UNIVERSITAT DE BARCELONA

Work 1: Clustering Exercice

Introduction to Machine Learning

Hussnain Shafqat

João Valério

Eirik Grytøyr

Date: 23/10/2022

1. INTRODUCTION

The main goal of the present work is to analyse six clustering algorithms, using three datasets from the UCI repository, according to their performance and algorithmic specifications.

From the repository, the three datasets selected, cmc, Hepatitis and SatImage, have, accordingly, the following specifications: a big dataset with a mix of numerical and categorical data, a small dataset with numerical and categorical features, and a big dataset only with numerical data. Furthermore, in order to perform the algorithms with diversified testing cases, it was intended that the datasets presented dissimilar characteristics in the deviation of the class distribution, the percentage of instances belonging to the majority and minority classes and the percentage of missing values.

The algorithms to implement are divided into two modes of approach. The first set of algorithms, Agglomerative Clustering and Mean Shift, are employed through the library scikit-learn, while the latter, K-Means, Bisecting K-Means, K-Harmonic Means and Fuzzy Clustering, are rebuilt during this project.

With the aim of performing clustering validation to infer the optimum outputs from the algorithms for different datasets and parameters, it was elected four validation techniques, with respect to internal and external criteria. In the first nature of the measure, the methods chosen are Davies-Bouldin Index (DBI) and Silhouette Coefficient. Those have the purpose of measuring the intra-cluster and inter-cluster similarities. In the latter, the Adjusted Mutual Information and Adjusted Rand have the goal of assessing clustering with respect to ground truth.

In addition to the aforementioned, it is also established as an objective to compare the different algorithms to understand the principal disparities using confusion matrices for the distinct datasets.

2. Data

a. Characteristics

According to the UCI repository, the datasets selected and the respective characteristics are:

- **cmc:** numerical and predominantly categorical data with relevant size;
- **Hepatitis:** numerical and categorical data with small size;
- **SatImage:** only numerical data with significant size.

Furthermore, it is intended that the remaining characteristics are dissimilar between datasets, in order to study different cases. Particularly, the contrasts in the number of classes and missing values were the most important considerations to acquire a diverse combination of datasets.

The attributes mentioned in each dataset are presented in table 1.

Table 1 – Characteristics of the datasets.

		Characteristics							
		Cases	**Num.**	**Nom**	**Cla.**	**Dev. Cla.**	**Maj. Cla.**	**Min. Cla.**	**MV**
Data	**cmc**	1473	2	7	3	8.26%	42.70%	22.61%	-
	Hepatitis	155	6	13	2	29.35%	79.35%	20.65%	6.01% (42.23%)
	SatImage	6435	36	-	6	6.19%	23.82%	9.73%	-

* The value in parenthesis indicates that the feature with the highest rate of missing values has a portion of 43 % missing values. This is considered to make a large impact on the result, especially since the data is relatively small.

b. Preprocessing

In order to obtain a proper dataset to feed a machine learning algorithm, it is important to pre-process the data. This domain is separated into three central categories: Missing Values (i.), Different Types (ii.) and Different Ranges (iii.), in which the sequential order referred is in consonance with the respective code.

i. Missing Values

In the first subject, it is crucial to handle the problem of missing data, where two approaches are possible: deletion or imputation. The only dataset with missing values is Hepatitis, with a rate value of 6.01%. Since the number of cases is only 155, the deletion of observations results in a considerable loss of information. Thus, it is chosen to implement imputation. Through that, it is expected to keep the information about the observations without inserting possible bias into the data.

Numerical Data:

For the numerical data, the considered metric is the K-Nearest Neighbours Imputer, in which each sample's missing values are imputed according to the mean of the k nearest neighbours considered. As the function considered from the sckit-learn library is optimized to the general cases of numerical imputation, the best parameters among the tested are the default ones. According to that, k assumes a value of 5, with uniform weight distribution between the neighbours.

Categorical Data:

In the case of categorical features, the missing nominal value, '?', is considered as a new feature named 'Unknown'. This way, the number of categorical values increases, implying higher dimensionality in One hot encoding. However, as this type of data will be considered binary a wrong categorical attribution could result in poor clustering since this pattern would be transversal to all the missing values.

ii. Different Types

Secondly, it is necessary to consider the types of algorithms that are going to be executed further. In this study, all the implementations, in their basic form, work with numerical data. So, as two (cmc and Hepatitis) of the three datasets are not exclusively numeric, it is necessary to implement a conversion between the categorical features into numerical ones. Thus, the three datasets evolve into solely numerical ones.

The approach considered is One hot encoding, where the nominal variables are converted into a binary vector. A prior disadvantage is the increasing dimensionality of the data, worsening the time and memory complexity. However, by attributing the same weight to all the categorical features, it is avoided the insertion of bias. Therefore, the main advantage is the quality of clustering and, consequently, the further interpretations.

Table 2 exposes the number of features priorly and after the One hot encoding execution.

Table 2 – Variation in the number of features.

	One hot Encoding	
	Previously	**After**
cmc Features	9	24
Hepatitis Features	19	42

According to the dimensional incrementations (15 and 23) and the sizes of the datasets (1473 and 155 observations), the trade-off between complexity and clustering quality is beneficial to further development.

iii. Different Ranges

In the last stage, the normalization of all the numerical data is executed. Since different features have distinct numerical ranges, the weights between them are not uniformly distributed, inserting biased information in the models. As there is no relevant information

pointing out that certain features should have more weight than others, it is implemented a uniform weight distribution along the attributes.

The method considered is Min-Max Scaling, in which each instance has a linear value attribution between 0 (minimum) and 1 (maximum). This way, the categorical data converted do not need to be rescaled.

3. ALGORITHMS

The algorithms implemented to study the data are divided into two forms of approach. The first group, constituted by Agglomerative Clustering and Mean Shift, are applied through the library scikit-learn, while K-Means, Bisecting K-Means, K-Harmonic Means and Fuzzy Clustering, are rebuilt during this project.

To select the best parameters for each model, it was decided to test different possibilities and evaluate the results with respect to the confusion matrix, in which F1-Score (equation 3) is the evaluation metric utilized. The selection made was based on the completeness of the scorer. Since it considers precision (equation 1) and recall (equation 2), it is possible to comprehend two perspectives: from the positive predictions, how many are actually positive (precision); and from all the positive cases, how many are labelled as positive (recall).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1) \quad , \quad Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

For all the algorithms, with the exception of the Agglomerative Clustering, it was studied the best average seed to start the initialization, which is 99999. The seed is held during the entire analysis, in order to compare the results without affection from random initialization. When the parameters are determined, the most satisfactory number of clusters is picked based on the validation metrics (Topic 4).

