International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 1, No. 1, September 2018, pp. 24-27

Effect of Selection of Classification Features C4.5 Algorithm in Student Alcohol Consumption Dataset

Tri Astuti [a,1,*], Pungky Dwi Putra Handoko [a,2]

[a] Informatics Engineering, Universitas AMIKOM Purwokerto
[1] tri_astuti@amikompurwokerto.ac.id*; [2] pungkycompong@gmail.com
* corresponding author

Abstract

Alcoholic beverages are psychoactive substances that are addictive. Psychoactive substances are a class of substances that work selectively, especially in the brain, which can cause changes in behavior, emotion, cognition, perception and awareness of one's and others. Police survey results in 2014 showed that users of narcotics and liquor. Most of the group of students, both junior, and senior student, which amounts to 70%, while only 20% of primary school graduates. In the modern era, especially in information technology, the need for information and the latest knowledge is multiplying. One of them is the user information of alcohol among teenagers is more accurate. Data mining is the process for extracting and identifying information useful and relevant knowledge from a variety of big data. In the data mining, there is a classification technique that assesses the data objects to include it in a particular class of several classes available, can be applied in the case - the case in the health sector, for example, in the case of alcohol addiction in adolescents. The algorithm that can be used in the classification is the C4.5 decision tree. The use of the decision tree algorithm to determine the level of alcohol use in teenagers using two methods, namely, the selection of attributes and without attributes.

Keywords: Alcohol; Youth; Data mining; Classification; Decision tree.

1. Introduction

Alcoholic beverages are psychoactive substances that have been used in many cultures for centuries and resulted in dependency properties. Harmful use of alcohol causes chronic illness, the social and economic burden on society [1]. Deviant behavior of adolescents to liquor a sight that has been commonly encountered and the rest is as already entrenched in every circle.

In western countries, for example, 90% of the population never drank alcohol, and 60-70% of them became drinkers remained until today [2][3]. Moreover, of the entire population in the world is noted that 40% experienced a temporary problem that consists of 20% is the abuse of alcohol and 20% longer experiencing alcohol dependence [4].

In Indonesia, drinking behavior in adolescents is a very complex problem and should be addressed immediately. This is due to the distribution of alcohol did not know the gender, age, class, religion, and economic status. According to the survey, the number of teenagers who consume alcohol reached 4.9%. The prevalence of alcohol 12 months and 1-month high start between the ages of 15-24 years is 5.5% and 3.5% increase to 6.7% and 4.3% in men aged 25-34 years but then down with age [5].

In the modern era, especially in information technology, the need for information and the latest knowledge is multiplying. Information and new knowledge can be found in the collection of large amounts of data types. To acquire knowledge of the pile of data, techniques that can be used is data mining. Data mining is the process that uses statistical techniques, mathematics, artificial intelligence, and machine learning for extracting and identifying information useful and relevant knowledge from a broad data range [6].

Based on the journal and the description above, the previous research in this study combined with the feature selection method for the analysis of datasets C4.5 algorithms of the consumption of alcohol in teenagers.

2. Research Methods

2.1. Decision Tree C4.5

According to [7] decision tree is a structure that can be used to divide the large data sets into the set - the set of records that more ketch by applying a set of decision rules. With each circuit division, members of the result set to be similar to one another.

Larose in [8] says many algorithms that can be used in the formation of the decision tree, such as ID3, CART, C4.5. The algorithm C4.5 is the development of ID3 algorithms.

2.2. Feature Selection

Feature or attribute selection is the process of identifying and removing information that is irrelevant and redundant as much as possible [9][14]. This enables the reduction of data dimensions so machine learning algorithms to work faster and more effectively. In some cases, the classification accuracy can be improved, but for some other cases, the result is more uncomplicated and more comfortable to learn and interpreted[10][11].

2.3. CfsSubseteval (Correlation-based feature selector)

Correlation-based feature selection from now on called correlation-based attribute selection or CFS is a simple filter algorithm that ranks subset based on heuristic evaluation function based on the correlation [12][13]. Based on the hypothesis that a good subset of attributes contains attributes that are strongly correlated to each class and not related to each other. A high correlation with one another attribute indicates the attribute redundant. Attribute the low correlation of the class attributes are irrelevant. Attributes that are not relevant and redundant should be removed[15].

