Automatic Group Labeling with Decision Trees: A Comparative Approach

Manoel Messias P. Medeiros[1], Vinicius P. Machado[2], Daniel de S. Luz[1], Rodrigo de Melo S. Veras[2]

[1]Federal Institute of Education, Science and Technology of Piauí (IFPI)
Av Pedro Marques de Medeiros, s/n - Parque Industrial, Picos - PI, 64605-500

[2]Department of Computer Science – Federal University of Piauí
Teresina-PI.

{mmessias,daniel.luz}@ifpi.edu.br, {vinicus,rveras}@ufpi.edu.br

Abstract. *The exponential growth in data volume demands efficient data analysis techniques, with data clustering being crucial but interpretation often posing a challenge. Automated group labeling using decision trees can alleviate this issue. This study compares four decision tree algorithms for automated group labeling, demonstrating that algorithm choice significantly influences performance. CHAID outperforms other algorithms in the Iris and Seeds datasets, while C4.5 excels in the Wine and Glass datasets. The proposed model's validity is confirmed, highlighting the importance of careful algorithm selection. These findings underscore the potential of automated group labeling models and emphasize the need for further research to refine and expand their applications across various domains.*

1. Introduction

The exponential growth of data generated by diverse sources, including sensor networks, commercial transactions, and social networks, is fueling the rapid advancement of data analysis. This growth has become particularly pronounced in recent years, driven by the rapid expansion of the technology industry and the increasing applicability of computational techniques, coupled with advancements in analytical tools. [Dimotikalis et al. 2021]. Consequently, the application of clustering algorithms followed by expert analysis is gaining prominence, attracting the attention of numerous studies in the field of Unsupervised Learning, a subdiscipline of Machine Learning (ML). The unsupervised formation of clusters is framed as a problem of classifying objects into distinct categories without predetermined category labels [Russell and Norvig 2016]. Clustering divides a set of data into smaller subsets, called clusters, which bring together objects with common characteristics [Lopes et al. 2016]. The labeling problem is defined as a summary identification for groups, naming them according to their characteristics [Lopes et al. 2016]. The group labeling process aims to uniquely identify a group through the tuple attribute and value range. Several techniques were proposed [Lopes et al. 2014], [Machado et al. 2015], [Filho et al. 2020], [Moura et al. 2022], [Silva et al. 2021]. This work presents a new approach using decision tree.

2. Related Work

The concept of group labeling was initially introduced in [Lopes et al. 2013]. In this work, a database is provided as input to an unsupervised learning algorithm, resulting in

the generation of clusters from the original elements. Subsequently, a supervised learning algorithm is applied to each previously formed cluster to select the most pertinent attributes. If the input database contains continuous values, a discretization method is employed prior to utilizing the supervised learning algorithm. The model comprises four steps. The Automatic Labeling Model (ALM) proposed by the authors achieved label generation agreement rates exceeding 90%. The ALM method was also successfully applied to other problems [Lopes et al. 2013], [Lopes et al. 2014], [de Lima et al. 2015], [Lopes et al. 2016], demonstrating promising results.

The study present in [Machado et al. 2015] introduces a novel group labeling model that utilizes the Fuzzy C-Means algorithm to assign membership degrees to data elements. A parameter-based approach is employed to refine labels and prevent overlap. The model achieves an average agreement rate of 96.61%, outperforming a previous method.

To address the reliance on the Fuzzy C-Means Algorithm, the approach present in [Filho et al. 2020] proposed a new method for group labeling that addresses the limitations of the Fuzzy C-Means Algorithm. The proposed method utilizes a distance-based algorithm, K-Means, to cluster data and then assigns labels to each cluster based on attribute values. The results of the proposed method were compared to a previous method and found to be comparable, with a slight difference of 0.36% for the Iris dataset.