As a final note, the number of clusters in this chapter is equal to the total of real classes in the data set, the ground truth. The exception is Mean Shift, which selects the best k autonomously.

a. Agglomerative Clustering

In the first algorithm, the parameters considered to study, by the respective order, are linkage and affinity.

Linkage:

This metric determines which distance to use between the clusters. The tested possibilities are ward, complete, average and single. Table 3 exposes the different F1-Scores obtained, leading to the conclusion that ward, by minimizing the variance of the clusters being merged, is the more suitable option.

Table 3 – F1-Score to the linkage values.

	F1-Score		
	Hepatitis	**cmc**	**SatImage**
Ward	0.75	0.40	0.50
Complete	0.50	0.30	0.48
Average	0.50	0.27	0.39
Single	0.51	0.20	0.18

Affinity:

This metric is the distance metric used to compute the linkage. Considering that the best linkage is 'ward', the only possible affinity is euclidean.

b. Mean Shift

In the algorithm Mean Shift provided by the scikit-learn library, the main parameters, bandwidth and seeds, are optimized automatically by the function.

For the remaining parameters, it is decided to work with the default values, since they do not present a relevant good implication in the accuracy. The important considerations are that all the points will be clustered and that the initial location of the kernel is defined by the data points.

c. K-Means

In the K-Means algorithm, it is compared two approaches, Forgy initialization and K-Means++, followed by the optimal distance metric (Euclidean, Manhattan and cosine).

Forgy Initialization vs K-Means++:

The initialization method in an algorithm has a substantial influence on the time to converge and the quality of the solution found.

To understand the best strategy, it is considered two different approaches, Forgy initialization and K-Means++. While the first computes a random sample from the data with k-size, in K-Means++ only the first centroid is determined through randomness, since the following

corresponds to the most distant point from the centroids already selected in the current iteration. Thus, the spatial distribution of the centroids increases.

Furthermore, with K-means++ is possible to apprehend which one of the first points (seed) in the dataset leads to better results. Then, this result can be replicated for the optimal seed and obtain always the same clustering. To achieve the same result with the Forgy initializer, it would be needed to test all the possible combinations in the data, which is not executable. So, it is not possible to mitigate the randomness of the Forgy method without a high time cost. One of the reasons why K-Means, in the most basic form, is not considered in the study, is that the level of randomness is even higher.

Table 4 demonstrates the F1-Scores for both metrics, suggesting that K-Means++ is able to provide better outcomes for the datasets.

Table 4 – F1-Score to Forgy initialization and K-Means++.

	F1-Score		
	Hepatitis	**cmc**	**SatImage**
Forgy	0.44	0.20	0.06
K-Means++	0.74	0.45	0.66

Distance Metrics:

In the distance metrics, Euclidean, Manhattan and cosine similarity, have the scores shown in table 5 indicating that the Euclidean metric is the selected one.

Table 5 – F1-Score to the distance metrics.

	F1-Score		
	Hepatitis	**cmc**	**SatImage**
Euclidean	0.74	0.45	0.66
Manhattan	0.44	0.20	0.06
Cosine	0.20	0.20	0.44

d. Bisecting K-Means

Since Bisecting K-Means is developed upon the K-means Algorithm, all the inferences related to the optimal parameters of the latter, logically, remain. On top of that, it is necessary to consider the best approach in the decision-making process of the worst cluster to split.

Cluster Decision-Making:

The three techniques considered and the respective grounds are the following:

- **Sum of Squares Error (SSE):** selects the cluster with the highest SSE value;
- **Largest Cluster:** designates the cluster with the most number of elements;
- **SSE + Largest Cluster:** nominates the largest cluster, which the SSE value is above the average SSE of all the clusters.

Table 8 symbolises the scores obtained, in which the sum of squared errors is the best approach.

Table 6 – F1-Score to determine the best Cluster Decision-Making.

	F1-Score		
	Hepatitis	**cmc**	**SatImage**
SSE	0.72	0.45	0.61
Largest Cluster	0.44	0.20	0.06
SSE + Larg. Cluster	0.44	0.20	0.06

e. K-Harmonic Means

In K-Harmonic Means, since the initialization metric has less effect on the results, it is considered the Forgy method to favour the time complexity. According to that, the remaining parameter is the power associated with distance (p).

Power associated with distance:

Since p is related to the distance calculation, by default, the value is two. In order to find the best result for p, it was tried values between two and seven.

Accordingly to the F1-Score quantities in table 7, the optimal power is equivalent to five.

Table 7 – F1-Score to determine the best p value.

	F1-Score			
	Hepatitis	**cmc**	**SatImage**	**Average**
2	0.72	0.35	0.7	0.59
3	0.75	0.42	0.67	0.61
4	0.75	0.42	0.65	0.61

5	0.75	0.44	0.65	0.61
6	0.73	0.45	0.62	0.60
7	0.73	0.45	0.30	0.49

f. Fuzzy C-Means Clustering

The version of Fuzzy clustering implemented is the Fuzzy C-Means Clustering, in which the parameter to select is also the degree of membership (m).

Membership:

As in point e., it is considered a range to the membership value, excluding a unitary one, since the intended result is a set of fuzzy clusters. The interval of values is maintained from the previous topic, as illustrated in table 8, where the optimal value is two.

Table 8 – F1-Score to determine the best membership value.

	F1-Score			
	Hepatitis	**cmc**	**SatImage**	**Average**
2	0.72	0.35	0.7	0.59
3	0.72	0.35	0.69	0.59
4	0.44	0.20	0.06	0.23
5	0.44	0.20	0.06	0.23
6	0.44	0.20	0.06	0.23
7	0.44	0.20	0.06	0.23

4. VALIDATION METRICS

Since in unsupervised learning the domain of a dataset is only constituted by the observations, $D=\{X_i\}$, the class labels, known for ground truth, are not available during the learning process. This way, there is no evaluation or testing data to evaluate or test the model.

Consequently, it is necessary to compare the outcomes of the distinct algorithms for various numbers of clusters, taking into account that the main goal is to obtain feasible representations that describe the unlabeled data, in order to discover the fundamental concepts. The metrics, internal and external criteria, pretend to evaluate, with a score from 0 to 1, the intra-class and inter-

class similarities, which are intended to be high and low, respectively. To achieve that, the metrics were normalized.

a. Internal Criterion

In the internal criterion, the main purpose is to evaluate the outputs without access to external information. So, the only knowledge considered by the internal indexes is the domain and the outcome clustering.