3. Results and Discussion

In this study, conducted some preliminary studies by studying literature relating to the study was to predict the consumption of alcohol in middle school and the selection of the appropriate algorithm for this study. From the results of research conducted by using feature selection methods and algorithms Decision Tree (C4.5) to analyze datasets.

In this study, the data will be used in research that is taking on the repository database. UCI Student Alcohol Consumption is student dataset. This dataset consists of 649 data. Such data can not be used directly to determine points in predicting the level of alcohol consumption among secondary school.

For Decision Tree, algorithm performance measurement without the selection attributes, gets the confusion matrix values as in Figure 1 below.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 141  34 |   a = no
  60 414 |   b = yes
```

Fig. 1 Results of Confusion Matrix for Decision Tree algorithm without attribute selection

From the confusion matrix obtained, next step is to find the value of precision, recall, and certainly value accuracy of the Decision Tree algorithm. From the calculation results obtained by value precision, recall, and accuracy gained by the Decision Tree algorithm, that can be seen in Table 2 below.

Table 2. The results of performance measurement algorithms without the selection attribute Decision Tree

Algorithm	*precision*	*recall*	*Accuracy*
decision Tree (No Selection Attributes)	70.1%	80.6%	85.5%

Performance measurement algorithms for Decision Tree with a selection of attributes in getting the confusion matrix values as in Figure 2 below.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 172   3 |   a = no
  67 407 |   b = yes
```

Fig. 2 Results of Confusion Matrix for Decision Tree algorithm with attribute selection

From the confusion matrix obtained next step is to find the value of precision, recall, and certainly value accuracy of the algorithm Decision Tree.

From the calculation results obtained by value precision, recall and accuracy gained by the Decision Tree algorithm shown in Table 3.

Table 3. The results of performance measurement algorithms with the selection attribute Decision Tree

Algorithm	*precision*	*recall*	*Accuracy*
decision Tree (Selection Attributes)	72%	98.3%	89.2%

From the calculations have been done on the Decision Tree algorithm with the two methods, the results obtained an accuracy of each - each algorithm that is 85.5162% with the precision and recalls 0701% 0806% on Decision Tree algorithm without using feature selection. While the accuracy of the Decision Tree algorithm using feature selection in the amount of 89.2142% with the precision and recall 0720% 0983%. Here in Figure 3 is a graph of the results of the calculation algorithm accuracy Decision Tree with a selection of attributes and without attributes the selection below.

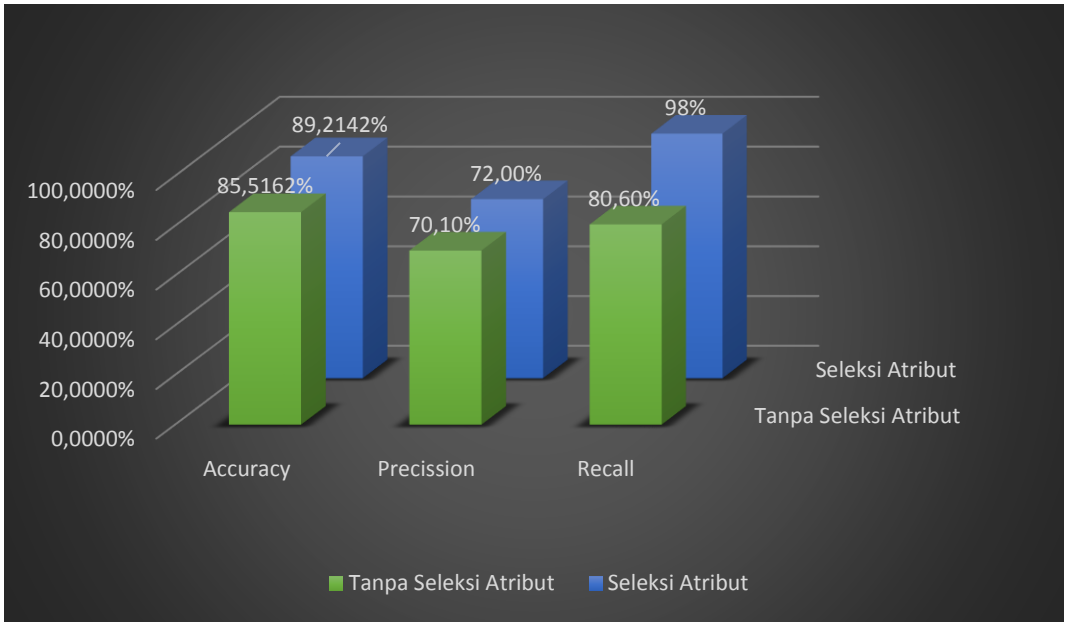

Fig 3. Comparison of Accuracy Results Decision Tree with a selection of attributes and no attribute selection

