Tree Algorithm Model on Size Classification Data Mining

Agis Abhi Rafdhili¹, Eddy Soeryanto Soegoto², Senny Luckyardi³, Chepi Nur Albar⁴
¹Department of Information System, Faculty of Engineering and Computer Science, Universitas Komputer Indonesia
²,³Department of Management, Faculty of Economic and Business, Universitas Komputer Indonesia
⁴Department of English, Faculty of Humanities, Universitas Komputer Indonesia
¹agis@email.unikom.ac.id, ²eddy.soeryantos@email.unikom.ac.id, ³chepi.nuralbar@email.unikom.ac.id

Abstract

The goal of this research is to use a tree algorithm to categorize student clothing in order to acquire an accurate size. This research is qualitative approach through descriptive analysis, while the analysis employed C.45 Tree algorithm classification. Manual calculations utilizing the tree algorithm formula revealed that the majority of students require XL-sized clothing. On the X5 (Shoulder length) characteristic, the maximum entropy and information gain values were obtained at 0.212642462. According to the forecast, the shoulder length attribute is the first calculation in developing a decision tree scheme since it has the largest entropy and information gain value. Lastly, the findings of this study analysis can be used as a mapping prediction to make decisions on the size of the student group’s clothing.

Keywords: classification; data mining; decision tree

1. Introduction

Data mining is a technique for generating patterns from a set of processed data in order to perform further analysis and produce reliable prediction results [1]. Data mining techniques can be used for a variety of purposes, including getting decision predictions and data distribution patterns, as well as sorting data into specific groups [2]. Classification is one of the data mining approach strategies used in data processing used to arrange datasets that have been separated into specified groupings [3]. This approach, for example, can categorize car types as Sedan, MPV, and SUV. The same rules can be used to categorize clothing sizes for undergraduate students. The datasets for this investigation were created based on the clothing sizes of 32 undergraduate students. To generate prediction patterns from the analyzed data, the classification stage employs a decision tree method.

Several studies use data mining classification techniques, one of which is in health sector carried out by Llias Tougui, Song, and Elhoseny [4]–[6] in forming decision patterns. The Naïve Bayes algorithm was utilized to classify the dataset and build data patterns in order to make decisions in health instances during the investigation. In the field of education, the algorithm can be used to determine the relationship and impact between a student’s academic achievements and their social abilities. This study assesses any qualities that may influence student’s potential to excel, both individually and as a group [7]. In another research, the Naïve Bayes classification technique was utilized to identify the typical pattern of consumer activity over a specific time period in order to assess the company’s inventory control. In the research, entropy and information gain calculations are employed to detect general customer behavior in order to generate output for the company’s policy-makers [8], [9], the research used the Naïve Bayes classification algorithm to quickly and simply determine the dataset pattern. However, in this study, data processing is carried out using a decision tree algorithm with the goal of conveniently visualizing the results of data processing so that decision making is relatively correct. The goal of this project is to use data mining classification algorithms to assess a dataset of students’ clothing sizes. A decision tree technique with entropy and information gain calculations was employed to process the dataset.

2. Research Methods

This study used descriptive analysis in conjunction with a qualitative method. Furthermore, Classification techniques were applied in the data processing, while the Decision tree algorithm was used in the algorithm approach. The algorithm is typically used to solve a problem by representing the criteria as interconnected
nodes in the form of a tree [10], [11]. The tree, also known as the Decision tree, is a predictive decision-making approach model that is represented in a tree-shaped structural hierarchy [12], [13]. The phases of data processing using a decision tree begin with defining the selected qualities and go to the decision. The gainration calculation was utilized in this study to determine the selected qualities. Formula 1 is the gainration computation formula.

\[
Gainratio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}
\]  
(1)

3. Results and Discussions

3.1 Data Collection

The dataset used in this study used 32 undergraduate students data as a test sample. Table 1 shows the dataset used in this study.

<table>
<thead>
<tr>
<th>Respondent</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>TARGET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VERY WIDE</td>
<td>VERY SHORT</td>
<td>SHORT</td>
<td>MEDIUM</td>
<td>XL</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>VERY WIDE</td>
<td>VERY SHORT</td>
<td>SHORT</td>
<td>MEDIUM</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>WIDE WIDE</td>
<td>LONG TALL</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WIDE WIDE</td>
<td>VERY TALL</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>WIDE WIDE</td>
<td>TALL WIDE</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>WIDE WIDE</td>
<td>TALL WIDE</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>WIDE WIDE</td>
<td>VERY TALL</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>WIDE WIDE</td>
<td>TALL WIDE</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>WIDE NARROW SHORT</td>
<td>SHORT</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>SMALL NARROW SHORT</td>
<td>SHORT</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>WIDE WIDE</td>
<td>LONG TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>VERY WIDE</td>
<td>LONG TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>VERY WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>WIDE NARROW SHORT</td>
<td>LOW</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>WIDE NARROW SHORT</td>
<td>SHORT</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>NARROW WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM XL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>WIDE WIDE</td>
<td>SHORT TALL</td>
<td>MEDIUM L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>WIDE WIDE</td>
<td>VERY TALL</td>
<td>WIDE XL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In Table 1, there are 5 main attributes from X1 to X5. X1 means Bust, X2 is Stomach Circumference, X3 is Arm Length, X4 is Body Length, and X5 means Shoulder Length. On the other hand, the Target Attribute is a label attribute that will be a decision prediction.