In this study, the two criteria considered are Davies-Bouldin Index (DBI) and Silhouette Coefficient. The classification score ranges from 0 to 1 for both metrics, but DBI indicates the best number of clusters with the minimum value in the function evolution, while Silhouette Coefficient uses the highest peak.

b. External Criterion

Unlike the previous method, the external criterion is not restricted to the information provided by the domain and the outcome clustering. This process is used when the number of class labels is known, as in the present work. Thus, it is possible to study the validity of the clusters by comparing the class (gold standard) and the clustering (model's output) labels.

The external indexes chosen are Adjusted Rand and Adjusted Mutual Information. Both perform scores on a unitary scale, in which the highest rise points out the best clustering numeral.

5. ANALYSIS OF THE VALIDATION METRICS

In the analysis procedure, it is intended to evaluate the outputs obtained from the different models in each dataset, in order to accomplish an understanding of the most satisfactory number of classes for each one. Posteriorly, the premises will be outlined and emphasized by the confusion matrices in topic 5.

In the subsequent subchapters, the plots acquired will be paired in the pursuing order:

- **Agglomerative Clustering:** the most isolated code from the set in terms of similarity.
- **K-Means + Bisecting K-Means:** algorithmic similarity purposes.
- **K-Harmonic Means + Fuzzy Clustering:** algorithmic similarity purposes as well.

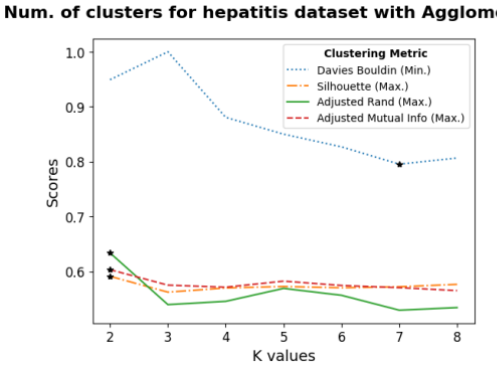
These plots have pointed out the optimal k-value in the algorithm. In Davies-Bouldin Index is the minimum of the function transition, while the remaining corresponds to the maximum.

As a last note, as Mean Shift determines automatically the optimal number of clusters, it is not necessary to effectuate the validation metrics.

a. Data - Hepatitis

In the first dataset (Hepatitis) the validation metrics present the subsequent marks.

Plot 1 – Validation metrics in Hepatitis (Agglomerative Clustering).



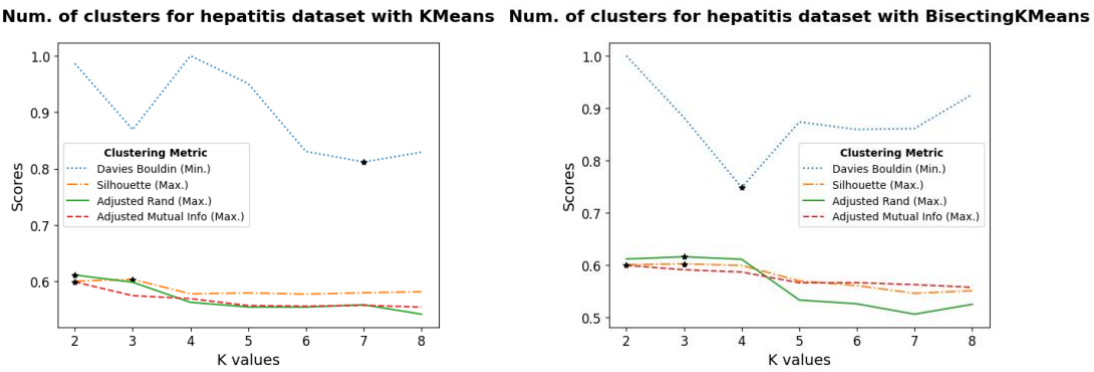
In plot 1, it is understandable that three metrics, the Silhouette Coefficient, Adjusted Mutual Information and Adjusted Rand, are in agreement that the best k-value is equal to two. Since the last two mentioned correspond to external criteria, thought that it is possible to validate the output obtained from the Silhouette Coefficient.

On the other hand, the DBI infers that seven is the optimal numeric. Precisely, it attributes 15% more score to seven clusters than two, which is a significant discrepancy. Particularly, the best number of clusters according to DBI is, in fact, the worst in Adjusted Rand. Even though the external indexes have access to the ground truth, the score oscillations between the different k-values are not significant, indicating that, methods that only access the domain and the output can easily have difficulties reaching the correct value.

In consonance with the analysis, the best cluster value using Agglomerative Clustering is two.

Plot 2 gives continuity to the present reasoning for K-Means and Bisecting K-Means algorithms.

Plot 2 – Validation metrics in Hepatitis (K-Means + Bisecting K-Means).



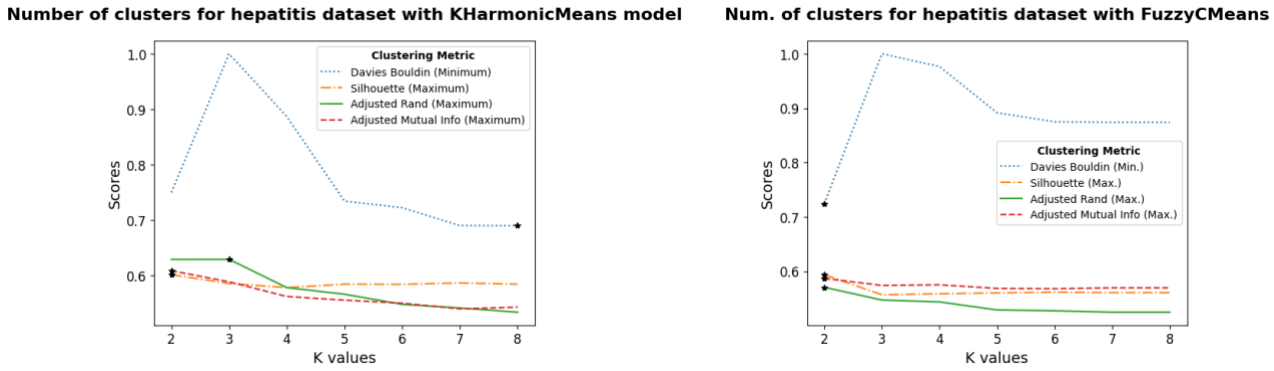
In K-Means the optimal value obtained with the external metrics, once again, is equal to two. However, neither one of the internal indexes support this premise. Silhouette Coefficient indicates that the global minima occurs when k is equal to three since two is only considered a local one. DBI displays that two is, actually, one of the worst possibilities to optimize the intra-cluster and inter-cluster relations. So, once again, it reveals that seven is the most satisfactory numeral.