4. Conclusions and Suggestions

4.1. Conclusions

From research that have been done then obtained some conclusions are:

a. Decision tree algorithm with the selection attribute has a more accurate result than the decision tree algorithm without the selection of attributes, the results of research that has been done indicates the level of accuracy of decision tree algorithm with attribute selection by 89.2% while the decision tree algorithm without selection only attribute by 85, 5%.

b. The results of calculations using a decision tree algorithm with a selection of attributes or no attribute selection are entirely accurate with the percentage of accuracy above 80% so that it can be used after the classification.

4.2. Suggestions

There is some suggestion that the author convey to the development of further research, including:

a. The addition of attributes used to determine whether the extent of the attributes affects the accuracy of the results of the decision tree algorithm.

b. Need comparisons decision tree algorithm with other algorithms to determine whether the decision tree algorithm has a pretty good result compared with other algorithms in the completion of the classification.

References

[1] WHO, "Global Status Report On Alcohol And Health". 2014.

[2] D.T. Larose, "Discovering Knowledgr in Data: An Intoduction to Data Mining". Hoboken : Wiley – Interscience, John Wiley & Sons, Inc, 2005.

[3] M.A. Hall, "Correlation-based Feature Selection for Machine Learning", Hamilton, The University of Waikato. NewZealand, 1999.

[4] Abellán,J., &Moral,S. (2006). An algorithm to compute the upper entropy for order-2 capacieties. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems,14(02), 141–154.doi:10.1142/S0218488506003911.

[5] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.,Ventura, S., Garrell, J., ... Herrera,F. (2009). Keel: a software tool to assess evolutionary algorithms for dataminingproblems.SoftComputing,13(3), 307–318.doi:10.1007/s00500-008-0323-y.

[6] Bernard,J. M.(2005). Anintroduction to the imprecise dirichletmodelfor multinomialdata.International Journal ofApproximateReasoning,39(2–3), 123–150.

[7] doi:10.1016/j.ijar.2004.10.002.Imprecise ProbabilitiesandTheirApplications

[8] Demšar,J. (2006).StatisticalcomparisonsofclassifiersovermultipledatasetsJournalofMachineLearningResearch,7, 1–30.

[9] Elouedi, Z., Mellouli, K., &Smets, P. (2001). Belief decision trees: theoretical foundations. International Journal of Approximate Reasoning, (28), 91–124.

[10] Frenay, B.,& Verleysen, M.(2014). Classificationin the presence of label noise:A survey. Neural Networks and Learning Systems, IEEE Transactions on,25(5), 845–869.doi:10.1109/TNNLS.2013.2292894.

[11] Klir,G.J. (2005). Uncertainty and information: Foundations of generalized information theory. John WileyAndSons,Inc. doi:10.1002/0471755575.

[12] Konˇ car,N.(1997). Optimisationmethodologiesfor directinverse neurocontrol. DepartmentofComputing,Imperial CollegeofScience,TechnologyandMedicine,UniversityofLondonDoctoraldissertation.

[13] Mantas,C.J., &Abellán,J. (2014a). Analysisandextensionofdecisiontrees based onimprecise probabilities:Applicationonnoisydata.ExpertSystemswithApplications,41(5), 2514–2525.doi:10.1016/j.eswa.2013.09.050.

[14] Mantas,C.J., &Abellán,J. (2014b). Credal-c4.5:Decisiontree based on impreciseprobabilitiesto classifynoisydata.ExpertSystemswithApplications,41(10),4625–4637.doi:10.1016/j.eswa.2014.01.017.

[15] Wang,Y.(2010). Imprecise probabilitiesbasedongeneralizedintervals for systemreliability assessment.International Journal ofReliabilityandSafety, 4(4), 319–342.doi:10.1504/IJRS.2010.035572