The CAIBAL model, introduced in [Moura et al. 2022], employs the CAIM discretization algorithm to address information loss in the discretization process. The model offers two methods for selecting value ranges: a standard method based on accuracy and an alternative method using the V parameter. The CAIBAL model achieved high agreement rates (98.49% and 97.33%) using these methods in three of five databases.

In the work of [Silva et al. 2021], the labeling model is structured into two phases: Phase I defines attribute-range pairs, utilizing regression techniques to minimize prediction errors within the domain of each group. Phase II selects pairs that most effectively differentiate and represent each group, forming distinct labels based on their predominant representation among elements.

3. The model

This study proposes a model for automated group labeling utilizing a decision tree. To identify the optimal decision tree algorithm, a comparative analysis was conducted among ID3, C4.5, CART, and CHAID algorithms to determine which generates superior labels. The rationale for selecting these algorithms is outlined below. When developing machine learning models, it is essential to consider the interplay between accuracy and interpretability. A growing number of critical domains, such as group labeling, emphasize the importance of understanding model outputs as much as their accuracy [Di Teodoro et al. 2024].

Decision tree ensemble models, including Random Forests and Boosted Trees, are widely employed in machine learning, particularly for prediction tasks. Their exceptional predictive performance makes them among the most recommended approaches for real-world problems. However, as noted in [Hara and Hayashi 2016], the primary limitation of tree ensemble models lies in their interpretability. They partition the input space into

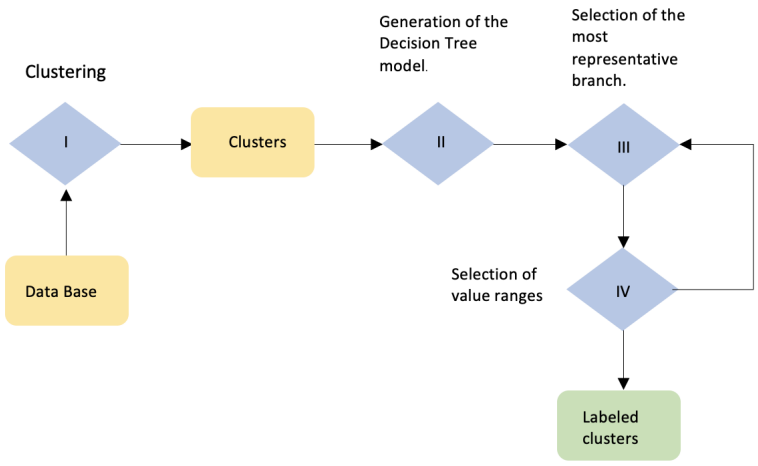

Figure 1. Flowchart of the proposed model.

numerous small regions and make predictions based on the corresponding region. Typically, the number of generated regions exceeds a thousand, implying thousands of distinct prediction rules, which can be challenging for non-experts to comprehend. In contrast, a simple decision tree is renowned for its high interpretability. Despite its relatively lower predictive capability, the number of regions generated by a single tree is significantly smaller, rendering the model transparent and understandable. These models divide the input space into a limited number of regions and make predictions based on the assigned region.

The proposed model analyzes the rules generated by decision trees, suggesting that more interpretable models are preferable for this application.

Given the group labeling problem outlined in [Lopes et al. 2013], the current model introduces a method for labeling clusters utilizing decision trees. Tree-based supervised learning techniques are well-suited for tasks prioritizing interpretability. The feature splits and decision paths of decision trees offer valuable insights into the distinguishing characteristics among members within each cluster [Bertsimas et al. 2020].

To compare the performance of each algorithm, a model was developed consisting of four phases: (I) employing an unsupervised clustering algorithm (k-means), groups are generated from the input database, (II) with the grouped dataset, a decision tree model is constructed, treating the generated cluster as a class attribute. At this stage, one of the four decision tree algorithms, ID3, C4.5, CART, or CHAID, is utilized, (III) the branches of the generated tree with the highest number of hits for each cluster used as a class are selected, and finally, (IV) within each selected branch, the attributes are listed, and their corresponding value ranges are calculated. The phases are illustrated in Figure 1

To illustrate the labeling process, the Iris dataset, initially introduced by [Fisher 1936], will be employed.