3.2 Data Analysis

The first step in analyzing the obtained dataset is to determine the value of entropy and information gain. Entropy is an information theory measure used to determine the characteristics of an existing dataset [14], [15]. The entropy value is a critical metric in determining the Information Gain (IG) value for each characteristic. The largest IG value of the current attributes is used to pick characteristics as nodes, both as roots and as branches [16]. The formula for computing the entropy value is shown in Formula 2.

\[
H(X) = - \sum_{i=1}^{n} P(X = i) \log_2 P(X = i) \tag{2}
\]

H is the Set entropy, Xis the Case set, n is the Number of X Partition, and Pi is the Number of samples for i and proportion of undergraduate students on X.

To calculate the entropy value in this dataset, there are 2 classes that will form the entropy value, namely L and XL. From this formula, the value of L class is 0.52439747 and the value of XL class is 0.371640762, resulting in entropy value of H(X) is 0.896038233. The next stage is the calculation of the IG value. Information gain is a measurement of attribute values used to test each node in the tree [17]. Formula 3 is a formula to get the IG value for each tested attribute:

\[
Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot entropy S_i \tag{3}
\]

S is the Case set, A is the Attributes, n is the Number of partitions on Attribute A, |Si| is the Number of cases on i-th partition, and |S| is the Number of cases on S.

The largest information gain is observed in the X5 attribute with a value of 0.212642462, hence the tree algorithm test in Figure 1 is performed on the X5 attribute with Medium and Wide instruments. The results of the tree algorithm test on the X5 Medium and X5 Wide attributes are shown below (see Figures 2 and 3).

In Figure 2, the X5 Medium characteristic was evaluated against other qualities to acquire the IG value so that further nodes for this decision tree technique could be obtained. According to the test findings, the X1 characteristic had the highest IG value, which is 0.519358614. Based on the value, the following node from the X5 Medium property, namely the X4 attribute, will be evaluated (see Figure 4-6). While testing the X5 Wide attribute in Figure 3, the results produced are homogenous in XL size on all other qualities. As a result, the X5 Wide attribute indicates that the decision obtained is the XL size.

![Figure 1. Tree Algorithm Examination on All Attributes](image-url)
In Figure 4, the X1 Wide attribute was evaluated on the X2, X3, and X4 attributes, with the X4 attribute yielding the highest IG of 0.170080659. As a result, the X4 attribute is a continuation node of the X1 Wide attribute (see Figures 7-9). While the X1 Very Wide attribute was evaluated on the X2, X3, and X4 attributes, the greatest value was recorded at the X3 attribute with an IG value of 0.789681414. However, the attribute value
was considered to be homogeneous at the XL size, hence the decision of attribute X1 Very Wide was XL size. Furthermore, the maximum value on the IG of X1 Small attribute is on the X3 attribute of 0.896038233, which is the same as the entropy value, implying that there is no option for the X1 Small attribute.

**Figure 5. Information Gain of Attribute X1 Very Wide**

**Figure 6. Information Gain of Attribute X1 Small**

**Figure 7. Information Gain of Attribute X4 Short**

In Figure 7, the X4 Short attribute was combined with the X2 and X3 attributes to find the next node; the X3 attribute had the maximum information gain with a value of 0.734213133. With these findings, the X4 Short attribute node is being tested on the X3 attribute. Furthermore, in Figure 8, the X4 Tall attribute test was performed, and the largest information gain on the X2 attribute was reached with a value of 0.470565953, indicating that the next node for the X4 Tall attribute is tested for the X2 attribute. Finally, a test of the X4 Very Tall attribute in Figure 9 reveals that the IG is the same for both traits and is homogeneous at the XL size. It can be concluded that the X4 Very Tall decision tree is XL size. The next test is the node of the X4 Short attribute shown in Figure 10 to 12.
In Figure 10, testing the X3 Very Long attribute is carried out on the X2 attribute because the remaining attributes that have not been tested are only the X2 attribute. The results of the X3 Very Long IG test resulted in all homogeneous values in the XL parameter and the information gain value was the same as the entropy. Therefore, the decision tree of the X3 Very Long attribute is the size of XL. Furthermore, based on Figure 11 test of the X3 Long attribute, it is found that the X2 attribute is still diverse, so it is necessary to carry out further testing on the X2 attribute. Figure 12 shows the same case as X2 attribute, so further testing needs to be done (see Figure 13-15).