In Bisecting K-Means, it is registered a new pattern in the functions' behaviour. Even though Adjusted Mutual Information keeps with the same judgment, Adjusted Rand, now, selects three as the best clustering. However, in the interval $k \in [2, 4]$, the latter illustrates an extremely soft variation in the score, which is not suitable to infer conclusions. This variation is also transversal to Silhouette Coefficient. DBI points out that four is undoubtedly a global minimum, which is not in complete disharmony with the remaining metrics.

Thus, while K-Means points out two as the optimal value, Bisecting K-Means is inconclusive between two and three.

In plot 3, it is illustrated the distinct metrics in K-Harmonic Means and Fuzzy Clustering.

Plot 3 – Validation metrics in Hepatitis (K-Harmonic Means + Fuzzy Clustering).



In K-Harmonic Means is verified the same pattern registered in plot 2 for Bisecting K-Means, since the external indexes are not in total agreement. However, this time, Adjusted Rand is nearly invariable in the interval $k \in [2, 3]$. On the internal indexes, DBI reports seven for the global minima, while the other is in consonance with Adjusted Mutual Information. The fact that DBI points out that three is, undoubtedly, the worst scenario, like in plot 1, can clarify the division on the external metrics.

In Fuzzy C-Means all the metrics, specially DBI, expose that two is the best value for k.

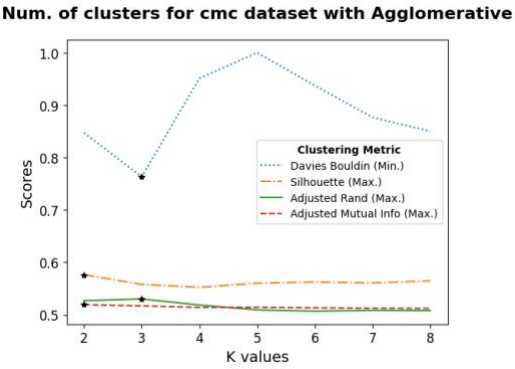
Additionally, it is important to refer that Mean Shift indicates that two is also the best number of clustering. So, according to the analysis conducted from plots one to three, the optimal

number of clusters is two, as indicated by the ground truth in table one. If the only possible analysis was through internal indexes, not accessing the gold standard, the conclusion would stay, however, not so evidently perceptible. These conceptions indicate the imputation of missing information was achieved at a satisfactory level.

b. Data - cmc

In the second dataset (cmc), numerical and categorical data with relevant size, the validation metrics are exhibited from plots 4 to 6, inclusively.

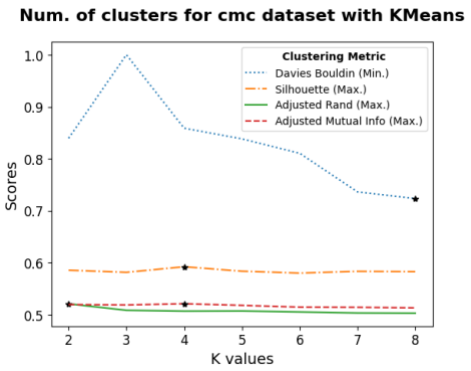
Plot 4 – Validation metrics in cmc (Agglomerative Clustering).

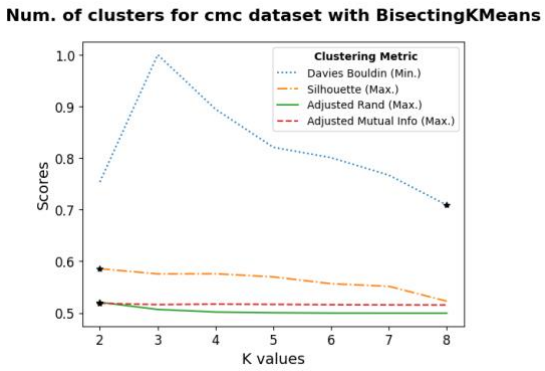


As shown in plot 4, the different metrics indicated that recurring to Agglomerative clustering, the most suitable value is between two and three. While Adjusted Mutual Information (external) and Silhouette Coefficient (internal) demonstrate that two is the preferable choice, Adjusted Rand (external) and DBI (internal) favour three as a selection.

Thus, only based only on plot 4, a unique inference is not unequivocal, so two and three showed to be the most satisfactory options. Even though it is very unbalanced, the relevant score difference that DBI points out might reveal util information to be considered in further analysis.

Plot 5 – Validation metrics in cmc (K-Means + Bisecting K-Means).





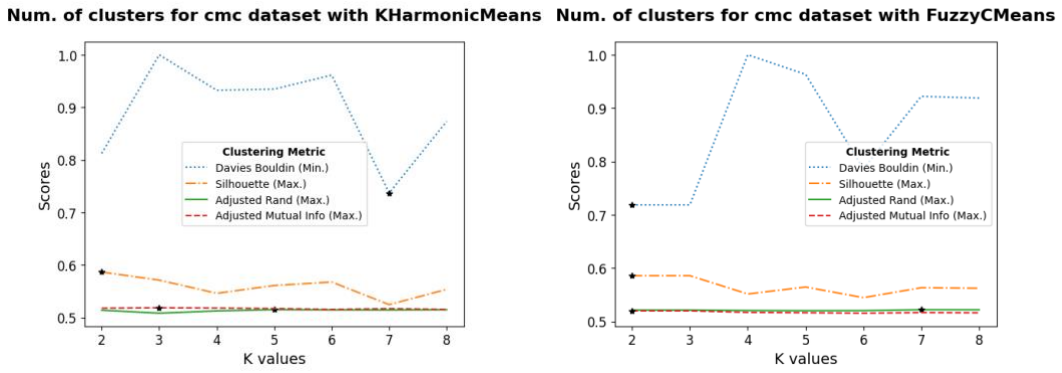
In the K-Means model as shown in plot 5, the external index Adjusted Mutual Information and Silhouette Coefficient conclude that the optimal number of classes is equal to four, while Adjusted Rand proposes two. So, as there is no consensus between the external indexes, and the three metrics mentioned have a score value roughly constant from k∈[2, 8], there is not a clear insight into the optimal clustering. Furthermore, the DBI recommendation is completely unsuited compared to the remaining metrics. Besides that, it seems a more unstable model than the remaining. Not only always registers notable variations in the y-axis, but also, in this plot considers three as the worst option, while in plot 4 it was the best.