3.1. Phase I - Generating clusters

The Iris database clustering process was carried out using the K-Means algorithm, defining a number for k(total cluster number) equal to 3.

3.2. Phase II - Generation of the decision tree model

In the second phase, the decision tree model is generated using the group number as a class attribute. To generate the trees shown in Figures 2, 3, 4 and 5 the default values for each respective algorithm were used. (ID3, C4.5, CART and CHAID) from the library *ChefBoost* [1] , which implements each of these four algorithms [Serengil 2021].

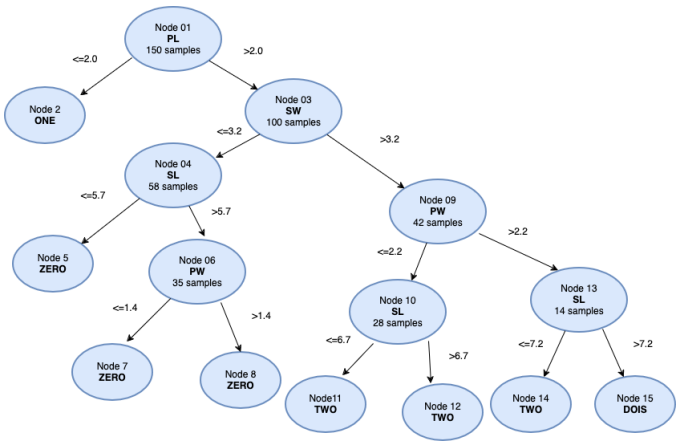

Figure 2. Tree induced with the Iris base from its clusters with the ID3 algorithm

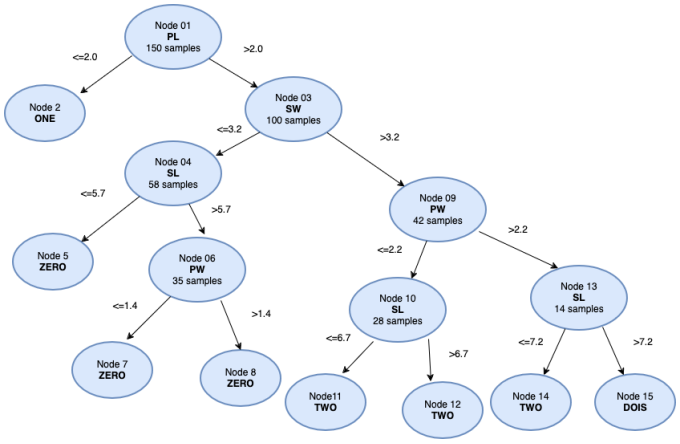

Figure 3. Tree induced from the Iris base from its clusters with the C4.5 algorithm

3.3. Phase III - Selection of branches with the best rating

In this phase, the binary tree generated in the previous phase is analyzed to identify the subtrees that yield the most favorable classification results for each cluster, as the clusters generated in step I were utilized as classes for inducing the classification tree. Consequently, at the conclusion of phase III, we have the set of branches that will be employed in phase IV. Each branch selected for each cluster is presented below.

For the sake of clarity, this demonstration will focus solely on the selection of branches from the tree generated by the CHAID algorithm, which, as illustrated in Tables 1, 2, 3, and 4, achieved the highest agreement rates compared to the other algorithms.

