Practical application 3: unsupervised classification

Álvaro García Barragán
alvaro.gbarragan@alumnos.upm.es
Universidad Politécnica de Madrid

December 12, 2023

Abstract

This paper proposes a comprehensive study aimed at predicting the survival states of patients diagnosed with liver cirrhosis with probabilistic and non-probabilistic supervised classification models. Data are sourced from a Mayo Clinic study on primary biliary cirrhosis. The objectives include interpreting clusters generated through Hierarchical, Partitional and Probabilistic Clustering.

1 Introduction

The data provided are sourced from a Mayo Clinic [1] study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984. The dataset is a multivariate dataset [2], which, after preprocessing, comprises 312 instances, each corresponding to an individual patient. It encompasses 20 distinct clinical features offering insights into demographics, treatment, and various other factors as displayed in Table 1. The study aims to achieve the following objectives:

- Explore relationships among patients through the application of Hierarchical, Partitional and Probabilistic Clustering.
- Evaluate and interpret the clusters obtained for each method.

2 Methodology

Clustering analysis entails few research questions to be resolve as related in subsection 2.1, one of them is how data how to be prepared that is explain in subsection 2.2. To resolve this questions I proposed a possible solution in subsection 2.3. Additionally, an exploration of the data is conducted in subsection 2.4, and the software tools used are detailed in subsection 2.5.

2.1 Problem of clustering

In the medical domain, clustering can be utilized to segregate datasets into distinct groups of patients, thereby enhancing our understanding of the entire population. This stratification can be instrumental in personalizing treatment based on the characteristics of each cluster. Additionally, it aids in identifying key features that distinguish different clusters, streamlining the process of feature selection. By summarizing various features into more compact representations, clustering not only simplifies the data but also facilitates the development of more interpretable classification models. Clustering models can then be used understanding patient demographics or health trends. However, the are some questions to respond:

- What methodologies determine the optimal number of clusters (k) in a dataset?
- What are the optimal criteria or thresholds in Hierarchical clustering?
- What metrics and methods are best for evaluating the effectiveness of clustering results?

Table 1: Data set

Feature	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**	**Type**
N_Days	2006.36	1123.28	41.00	1191.00	1839.50	2697.25	4556.00	continuous
Drug	0.49	0.50	0.00	0.00	0.00	1.00	1.00	binary (Y:0, N:1)*
Age	50.05	10.59	26.30	42.27	49.83	56.75	78.49	continuous
Sex	0.12	0.32	0.00	0.00	0.00	0.00	1.00	binary (F:0, M:1)
Ascites	0.08	0.27	0.00	0.00	0.00	0.00	1.00	binary (N:0, Y:1)
Hepatomegaly	0.51	0.50	0.00	0.00	1.00	1.00	1.00	binary (N:0, Y:1)
Spiders	0.29	0.45	0.00	0.00	0.00	1.00	1.00	binary (N:0, Y:1)
Bilirubin	3.26	4.53	0.30	0.80	1.35	3.42	28.00	continuous
Cholesterol	369.51	221.26	120.00	255.75	322.00	392.25	1775.00	continuous
Albumin	3.52	0.42	1.96	3.31	3.55	3.80	4.64	continuous
Copper	97.65	85.34	4.00	41.75	73.00	123.00	588.00	continuous
Alk_Phos	1982.66	2140.39	289.00	871.50	1259.00	1980.00	13862.40	continuous
SGOT	122.56	56.70	26.35	80.60	114.70	151.90	457.25	continuous
Tryglicerides	124.70	61.93	33.00	87.00	114.00	145.25	598.00	continuous
Platelets	261.94	94.99	62.00	200.00	258.50	322.00	563.00	continuous
Prothrombin	10.73	1.00	9.00	10.00	10.60	11.10	17.10	continuous
Stage	3.03	0.88	1.00	2.00	3.00	4.00	4.00	ordinal
Edema_N †	0.84	0.36	0.00	1.00	1.00	1.00	1.00	binary
Edema_Y †	0.06	0.25	0.00	0.00	0.00	0.00	1.00	binary
Edema_S †	0.09	0.29	0.00	0.00	0.00	0.00	1.00	binary

[*] Y=Drug=Dpenicillamine; N=Placebo;

[†] Edema_N=No edema and no diuretic therapy for edema; Edema_Y=Edema despite diuretic therapy; Edema_S=Edema present without diuretics, or edema resolved by diuretics;

- Which specific algorithms are most effective for different clustering methods?
- What distance metrics should be used in various clustering scenarios?
- How should cluster solutions be evaluated and interpreted?
- How to deal with qualitative data in clustering analysis?