In Bisecting K-Means both external indexes and the Silhouette Coefficient agree that k=2 produces the outcome desired. Additionally, even though DBI does not agree completely, it indicates two as a local minimum for the solution.

In summary, for K-Means the optimal number would be two or four and for Bisecting K-Means two is the chosen one.

Follows plot 6 with K-Harmonic Means and Fuzzy Clustering tests.

Plot 6 – Validation metrics in cmc (K-Harmonic Means + Fuzzy Clustering).



Relatively to the K-Harmonic Means there is not a possible inference related to the k-value, since each metric points out a distinct suggestion. Furthermore, the evolution of the external metrics is too steady, in order to give a clear understanding of the gold standard.

When it comes to the Fuzzy C-Means algorithm, the external indexes demonstrate the same behaviour seen priorly. However, in this turn, DBI, Silhouette Coefficient and Adjusted Mutual Information hypothesis are in harmony.

Thus, K-Harmonic Means analysis is inconclusive, while Fuzzy C-Means specifies k=2, even though the deduction is not evident due to the external indexes.

According to Mean Shift, the best clustering is in fact three, which corresponds to the gold standard. By merging all the analyses, the number of optimal clustering is inconclusive, mainly

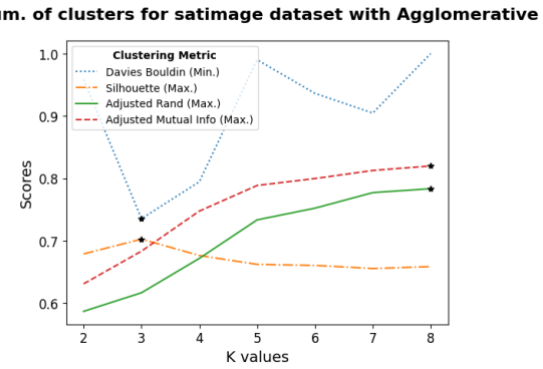
due to the fact that the external indexes, that have access to the gold standard, are, in the majority of the times, in disagreement.

The indefinite results might be due to the data. In fact, the only processes applied were normalization and One-hot Encoding, which, for this example, could not produce this level of disparities. Thus, the quality of the data, with inherently biased information, produces the observed outcomes. In order to converge to a unanimous conclusion, it would be needed to perform feature analysis to improve the quality of cmc.

c. Data - SatImage

The last dataset (SatImage), only numerical data with significant size, is analysed in the plots between 7 and 9.

Plot 7 – Validation metrics in SatImage (Agglomerative Clustering).

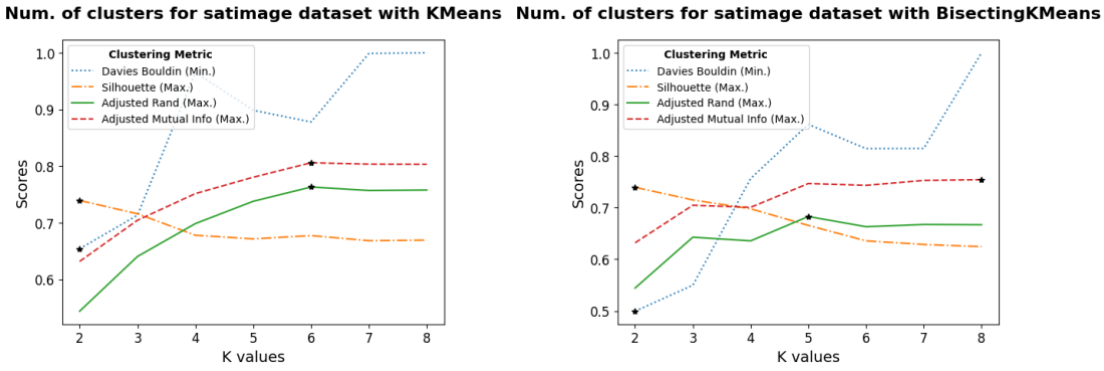


In plot 7, it is registered that both external indexes are in full conformity about 8 being the optimal clustering with the Agglomerative model. This is shown by a clear discrepancy of around 20% between the optimal choice (8) and the worst one (2). Both internal indexes point out that three is the best clustering, a significant contrast with the previous assumption.

Thus, since internal and external indexes are not in conformity, it is not possible to comprehend the most promising clustering. However, since the latter is based on the gold standard and results in the illustrated development, it is foreseeable that the clustering number will be high.

Plot 8 illustrates the metrics with respect to K-Means + Bisecting K-Means.

Plot 8 – Validation metrics in SatImage (K-Means + Bisecting K-Means).



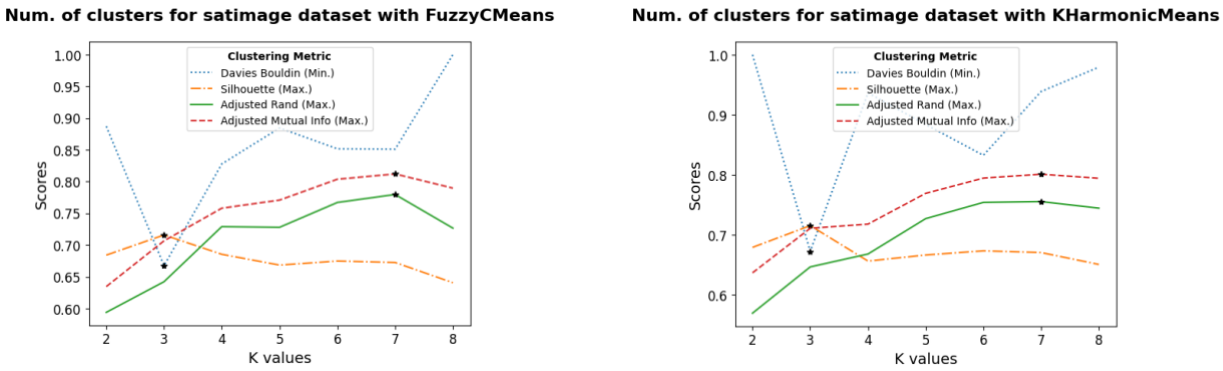
As already witnessed in the previous analysis, also K-Means in plot 8 reports a visible difference between internal and external indexes. While the first, specially DBI, indicates that the optimal value is two, the second obtains the number six. However, it is important to note that for k∈[6, 8], the external indexes are approximately steady, so a precise conclusion can be erroneous.

Bisecting K-Means shows the same behaviour pattern for the internal indexes, in which the graphical evolutions are extremely similar. The external indexes do not suggest the exact same value. Nonetheless, they denote, once again, that the optimal number of clusters is not low, as the internal indexes indicate constantly.