[1] available at https://github.com/serengil/chefboost

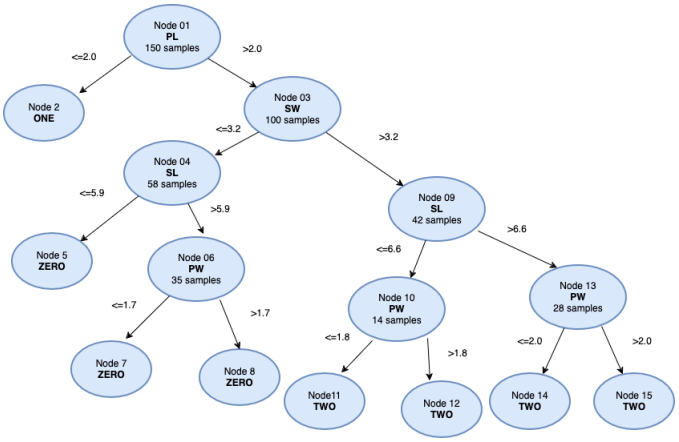

Figure 4. Tree induced from the Iris base from its clusters with the CART algorithm

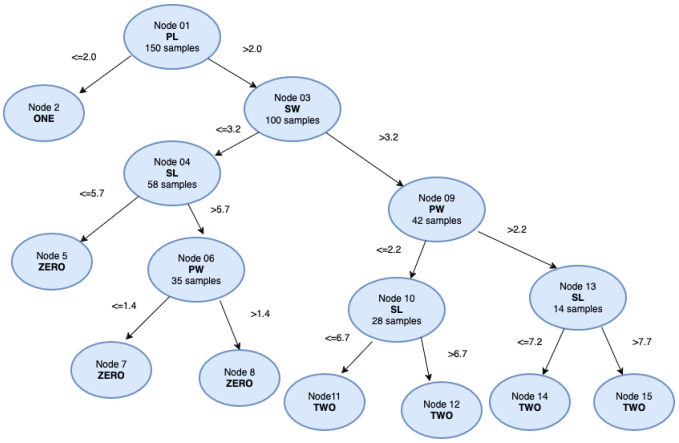

Figure 5. Tree induced from the Iris base from its clusters with the CHAID algorithm

For cluster 0, the branch composed of nodes (1,3,4, 6, and 7) of the tree was selected, as depicted in the subtree in Figure 6. In this instance, there were two branches that could lead to the classification of group 0, and the one that resulted in the highest number of correct classifications was chosen, specifically, the branch with the largest number of samples in the leaf node.

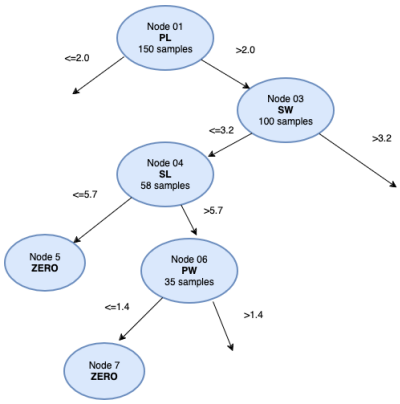

Figure 6. Subtree selected for cluster zero.

For cluster 1 (one), nodes (1 and 2) were selected. In this case, it is the only path that leads to the leaf node of cluster one. This branch is demonstrated in Figure 7.

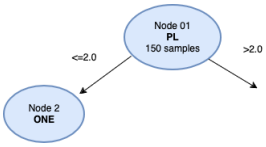

Figure 7. Subtree selected for cluster one.

For cluster 2 (two), the selected branch was composed of nodes (1, 3, 9, 10 and 11) of the decision tree. In this case, it was also necessary to choose the path that leads to the highest number of hits. This subtree is demonstrated in Figure 8.

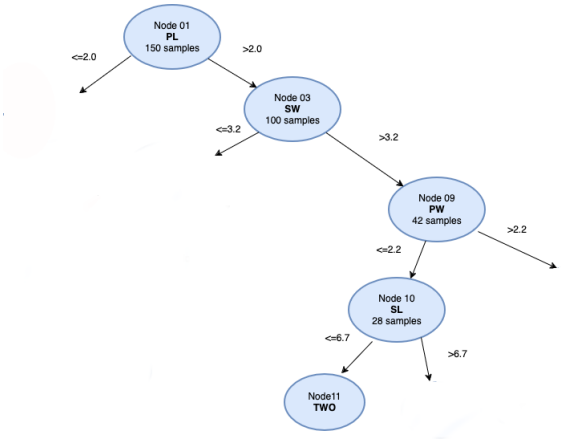

Figure 8. Subtree selected for cluster two.