2.2 Data preparation

In this section, I explain how the data is processed to enable effective clustering. The majority of clustering algorithms use Euclidean Distance, which implies that there is a measurable numerical proportion between two features. Qualitative data encoded as strings cannot be accurately calculated using this method; therefore, preprocessing is necessary. The dataset undergoes the subsequent stages in this process:

1. **Remove Metadata Information:** Remove metadata information, such as the ID column, which is not relevant for modeling purposes.
2. **Convert Age to Years:** Convert the age values from days to years for improved model interpretability.
3. **Remove Null Values in Drug Feature:** Eliminate instances with missing values in the `Drug` feature to maintain data integrity, 106 instances are removed.
4. **Replace Missing Numerical Values:** For numerical features with missing values, apply the imputation method by replacing these missing values with the arithmetic mean of the respective feature. The following features have been inputted Cholesterol, Copper, Triglycerides, and Platelets with 28, 2, 30, and 4 null values, respectively.
5. **One-Hot Encoding for Categorical Features:** Perform one-hot encoding on categorical features to convert them into a numerical format suitable for modeling purposes.

2.3 Approach

In this section, I explain how the clustering is carried out. This approach try to adapt clustering to the business, meaning that the configuration of k and the thresholds are manually adjusted until the results are interpretable. This entails viability and comprehensibility by experts in the relevant field, in our case, the medical field. As the dataset contains more than two features, a dimensionality reduction technique is required for visualizing the results. To interpret the clustering, I trained a Decision Tree classifier to summarize the features of each cluster. The model was trained using 70% of the data and evaluated on the remaining 30%, to assess how accurately it predicts the labels of the clusters. For dimensionality reduction, I utilize Principal Component Analysis [3] (PCA). This algorithm, similar to other clustering methods, assumes that the data is normalized. However, training a Decision Tree with this normalized dataset lead to loss of interpretability. To address this, I generate two datasets: *DS* and *DS_normalized*. *DS* is used to train the classifier model and clustering models that utilize Mahalanobis distance [4]. In contrast, *DS_normalized* is employed for PCA and other clustering methods that require normalization. Additionally, it is important to note that the Decision Tree has been trained with a $maximum\_depth = 6$ to ensure that the resulting visualization is understandable for a human.

2.4 Chernoff faces

In this section, it is demonstrate the use of Chernoff Faces [5, 6]. This visualization method, Figure 1, can effectively illustrate the differences between patients. To interpret these differences, one needs to compare the features individually. Nevertheless, is effective way to visually identify common features among patients, such as with Patients 5, 6, and 7.

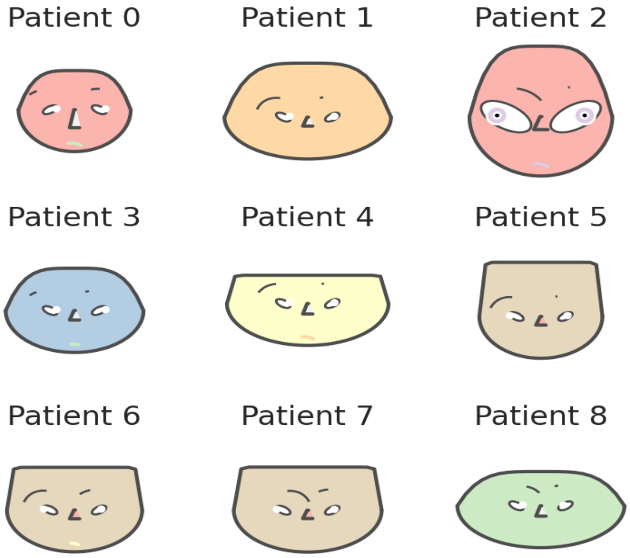


Figure 1: Chernoff faces with a subset of patients

2.5 Software Tools

The software utilized for the execution of this project was implemented within the Python Jupyter Notebooks environment to systematically document all experimental procedures. The codebase and corresponding experiment records are accessible via the provided repository link: https://github.com/Alvaro8gb/PracticalApplicationsML. The primary libraries utilized are [7, 8].