Despite the ambiguity between internal and external information, the optimal number for k is k∈[5, 8]. This conclusion is reached, because of the knowledge base richness of Adjusted Rand and Adjusted Mutual Information.

The last plot illustrates the metrics in the study applied to K-Harmonic Means + Fuzzy Clustering.

Plot 9 – Validation metrics in SatImage (K-Harmonic Means + Fuzzy Clustering).



By observing both results from the algorithms, it is comprehensible that the metric results are extremely similar. This is due to the proximity of the algorithms, that have been visible in every dataset.

According to the outcomes, once again, the judgements from internal and external methods are totally mismatched. Nevertheless, it is inferable that the optimal k registers values between 6 and 7, accordingly to arguments already reasoned previously.

The Mean Shift implementation reveals, correctly, that the right quantity of classes is six. Through the plots, it would be feasible to deduct by the external indexes that the truthful value of k is between six and eight. Since the internal ones always indicate eight as the worst optimization, the range can be reduced to six and seven. Further than that is not possible to conclude.

Additionally, without the external indexes, it would seem evident that the best k would be two or three, even though this is completely wrong. This indicates that for some metrics, only the intra-cluster and inter-cluster similarity as information, in certain datasets can be misleading.

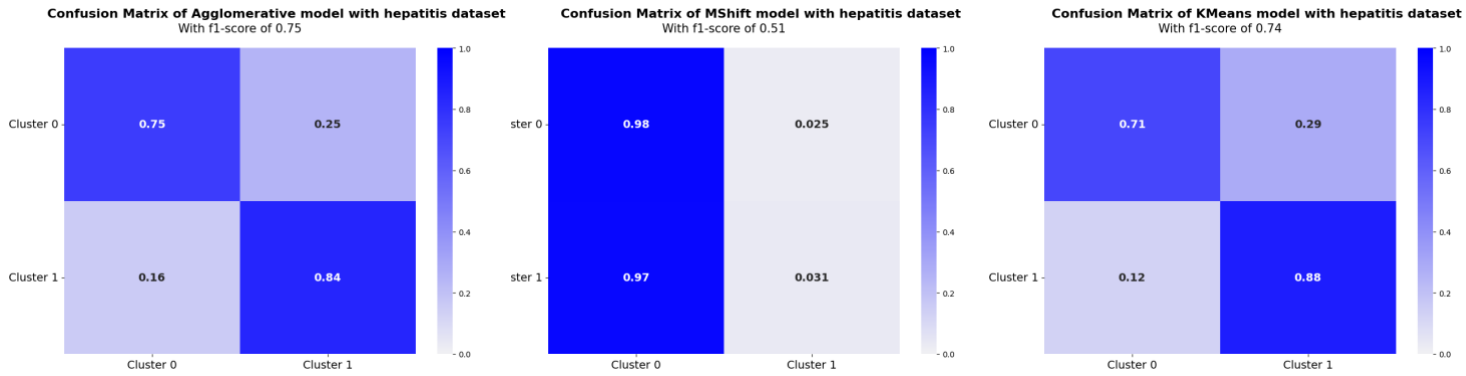
6. ANALYSIS OF THE CONFUSION MATRICES

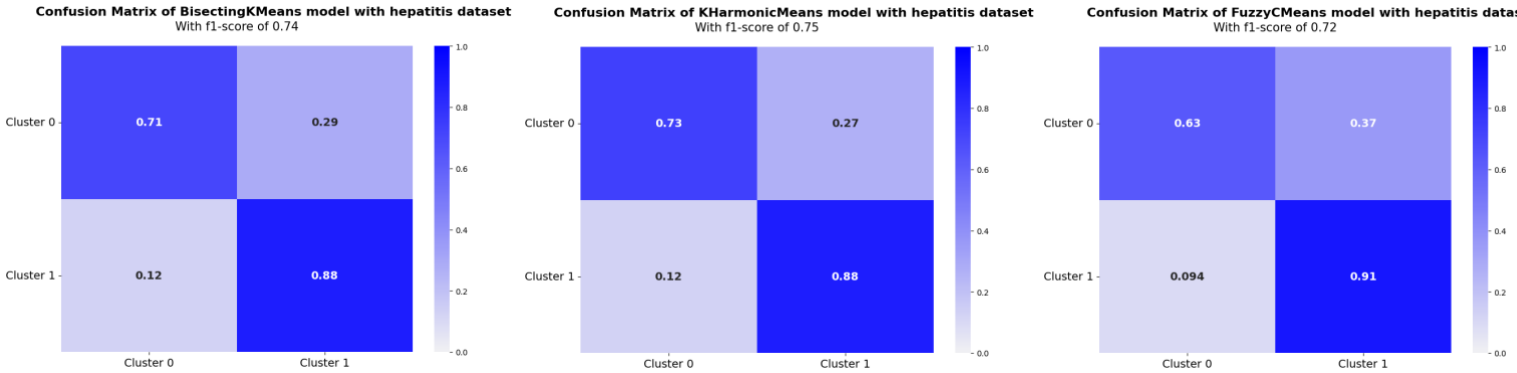
In other to complement the aforementioned study, the succeeding topics illustrate the respective confusion matrices in order to summarize the overall comprehension of the best number of clusters and methods. The scorer employed is the F1-Score, as explained in topic 3, and the number of clusters corresponds to the real number of classes in the dataset.

a. Data - Hepatitis

The first dataset (Hepatitis), numerical and categorical data with small size, presents the confusion matrices in plot 10.

Plot 10 – Confusion Matrices in Hepatitis (all the models).





Plot 10 represents how the distinct models are behaving on the Hepatitis dataset. According to the illustrations, all the algorithms follow the same tendency and are able to compute accurately over 70 % of cases.

The worst model presented is the Mean Shift, where the F1-Score only reaches 51%. This value is extremely inferior when compared with the remaining since the second worst model is able to perform 21% more accurately. On the other hand, the preferable models are Agglomerative Clustering and K-Harmonic Means with an F1-Score of 0.75. Excluding Mean Shift, the remaining models set nearby scores, by misclassifying simply a few more data.