3.4. Phase IV - Selection of attributes and assignment of value ranges

In this phase, the attributes constituting the subtrees identified in the previous phase are analyzed to determine the subset of attributes that will comprise the label for each cluster. Each algorithm employs a specific metric to identify decision points within the decision tree.

The attributes that make up a label are those that belong to the respective branches selected in phase III. Within each decision node generated along the branch, in addition to the attribute, we find the limits that will be used to compose the label value ranges. In this case, the number of bands is only two, because binary trees are always generated from the bases used by the decision tree algorithms used in the study.

The rules for defining the limits are as follows: if the rule node is the child to the left of the parent node, the rule of that node determines the range with the upper limit, and the lower limit will be selected in the database, accordingly. according to the rule; if the child node is on the right, the rule of that node determines the range with the lower limit, and the upper limit will be chosen from the database according to the rule.

3.5. Value range selection process for zero cluster labels.

To determine the value range of the label attributes, the subtree to which the attribute node belongs is identified. For this cluster, the range of values is defined by the attributes: Petal Length, Sepal Width, and Sepal Length. The following criteria are applied: Petal Length: The first value exceeding 2, which is 3.0 in this case, is considered the lower limit. As this node leads to the right branch (node number 3 in the tree), the upper limit is found in the database as 6.3; Sepal Width: The value 3.2 is considered the upper limit, as this node leads to the left branch of the subtree. The lower limit is found in the database as 2.9; Sepal Length: The lower limit is the first value exceeding 5.7, which is 5.8 in the database. The upper limit selected from the database is 7.6.

3.6. Value range selection process for cluster one labels.

For this cluster, Petal Length is the sole attribute. To establish the value range, the following observation is made: the value 1.9 is designated as the upper limit, as the tree branch rule defined for the attribute is ”less than or equal to 2.0.” Since the leaf node is located in the left subtree, the value 1.0 is selected from the database as the lower limit.

3.7. Value range selection process for cluster two labels.

For cluster two, the attributes Petal Length, Sepal Width, and Petal Width constitute the label. The label value ranges are defined as follows: Petal Length: The lower limit is the first value exceeding 2.0, as the next node belongs to the right subtree. The base values 4.5 and 6.9 serve as the lower and upper limits, respectively; Sepal Width: The lower limit is the constant value of node 3, which is greater than 3.2. Since the next node in the branch belongs to the right subtree, the base values 3.3 and 4.4 define the lower and upper limits, respectively; Petal Width: The upper limit is ”less than or equal to 2.2,” and the lower limit is the base value 1.4.

Tables 1, 2, 3, and 4 present an analysis of the generated labels, comparing the performance of each of the four decision tree algorithms on the Iris dataset. The following fields provide information about the labels: Cluster: Indicates the group number to

which the label belongs; Elem: Indicates the total number of elements within that cluster; Attri.: Indicates the attribute being labeled; Value Range: Indicates the range of values that define an attribute label for a given cluster; Errors: Indicates the number of label errors; Agree. Rate (Agreement Rate) %: Indicates the percentage of label hits, which is calculated as the percentage of elements falling within the label's value range relative to the total number of elements in the respective cluster.

Table 1. Labeling analysis for the Iris database with the ID3 algorithm.

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree. Rate%
0	62	PL	3.0 ∼ 6.3	0	100
		SW	2.9 ∼ 3.2	8	87.09
		SL	5.8 ∼ 7.6	23	62.90
1	50	PL	1.0 ∼ 1.9	0	100
2	38	PL	4.5 ∼ 6.9	0	100
		SW	3.1 ∼ 4.4	4	89.47
		PW	1.4 ∼ 2.2	13	54.78

Table 2. Labeling analysis for the Iris database with the C4.5 algorithm.