3 Experiments and Results

In this section, I present representative methods from Hierarchical, Partitional, and Probabilistic Clustering. In Hierarchical clustering, a threshold is proposed to determine where the dendrogram is cut. Using this threshold, k is obtained. For Partitional and Probabilistic methods, k is chosen as a parameter of the model. The Table 2 display this configurations for each method with the distance used and the F-score of the evaluation explained in subsection 2.3. To view the gallery of figures in an expanded format or explore visualizations that are not included here due to space constraints, please refer to the repository in subsection 2.5.

Table 2: Clustering methods and parameters

Method	Distance	threshold	k	F-score	Ref
Hierarchical Single	mahalanobis	5.6	16	0.9200	Figure 2
Hierarchical Complete	mahalanobis	12	5	0.7965	Figure 3
Hierarchical Complete	mahalanobis	10	10	0.6663	-
Hierarchical Centroid	euclidean	6	13	0.9015	-
Hierarchical Ward	euclidean	15	9	0.8792	Figure 4
Hierarchical Ward	euclidean	18	7	0.8940	-
Partitional KMeans	euclidean	-	3	0.8617	-
Partitional KMeans	euclidean	-	5	0.8243	Figure 5
Probabilistic Gaussian Mixture	-	-	3	1.0000	Figure 6
Probabilistic Gaussian Mixture	-	-	5	0.8901	-

4 Discussion

In this section, the discussion of the results is presented. The results are discussed in three ways: firstly, through a comparative analysis with other methods (subsection 4.1), secondly, by providing individual comments the visualization of each method (subsection 4.2) and finally by providing a feature interpretation of one selected clustering method (subsection 4.3). However, this paper has certain limitations. The interpretation of the cluster may not be entirely accurate, as I am not an expert in the domain. The appropriate decision regarding whether the cluster makes sense should be made by a doctor in this field.

4.1 Comparative Analysis

The Table 2 illustrates the insights gained from practicing clustering. It can be observed that Hierarchical Single Clustering yields a very high F-score with numerous labels (clusters), primarily due to one cluster having many instances. Additionally, it is evident that in Hierarchical Complete Clustering, a slight decrease in the threshold leads to a duplication in the number of clusters. Moreover, It can also be observed that the threshold is indirectly correlated with the number of clusters (k). For methods such as Single or Centroid with a lower threshold, the value of k tends to be higher. Conversely, Ward, with the highest threshold, is the method associated with the lowest value of k. Finally, we can appreciate that as we increase k, the classifier has less F-score, as demonstrated by Complete, KMeans and Gaussian Mixture clustering.

4.2 Individual Analysis

In this section It is discuss the most relvant figures to empahis only in remarkable clustering methods. The selected figures are the following:

- Figure 2: there are 16 clusters exhibiting a non-stratified pattern, where one cluster contains a substantial number of instances, while others have very few. Additionally, it is apparent that the algorithm does not adjust well for outliers; instead, it groups them into individual clusters.

- Figure 3: show a setup with 5 clusters, where one cluster has a larger number of patients compared to the others, each of which has only one patient. The PCA does not reveal a recognizable margin that separates one cluster from the others. Furthermore, the tree does not exhibit distinctive features for each cluster.
- Figure 4: display a well-balanced configuration of 9 clusters with a more homogeneous structure and flexible boundaries, where each cluster contains a similar number of instances. The tree displays an high number of leaf nodes with a well-balanced structure; however, it lacks clear branches delineating individual clusters.
- Figure 5: the clustering boundary margins are rigid, with each cluster containing a similar number of instances. The tree exhibits many leaf nodes, forming a balanced structure, and reveals appreciable patterns within its overall architecture.
- Figure 6: the method have flexible clustering boundary margins, with each cluster containing a similar number of instances.