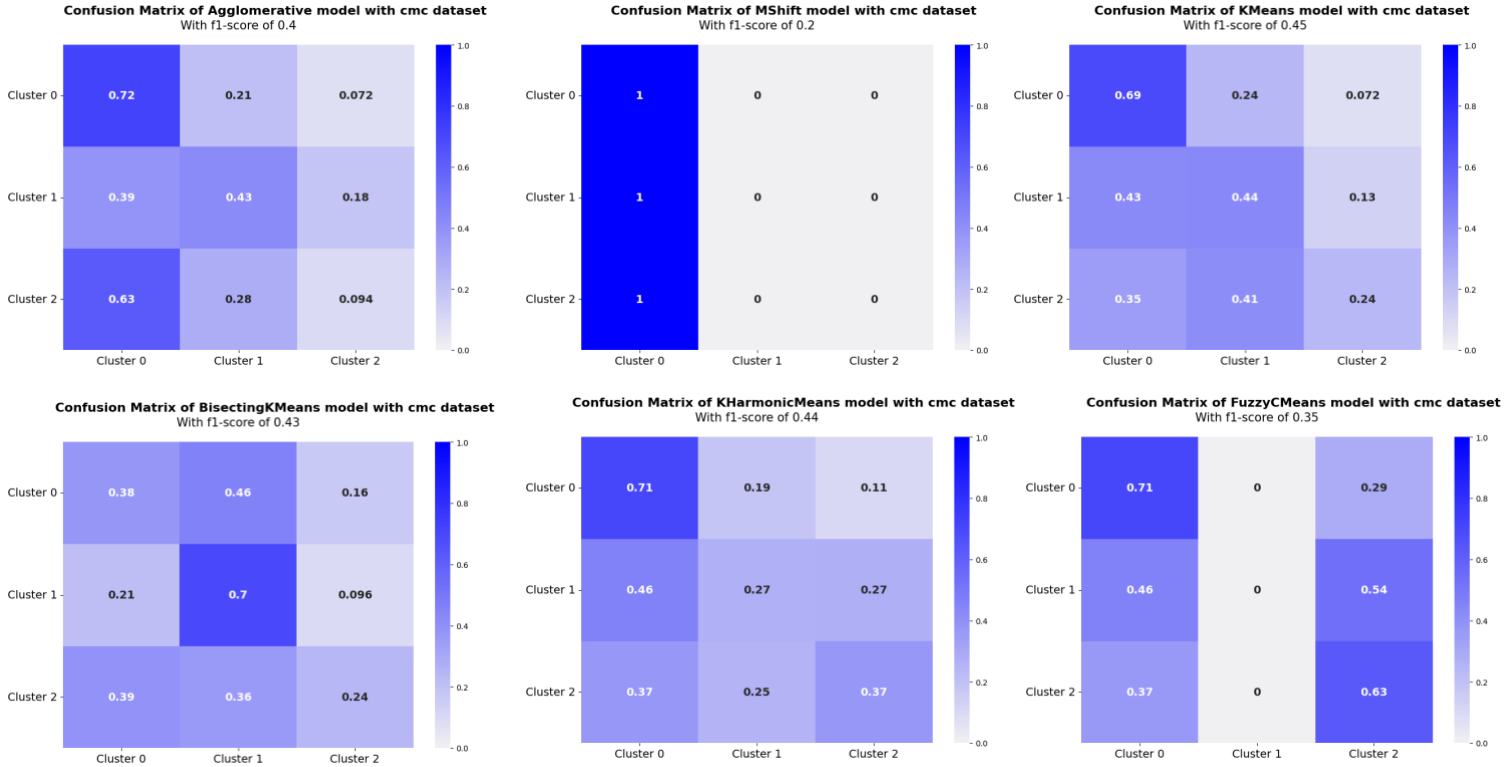
Notably, as expected, K-Means and Bisecting K-Means reach the same values in F1-Score. This occurs because in K-Means the seed defined is always the same and the selection of centroids is done by K-Means++. So, the initial division is always performed in the same manner. As the number of clusters is only two, the algorithm Bisecting K-Means only performs the call of K-Means and outputs the clusters given, without any influence. Additionally, K-Harmonic Means and Fuzzy C-Means reach also extremely similar scores, due to the closeness of the algorithms.

Overall, cluster 1, in the great majority of the cases, has the more accurate classifications of the observations. Finally, it is important to mention that those results are particularly noteworthy, considering that the study is about Unsupervised Learning. Thus, a quite good explanation regarding the patterns of the data could be provided. This theme is not explored taking into account the limitations of the project.

b. Data - cmc

The second dataset (cmc), numerical and categorical data with relevant size, illustrates the confusion matrices in plot 11.

Plot 11 – Confusion Matrices in cmc (all the models).



In plot 11, throughout the analysis of the overall F1-Scores, it is inferable that none of the models in the study performs satisfactorily in cmc dataset, since the maximum value is merely 45%. This idea is supported by the observations referred on point b. Chapter 5, where was not conceivable to comprehend the optimal clustering.

As well as in plot 10, once again the worst model is Mean Shift, with an extremely poor score of 20%. On the contrary, the best performances are set by, in increasing order, Bisecting K-Means, K-Harmonic Means and K-means with, respectively, 43%, 44% and 45%.