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree. Rate%
0	62	PL	3.0 ∼ 6.3	0	100
		SW	2.9 ∼ 3.2	8	87,09
		SL	5.4 ∼ 5.9	29	53,22
1	50	PL	1.0 ∼ 1.9	0	100
2	38	PL	4.5 ∼ 6.9	0	100
		SW	3.3 ∼ 4.4	4	89.47
		SL	6.7 ∼ 7.9	12	68,42

Table 3. Labeling analysis for the Iris database with the CART algorithm.

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree. Rate%
0	62	CP	3.0 ∼ 6.3	0	100
		LS	2.9 ∼ 3.2	8	87.09
		CS	5.4 ∼ 5.9	29	53.22
1	50	CP	1.0 ∼ 1.9	0	100
2	38	CP	4.5 ∼ 6.9	0	100
		LS	3.3 ∼ 4.4	4	89.47
		CS	6.7 ∼ 7.9	12	68.42

Table 4. Labeling analysis for the Iris database with the CHAID algorithm.

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree. Rate%
0	62	PL	3.0 ∼ 6.3	0	100
		SW	2.9 ∼ 3.2	8	87,09
		SL	5.8 ∼ 7.6	23	62,90
1	50	PL	1.0 ∼ 1.9	0	100
2	38	PL	4.5 ∼ 6.9	0	100
		SW	3.3 ∼ 4.4	4	89.47
		PW	1.4 ∼ 2.2	13	65.78

Based on the average agreement rates presented in Tables 1, 2, 3, and 4, the CHAID algorithm exhibited the highest performance with an average agreement rate of 86.46%. Consequently, this algorithm was selected for phase II of the proposed model. However, this result falls short of the average agreement rate achieved by the model demonstrated in [Lopes et al. 2016], which was 95.43%.

Table 5. Labeling analysis for Iris database present in [Lopes et al. 2016].

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree. rate%
1	50	PW	0.1 ∼ 1	0	100
		PL	1 ∼ 3.7	0	100
2	62	PL	3.7 ∼ 5.1	6	90.32
3	38	PL	5.1 ∼ 6.9	3	92.10
		PW	1.7 ∼ 2.5	2	94.73

4. Results

This section presents the results of applying the proposed data grouping model to three datasets: Seeds, Glass, and Wines, obtained from the UCI repository [Dua and Graff 2017]. These datasets were selected due to their frequent use in related research, enabling a comparison of the results achieved here with those reported in the literature. The results for the Iris dataset were previously presented in the model section.

4.1. Seeds Database

The CHAID algorithm-based model demonstrated superior performance, achieving an average agreement rate of 92.23% as can be seen in table 6. This result is marginally lower than the 92.65% reported in [Lopes et al. 2013], as illustrated in Table 7.

Table 6 presents the results of the model using the CHAID algorithms.

Table 6. Labeling analysis for Seeds database using the CHAID algorithm

Cluster	# Elem	Label			
		Attrib	Range of values	# Erros	#Agree. Rate. %
0	72	A	12.0 ∼ 16.6	0	100
		SL	3.1 ∼ 15.4	7	90,27
1	61	A	17.9 ∼21.1	13	78.68
2	77	A	10.5 ∼14.2	0	100
		SL	2.6 ∼3.0	1	92.20

4.2. Wine Database

Table 8 presents the analysis of the labels generated by the C4.5 algorithm, as a decision tree algorithm of the proposed model. The decision tree algorithm that performed best using the proposed model was C4.5, with an average agreement rate of 99.67%, being only slightly below (0.33%) the result achieved on the same basis by the model presented [Silva et al. 2021] as shown in Table 9.

4.3. Glasses Database

Comparing the results achieved by the four decision tree algorithms, as shown in tables 10, it can be seen that the best result was achieved using the C4.5 algorithm as a supervised algorithm, with an average agreement rate of 99.28%. This result was above the result [Lopes et al. 2016] which was 95.54% as shown in table 11

Table 7. Labeling analysis for Seeds database present in [Lopes et al. 2016].