4.3 Feature Clustering Analysis

Since every tree model inherently involves a complex interpretation of selected features, I chose to evaluate the model shown in Figure 6 because it comprises only three clusters, exhibits a structured pattern and F-score of 1.00 asses that all population are represent by this clusters. The summary of each cluster is the following:

- **Cluster 0 (Class 0, 92 patients):** This cluster may be characterized by patients with moderate-severity biliary cirrhosis, as indicated by the absence of edema (`Edema_N` $= 0$) in 45 out of 60 patients. Only 15 out of 60 have `Edema_S` $= 1$ and no ascites (`Ascites` $= 0$)[1]. Moreover, considering an average bilirubin level of 3.26 and a range reaching up to 28.00 (see Table 1), individuals in this cluster generally exhibit bilirubin levels ranging from 9.4 to 15.85. Consequently, a significant portion of patients in this group (26 out of 60) tends to have elevated bilirubin levels.
- **Cluster 1 (Class 1, 194 patients):** Patients in this cluster appear to exhibit less severe biliary cirrhosis, since all not having edema (`Edema_N` $= 1$). The majority are females (`Sex` $= 0$), wide range of bilirubin (`Bilirubin` ≤ 9.4, $\mu = 3.26$), a broad spectrum of cholesterol (`Cholesterol` ≤ 7893, $\mu = 369.51$), and a wide range of Copper (`Copper` ≤ 271.0, $\mu = 97$).
- **Cluster 2 (Class 2, 26 patients):** This cluster may represent patients who exhibit significant symptoms of edema despite diuretic therapy (`Edema_Y` $= 1$), which could imply a more advanced stage of biliary cirrhosis. Additionally, Ascites is present (`Ascites` $= 1$) in 4 out of 18 patients.

5 Conclusion

This project has effectively showcased the application of various clustering methods on a dataset related to biliary cirrhosis. The findings indicate a diverse range of algorithms and configurations. However, the crucial aspect lies in the ability to interpret the meaning of the clusters in real life. Emphasizing the significance of meaningful interpretation over relying solely on a doctor in this domain to validate the clusters is paramount. In interpreting the results, it is concluded that a smaller number of clusters and a reduced dimensionality of the decision tree, summarizing the clusters, enhance the interpretability of each cluster.

The paper also demonstrates the parsimony [10] of decision tree classifiers, suggesting that those with fewer target classes exhibit higher F-score. Furthermore, for this dataset, methods such as Ward or KMeans exhibit a more homogeneous division of clusters, while others, like Single, are less stratified.

Finally the paper proposed that Probabilistic clustering divide the patients in three distinct cohorts: moderate-severity of biliary cirrhosis, less severity of biliary cirrhosis and advanced stage of biliary cirrhosis. Each cluster is related with knowledge of the disease.

[1] While both ascites and edema involve fluid retention, ascites specifically refers to the accumulation of fluid in the abdominal cavity, often associated with liver disease [9].

References

[1] M. Clinic, 2023. https://www.mayoclinic.org.

[2] D. E., G. P., F. T., F. L., and L. A., "Cirrhosis patient survival prediction." UCI Machine Learning Repository, 2023. https://doi.org/10.24432/C5R02G.

[3] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[4] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.

[5] A. Antonov, "Making chernoff faces for data visualization," 2016. https://mathematicaforprediction.wordpress.com/.

[6] H. Chernoff, "The use of faces to represent points in k-dimensional space graphically," *Journal of the American statistical Association*, vol. 68, no. 342, pp. 361–368, 1973.