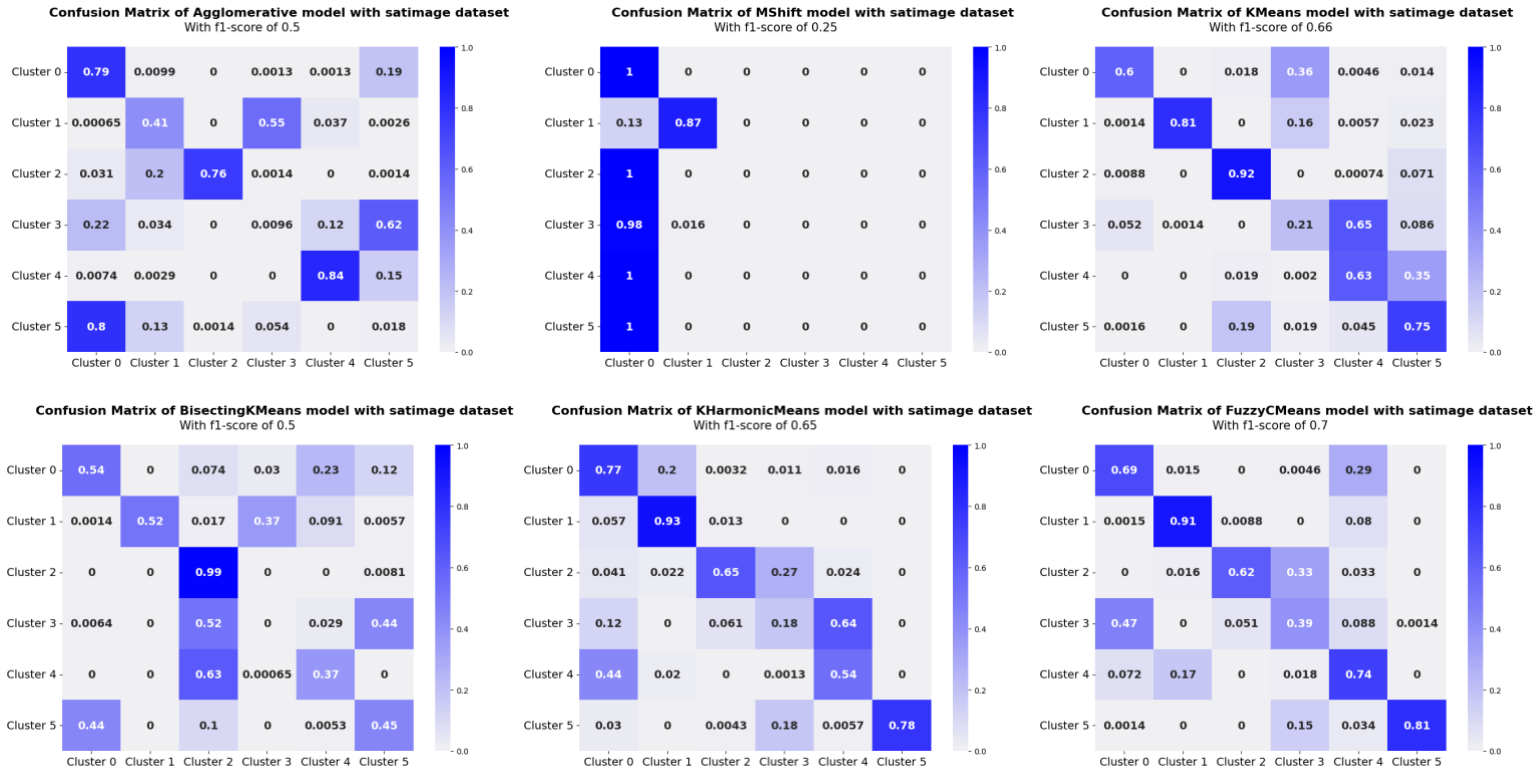
Particularly, Fuzzy C-Means marks really adequate values on clusters zero and two, especially in the second where the remaining algorithms are remarkably insufficient. However, with cluster one, no element is correctly classified, leading to a non-sufficient F1-Score. The same pattern is observed with Mean Shift in clusters one and two.

Overall, with these unsatisfactory outcomes, it is not possible to provide a complete description of the data, since the most promising models only can do it partially.

c. Data - SatImage

The last dataset (SatImage), only numerical data with significant size, has the following confusion matrices in plot 12.

Plot 12 – Confusion Matrices in SatImage (all the models).



In the final dataset, SatImage, the variety in the F1-Scores achieved increases, compared to the previous datasets. As observed in the prior datasets, also in SatImage the Mean Shift performs the worst accuracy of 25%, half of the second worst models (Agglomerative Clustering and Bisecting K-Means). In fact, it is not able to recognize four out of the six total clusters. This insufficient performance might be due to the dimensions and complexity of the data.

On the other hand, K-Harmonic Means, K-means and Fuzzy C-Means achieve, respectively, 65%, 66% and 70%, which are satisfactory scores. Particularly, among these models, cluster number three has significantly the most misclassifications, while one and five always reach considerably good accuracies. According to the previous marks, these algorithms would be suitable to present a feasible description of the data.

Particularly, it is evidently noticeable the link between Bisecting K-Means and K-Means algorithm, since the relation of the scores in the diagonal stay similar. For example, two is the most accurate cluster, while number three is the worst. The pattern is roughly similar between K-Harmonic Means and Fuzzy C-Means.

In general, as in point a, the results reached, for three models, are good and able to describe the present dataset in a feasible way. Furthermore, as seen in all points, different models provide distinct descriptions of the data, even when the accuracies and the models are extremely similar.

7. CONCLUSION

All the goals initially proposed were achieved with success and will be explained during the conclusion.

Firstly, it was comprehended that parameter optimization is crucial to reach the most acceptable possible outputs. Particularly, working on the initialization process, like applying K-Means++ and selecting the initial seed, increases significantly the clustering.

In order to conclude the right quantity of clusters, the implementation of internal and external metrics is essential. However, it was inferable that these do not always allow for reaching the real number of clusters. Furthermore, when the underlying characteristics of the data are not already optimal for the analysis, not even the external indexes are in agreement. As a last note, when external indexes are not available, the analysis lies upon the internal ones, in which the results are more unstable and less reliable.

By analysing the behaviour of the algorithms, it was inferable that the outputs achieved are always different, even though when the accuracies are roughly similar. So, it is concluded that distinct algorithms lead to dissimilar comprehensions of the patterns of the data.

Furthermore, the applicability of the model depends on the inherent characteristics of the dataset. So, even though the parameters are optimized for a particular dataset, that's not sufficient to accomplish adequate outcomes.

Among the algorithms, it was perceptible that the underlying mathematical similarities reproduce identical, but not exact, behaviours in the datasets, not only in the general accuracy values but also in the particular clustering results. This allows deducting that these models extract similar patterns.

According to the results achieved, for the numerical dataset, SatImage, the Fuzzy C-Mean is the best model, in mixed data, Hepatitis, the Agglomerative Clustering achieves the most suitable results, mainly due to the linkage metric (ward); and in cmc, predominantly nominal, the preferable algorithms are K-Means or K-Harmonic Means.

To conclude, in clustering, every dataset has to be seen as a new problem, where different techniques and algorithms need to be implemented, in order to achieve satisfactory results and extract the most valuable information from the data.

8. BIOGRAPHY

- MACQUEEN, J. (1967) - *SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS.* USA: University of California;
- STEINBACH, Michael *et al.* (2000) - *A Comparison of Document Clustering Techniques*. USA: University of Minnesota;
- COMANICIU, Dorin; MEER, Peter (2002) - *Mean Shift: A Robust Approach Toward Feature Space Analysis*. USA: IEEE;
- DOLFING, Henrico (2004) - *k-Harmonic Means Clustering Algorithm*. Germany: University of Konstanz;
- ARTHUR, David; VASSILVITSKII, Sergei (2006) - *k-means++: The Advantages of Careful*. USA: SIAM;
- ZHI, Xiao-bin; FAN, Jiu lun (2010) - *Some Notes on K-Harmonic Means Clustering Algorithm*. USA: Springer;
- CELEBI, M.Emre *et al*. (2012) - *A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm.* USA: Cornell University;
- AHMAD, Amir; HASHMI, Sarosh (2016) - *K-Harmonic means type clustering algorithm for mixed datasets*. Netherlands: Elsevier.