In Figure 12, the X2 Very Fat attribute test was carried out on the X3 attribute and obtained homogeneous results in the XL size. While in Figure 13, all homogeneous attribute of X2 Fat tested on the XL size. Likewise, the homogeneous X2 Flat attribute was tested at size L. It can be concluded that for testing the X2 attribute, the decision tree obtained for the X2 Very fat attribute is XL, the X2 Fat and X2 Flat attributes are L size. Figure 15 is the result of the decision tree, based from all the tests that have been carried out on all attributes and obtained the following decision tree.

DOI: https://doi.org/10.29207/resti.v7i4.4572
Creative Commons Attribution 4.0 International License (CC BY 4.0)
From the results of the decision tree obtained in Figure 16, it can be described. The results of the decision tree algorithm were processed manually using theoretical formulas.

1. IF X5="MEDIUM" AND X1="SMALL" THEN Ukuran = "XL"
2. IF X5="WIDE" THEN Ukuran = "XL"
3. IF X5="MEDIUM" AND X1="WIDE" AND X4="VERY TALL" THEN Ukuran = "XL"
4. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="VERY LONG" THEN Ukuran = "XL"
5. IF X5="MEDIUM" AND X1="WIDE" AND X4="TALL" AND X2="VERY FAT" THEN Ukuran = "XL"
6. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="LONG" AND X2="VERY FAT" THEN Ukuran = "XL"
7. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="LONG" AND X2="FAT" THEN Ukuran ="L"

8. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="LONG" AND X2="FLAT" THEN Ukuran ="L"

9. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="SHORT" AND X2="VERY FAT" THEN Ukuran ="XL"

10. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="SHORT" AND X2="FAT" THEN Ukuran ="L"

11. IF X5="MEDIUM" AND X1="WIDE" AND X4="SHORT" AND X3="SHORT" AND X2="FLAT" THEN Ukuran ="L"

12. IF X5="MEDIUM" AND X1="VERY WIDE" THEN Ukuran ="XL"

In future research, it can be done using an application that can generate and display more accurate results. This manual test aims to explain the flow of calculations in the prediction of decision tree decision making [18]–[20]. Data mining techniques, such as the tree algorithm formula, can be used to evaluate business activities, such as mapping the most popular products and examining the relationships between those products and other products. Another example is that data mining can be used to evaluate the market in order to discover the link between products purchased by customers in the transaction database. The output of data mining assists the owner in developing the marketing strategy and procurement system [21]. Furthermore, data mining can be coupled with other technologies such as 3D Virtual Reality to provide accurate data and lifelike visuals. For example, it was discovered in the examination of clothing comfort from the numerical garment viewpoint that different regions of the human body affect the amount of comfort differently. The study aided in the development of a pattern for measuring the ease of quantitative calculations for each data sample [22].

The tree algorithm model works by recursively partitioning the feature space into subsets based on the values of the features. This is done by constructing a tree-like structure where each node represents a subset of the feature space and each branch represents a decision based on a feature value. At each node, the algorithm selects the feature that provides the most significant information gain and splits the data based on its values. The process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples in a leaf node.

To apply the tree algorithm model for size classification, the features used can be physical characteristics of the object, such as length, width, and height. These features can be used to train the model using a dataset of objects with known sizes. Once the model is trained, it can be used to predict the size of new objects based on their features.

Overall, the tree algorithm model is a powerful and flexible tool for size classification, and it has been successfully applied in many fields, such as agriculture, industry, and medical imaging. However, like all machine learning models, it requires careful data preprocessing and validation to ensure accurate and reliable results.

4. Conclusion

Based on the results of the Decision tree test from the undergraduate student data, it was observed that the majority of predictions resulted on the XL size with 8 out of 12 decisions. Its entropy value of 0.896038233. On the Information Gain, testing begins with the X5 attribute after attaining the highest Information Gain when testing all attributes, which is 0.212642462. Meanwhile, the next node is tested on the X1 attribute with a value of 0.519358614 for Information Gain. After testing on the X1 Wide instrument, the third node is tested on the X4 attribute with an Information Gain value of 0.170080659. Furthermore, after receiving an Information Gain of 0.73421333 on the X4 Short attribute, the X3 attribute became the second last test. Finally, the X2 attribute is the final test that leads to the decision.

References


DOI: https://doi.org/10.29207/resti.v7i4.4572

Creative Commons Attribution 4.0 International License (CC BY 4.0)