[7] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

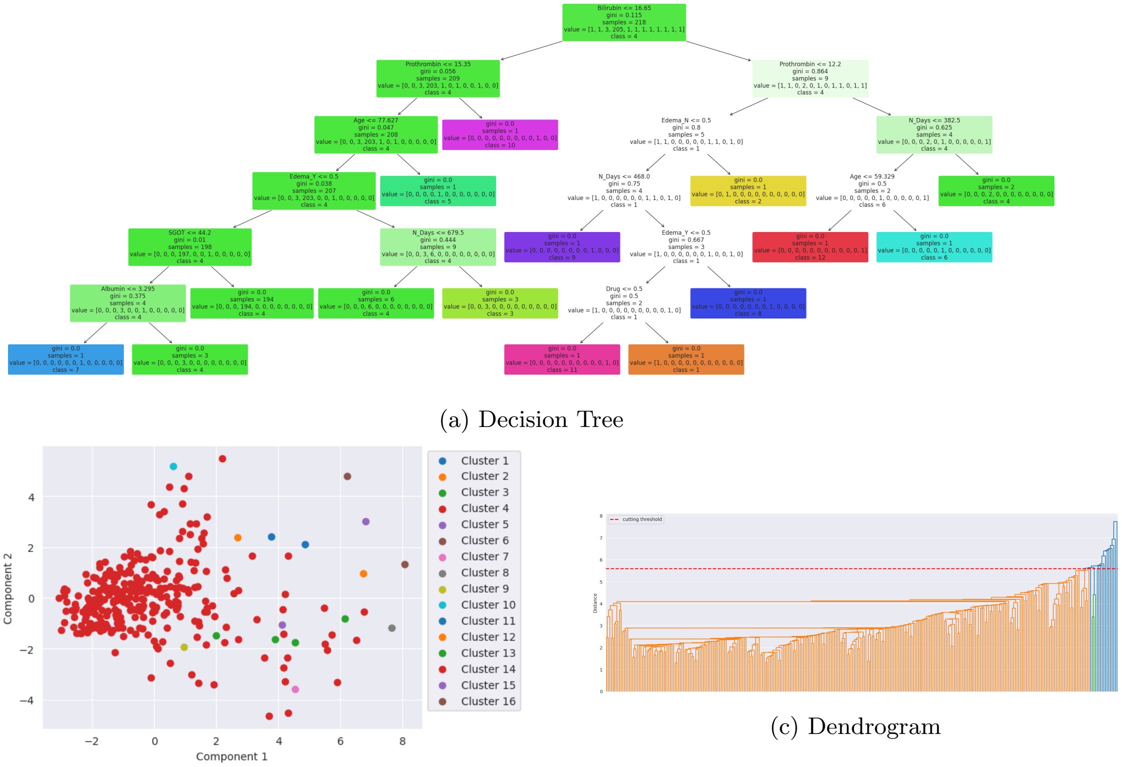
[8] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[9] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), "Medlineplus - cirrhosis," 2019. https://www.ncbi.nlm.nih.gov/books/NBK470482/.

[10] P. Domingos, "The role of occam's razor in knowledge discovery," *Data mining and knowledge discovery*, vol. 3, pp. 409–425, 1999.

[11] D. Arthur and S. Vassilvitskii, "How slow is the k-means method?," in *Proceedings of the twenty-second annual symposium on Computational geometry*, pp. 144–153, 2006.

