the duration of using Instagram attributes. A dataset in the decision tree that appears became overlap. A dataset split by the duration of using Instagram: an average of 18.6 is 15 to 30 minutes, 30 to 60 minutes, or 5 to 15 minutes. In the sieve diagram, 18.6 approaching 19 years of age, which means datasets have positive residue, they are 15 to 30 minutes, 30 to 60 minutes, and 5 to 15 minutes. However, dataset split again with the same attributes, when an average of 18.0: 5 to 15 minutes, 18.6: 15 to 30 minutes or 30 to 60 minutes. In the sieve did not change, excepted 30 to 60 minutes for 18.0 that caused it had null value (observed frequency is equal to expected frequency), and >60 minutes.

Therefore, it makes a dataset cannot identify. In Figure 5, the duration of using Instagram for 19 years old is 5 to 60 minutes, and the other one using Instagram that the time was inconsistent, such as respondents who are 18 years old that using Instagram in 5 to 15 minutes, 15 to 30 minutes, or >60 minutes. Twenty years old respondents using in < 5 minutes or >60 minutes. >=20 years old respondents using in < 5 minutes or 30 to 60 minutes.

The mistakes can occur when the design of questioner is not correctly, especially the characteristics habit using Instagram based on age. Several attributes, when associating it with age, is a similar pattern (shown in Figure 2). We suggest that the questions for respondents need to be improved, such as Tiered questions. Therefore, The classification of decision trees would become more precise in the subsequent studies.

4. Conclusion

Based on exploring data in Sieve Diagram, several attributes cannot identify because they had inconsistent data. In the Duration Of Using Instagram attribute's sample, based on the age of 18, 20, and >=20 using Instagram at a random time. We deduced that several attributes for the habit of using
Instagram are not unique characteristics because the age-based attributes had a similar pattern. We suggested that questions need improving for the respondents, such as the design of Tiered questions. Thus, the classification of decision trees in the subsequent studies would become more precise.

References

[1] Y. L. Chen, H. W. Hu, and K. Tang, "Constructing a decision tree from data with hierarchical class labels," Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 4838–4847, Apr. 2009.
[2] E. W. Ambarsari, A. Ar Rakhman Awaludin, A. Suryana, P. M. Hartuti, and R. Rahim, "BASIC CONCEPT PYTHAGORAS TREE FOR CONSTRUCT DATA VISUALIZATION ON DECISION TREE LEARNING," J. Appl. Eng. Sci., vol. 17, no. 4, pp. 468–472, 2019.
[3] A. S. Fitrani, M. A. Rosidi, Y. Findawati, Y. Rahmawati, and A. K. Anam, "Implementation of ID3 algorithm classification using web-based weka," in Journal of Physics: Conference Series, 2019, vol. 1381, no. 1, p. 012036.
[4] E. Indra, K. Ho, Arlinandra, R. Hakim, D. Sitanggang, and O. Sihombing, "Application of C4.5 Algorithm for Cattle Disease Classification," in Journal of Physics: Conference Series, 2019, vol. 1230, no. 1, p. 012070.
[5] S. PANG and J. GONG, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," Syst. Eng. - Theory Pract., vol. 29, no. 12, pp. 94–104, Dec. 2009.
[6] E. W. Ambarsari and H. Herlinda, "Pythagoras Tree Applied For Determined Instagram Usage Habit Decision," SinkrOn, vol. 4, no. 2, pp. 56–61, 2020.
[7] E. Windia Ambarzari, R. Avrizal, E. Doni Sirait, S. Dwiasnati, and R. Rahim, "Regression Tree Role for Interpret Monetizing of Game Live Streaming," J. Phys. Conf. Ser., vol. 1424, p. 012014, Dec. 2019.
[8] R. A. Rensink, "The nature of correlation perception in scatterplots," Psychon. Bull. Rev., vol. 24, no. 3, pp. 776–797, Jun. 2017.
[9] M. Friendly, "Conceptual Models for Visualizing Contingency Table Data," in Visualization of Categorical Data, Elsevier, 1998, pp. 17–I.
[10] P. M. E. Altham, "Contingency tables," in Wiley StatsRef: Statistics Reference Online, Chichester, UK: John Wiley & Sons, Ltd, 2014.
[11] M. Kateri, "Analysis of Two-way Tables," in Contingency Table Analysis, New York, NY: Springer New York, 2014, pp. 17–61.
[12] N. Pandis, "The chi-square test," Am. J. Orthod. Dentofac. Orthop., vol. 150, no. 5, pp. 898–899, Nov. 2016.
[13] A. Agresti, "Introduction: Distributions and Inference for Categorical Data," in Categorical Data Analysis, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003, pp. 1–35.

Acknowledgments

Allow us to thanks Mrs. Herlinda as our co-worker at Universitas Indraprasta PGRI to her permit in observed the data for our study.