Cluster	# Elem	Label		Analysis	
		Attri.	Range of values	# Errors	Agree.Rate %
1	67	A	12.78 ∼16.14	8	88.05
		P	13.73 ∼15.18	9	86.56
2	82	A	10.59 ∼12.78	12	85.36
		P	12.41 ∼13.73	10	87.80
3	77	P	15.18 ∼17.25	0	100
		SW	3.465 ∼ 4.033	3	95.08
		SL	5.826 ∼ 6.675	1	98.36
		A	16.1 ∼ 21.18	0	100

Table 8. Labeling analysis for Wine database using the C4.5 algorithm

Cluster	# Elem	Label			
		Attrib	Range of values	# Errors	#Agree. Rate %
0	62	*Proline*	600 ∼937	0	100
		Alcalinity of ash	14.8 ∼30.0	1	98.38
1	47	*Alcalinity of ash*	17.5 ∼27.0	0	100
2	69	*Proline*	278 ∼590	0	100
		Alcalinity of ash	10.6 ∼25.0	0	100

Table 9. Labeling analysis for the Wines database present in [Silva et al. 2021]

Cluster	# Elem	Label		Análise	
		Attri.	Range of values	# Erros	Agree. Rate %
1	62	*Proline*	600.0 ∼937.0	0	100
2	47	*Proline*	953.5 ∼1680.0	0	100
3	69	*Proline*	278.0 ∼598.0	0	100

Table 10. Labeling analysis for Glass database using the C4.5 algorithm

Cluster	# Elem	Lbel			
		Attrib	Range of values	# Errors	#Agree. Rate. %
0	35	Ca	8.6 ∼11.6	0	100
		Mg	1.8 ∼ 4.4	0	100
		K	0.0 ∼0.7	0	100
		Si	70.2 ∼72.7	0	100
1	124	Ca	7.0 ∼ 9.4	0	100
		Mg	2.71 ∼3.98	0	100
		K	0.06 ∼1.10	0	100
		Na	10.7 ∼14.8	4	96,77
		Ba	0.0 ∼0.15	0	100
2	5	Ca	5.43 ∼ 6.96	0	100
		K	1.46 ∼6.21	0	100
		Si	69.8 ∼72.8	0	100
		Al	1.8 ∼3.5	0	100
3	17	Ca	8.9 ∼12.5	0	100
		K	0.0 ∼0.9	1	94,11
4	26	Ca	6.4 ∼9.9	0	100
		K	0.0 ∼0.14	1	96,15
		RI	1.515 ∼1.526	0	100
5	7	Ca	13.2 ∼16.1	0	100
		K	0.0 ∼0.8	0	100
		Si	69.8 ∼73.2	0	100

5. Conclusions

Group labeling models are essential tools for data specialists, providing concise definitions of key group characteristics [Silva et al. 2021]. This study evaluated an automatic

Table 11. Labeling analysis for Glass database found in [Lopes et al. 2013]

Cluster	# Elem	Label		Análise	
		Attri.	Range of values	# Errors	Agree. Rate %
1	74	Ba	0 ∼0.7875	0	100
		K	0 ∼1.5525	0	100
		Si	72.61 ∼74.01	2	97.29
		Na	12.3925 ∼14.055	3	95.94
2	5	Fe	0 ∼0.1275	0	100
		Ca	5.43 ∼8.12	0	100
3	19	K	0 ∼1.5525	0	100
		Ba	0 ∼0.7875	1	94.73
4	32	K	0 ∼1.5525	0	100
		Ba	0 ∼0.7875	1	96.87
		Ca	8.12 ∼10.81	1	96.87
5	56	Ba	0 ∼0.7875	0	100
		K	0 ∼1.5525	0	100
		Na	12.3925 ∼14.055	2	94.62
		Al	1.0925 ∼1.895	4	92.85
		Mg	3.3675 ∼4.49	6	89.28
6	28	Fe	0 ∼0.1275	0	100
		K	0 ∼1.5525	1	96.42

group labeling model using decision trees, testing ID3, C4.5, CART, and CHAID algorithms. The results revealed significant performance variations among algorithms based on dataset characteristics.