A Cluster gallery

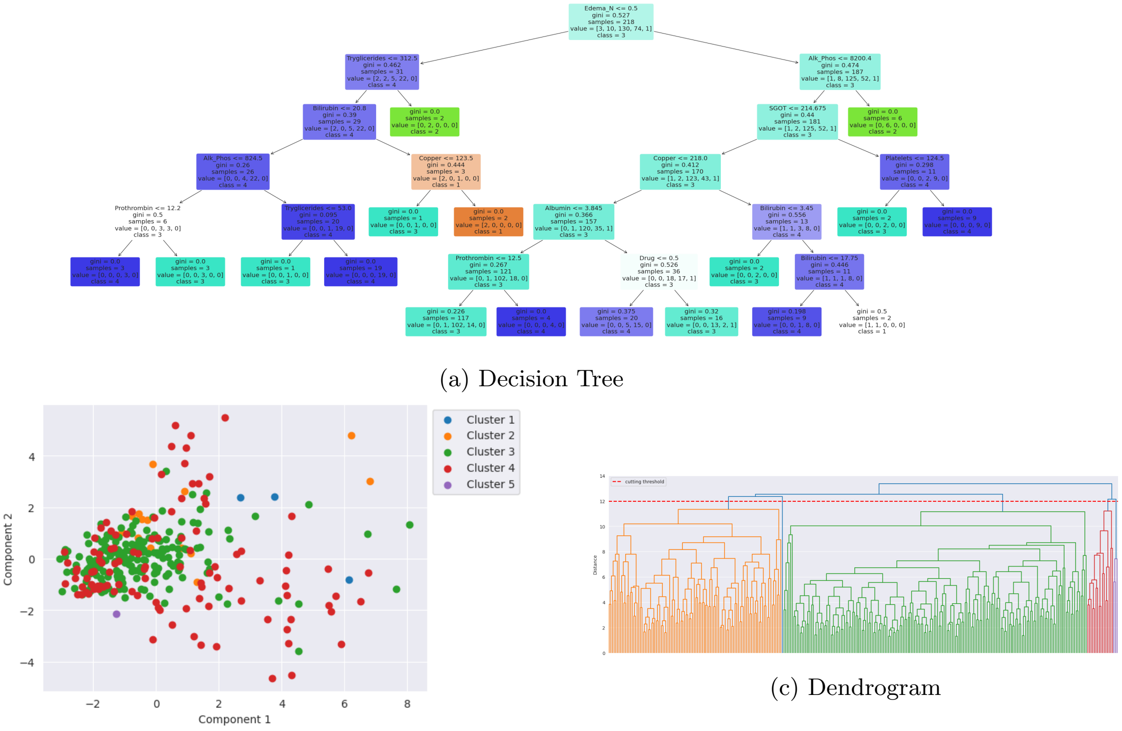


(a) Decision Tree

(b) PCA with clusters

(c) Dendrogram

Figure 2: Hierarchy Single

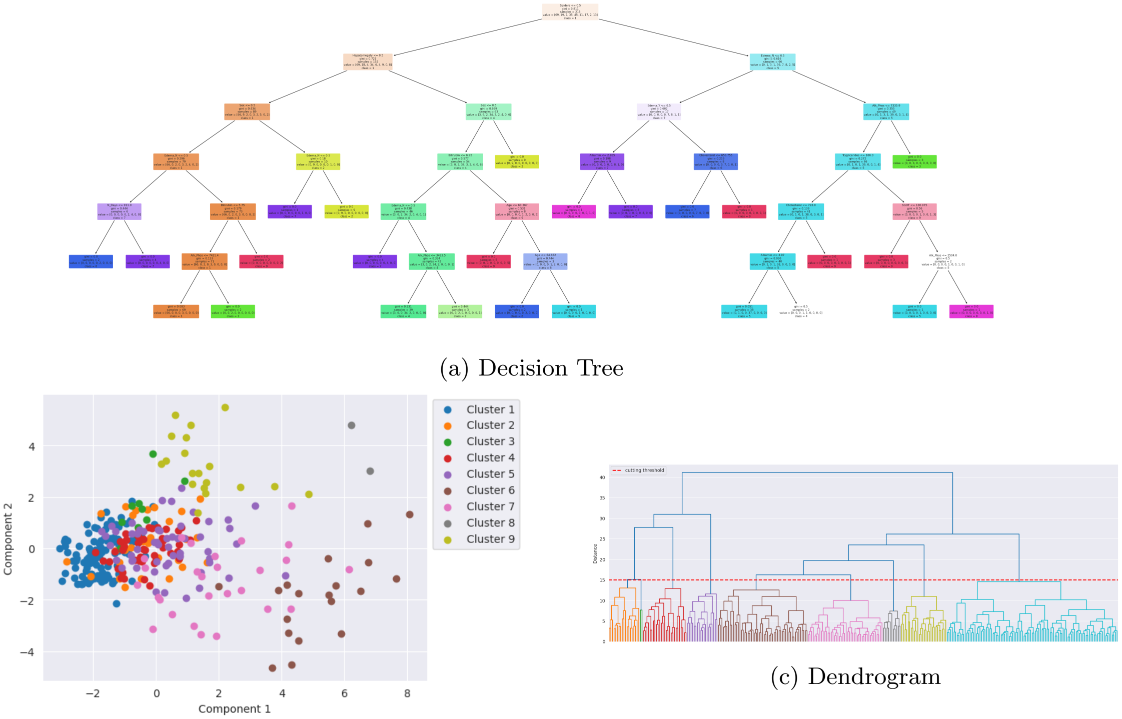


(a) Decision Tree

(c) Dendrogram

(b) PCA with clusters

Figure 3: Hierarchy Complete

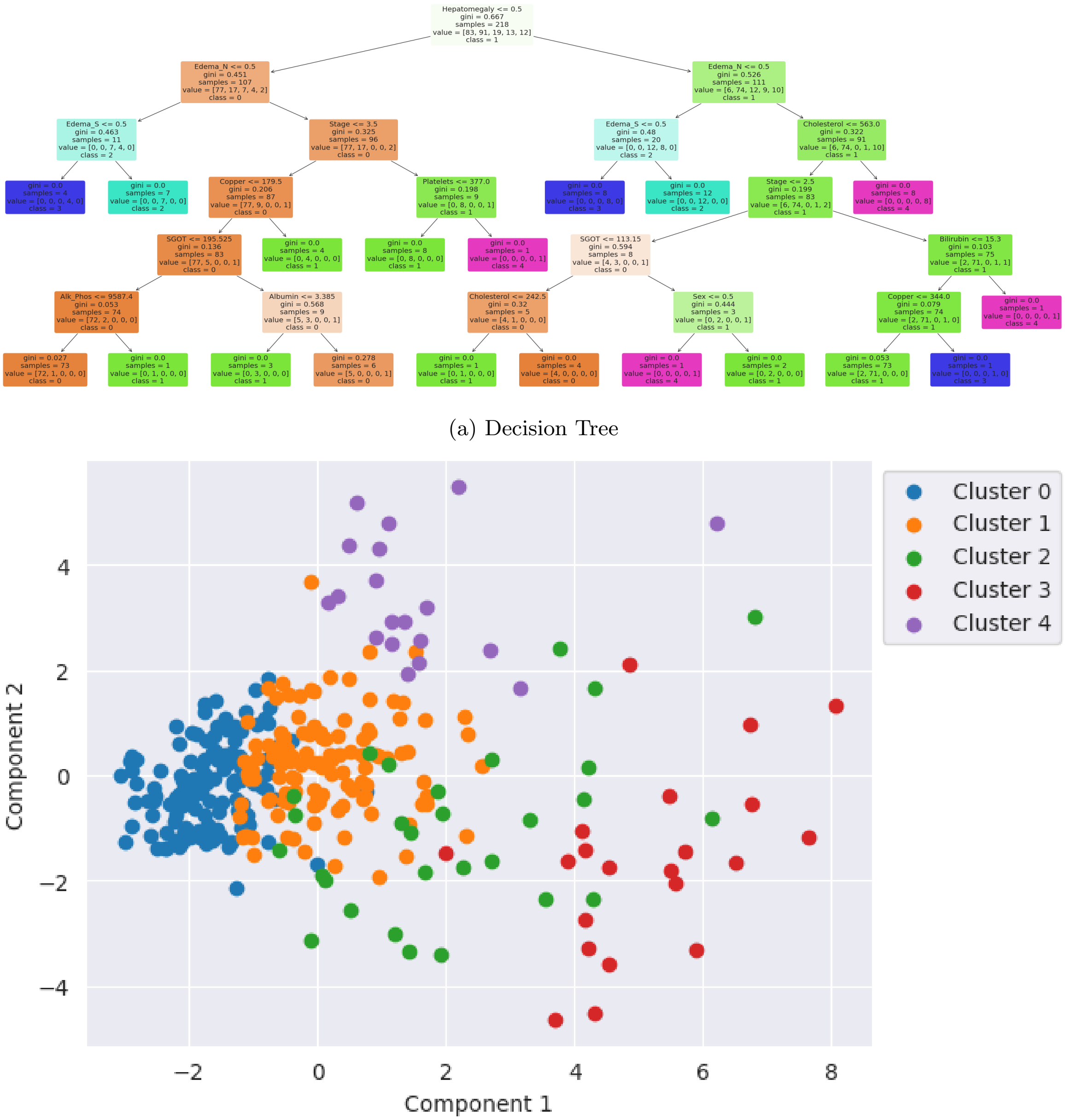


(a) Decision Tree

(c) Dendrogram

(b) PCA with clusters

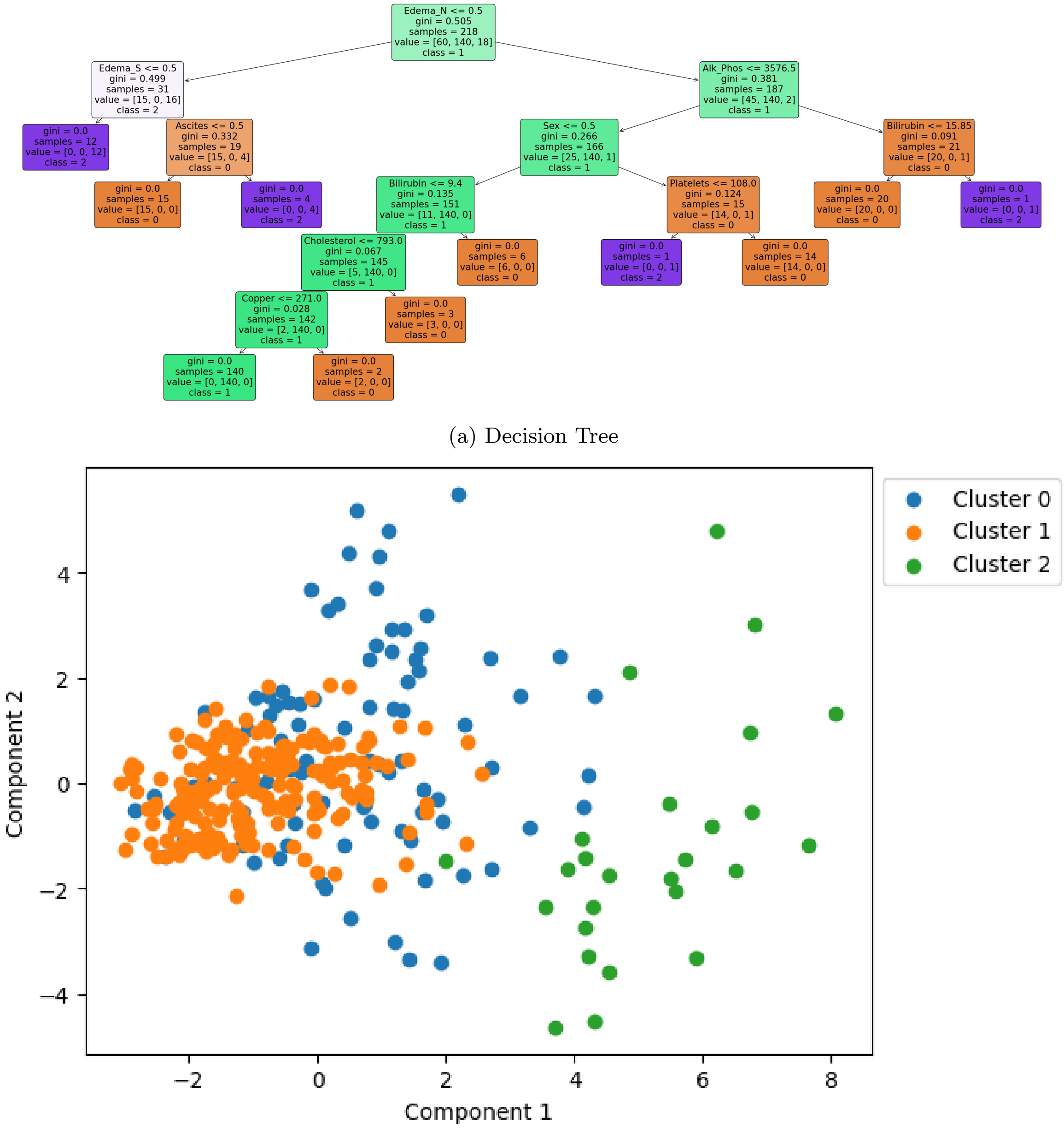
Figure 4: Hierarchy Ward



(a) Decision Tree

(b) PCA with clusters

Figure 5: KMeans Lloyd's algorithm [11]



(a) Decision Tree

(b) PCA with clusters

Figure 6: Gaussian Mixture