CHAID outperformed the others on the Iris and Seed databases due to its efficiency with smaller datasets. By employing chi-squared tests for variable selection, CHAID demonstrated robustness against outliers and noise. In contrast, ID3, C4.5, and CART, which rely on information gain, were more susceptible to outliers in smaller datasets like Iris and Seed. CHAID's tendency to create simpler decision trees helps to prevent overfitting, especially in datasets with limited records.

C4.5 excelled on the Wines and Glass databases, likely due to its pruning capability, which leads to more generalizable models. This is particularly advantageous in datasets with class imbalances and high dimensionality, such as the Glass data.

These findings emphasize the importance of algorithm selection in model accuracy. Careful consideration of data characteristics and algorithm strengths is crucial for optimal performance. The study confirms the feasibility of the proposed automatic group labeling model and highlights the significant impact of algorithm choice on model effectiveness and applicability.

Moreover, the study seeks to go beyond merely improving agreement rates. Its objective is to uncover novel labels that have been overlooked in prior research, as exemplified by the Wine dataset

References

Bertsimas, D., Orfanoudaki, A., and Wiberg, H. M. (2020). Interpretable clustering: an optimization approach. *Machine Learning*, 110:89–138.

de Lima, B. V. A., Machado, V. P., and Lopes, L. A. (2015). Automatic labeling of social network users scientia. net through the machine learning supervised application. *Social Network Analysis and Mining*, 5:1–10.

Di Teodoro, G., Monaci, M., and Palagi, L. (2024). Unboxing tree ensembles for interpretability: a hierarchical visualization tool and a multivariate optimal re-built tree. *EURO Journal on Computational Optimization*, 12:100084.

Dimotikalis, Y., Karagrigoriou, A., Parpoula, C., and Skiadas, C. H. (2021). *Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools*. John Wiley Sons, Incorporated, Newark.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Filho, F. I., Machado, V. P., Veras, R. D. M. S., Aires, K. R. T., and Montenegro Leal Silva, A. (2020). Group labeling methodology using distance-based data grouping algorithms. *Revista de Informática Teórica e Aplicada*, 27(1):48–61.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Hara, S. and Hayashi, K. (2016). Making tree ensembles interpretable. *arXiv preprint arXiv:1606.05390*.

Lopes, L. A., Machado, V. P., Rabêlo, R. A., Fernandes, R. A., and Lima, B. V. (2016). Automatic labelling of clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, 106:231–241.

Lopes, L. A., Machado, V. P., and Rabêlo, R. A. L. (2013). Automatic labeling of groupings through supervised machine learning.

Lopes, L. A., Machado, V. P., and Rabelo, R. D. A. L. (2014). Automatic cluster labeling through artificial neural networks. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 762–769. IEEE.

Machado, V. P., Ribeiro, V., and RABêLO, R. (2015). Rotulacao de grupos utilizando conjuntos fuzzy. In *XII Simposio Brasileiro de Automacao Inteligente-SBAI*, number 12, pages 355–360.

Moura, M., Veras, R., and Machado, V. (2022). Caibal: Cluster-attribute interdependency based automatic labeler. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 1109–1116, New York, NY, USA. Association for Computing Machinery.

Russell, S. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson.

Serengil, S. I. (2021). Chefboost: A lightweight boosted decision tree framework. https://doi.org/10.5281/zenodo.5576203.

Silva, L. E. S., Machado, V. P., Araujo, S. S., de Lima, B. V. A., and Veras, R. d. M. S. (2021). Using regression error analysis and feature selection to automatic cluster labeling. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 376–388. Springer.