Big Data Mining

Assignment 3: Mining Over Datasets

Theodoros Efthymiadis
Maria Moutti

Theodoros Efthymiadis
Maria Moutti

Contents

Theodoros Efthymiadis
Maria Moutti

Data Set 1: Police – Crime Types

This data set is provided by the Greek Statistical Service Organization. It includes information about attempted and committed crimes in Greece in year 2016.

1.1 Preprocessing

The data set contains 6 features and 42 instances. Each instance stands for an individual crime type. There was a series of necessary preprocessing steps that were followed in order to proceed with the data analysis:

1. Manually readjust the columns of excel file (superclass became extra column "Κατηγορία" with values "ΕΠΙΚΡΑΤΕΙΑ", "ΚΛΟΠΕΣ - ΔΙΑΡΡΗΞΕΙΣ", "ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ" and "ΛΗΣΤΕΙΕΣ"). Drop the "Ληστείες ΔΟΥ" record -> only missing values
2. Fill missing values: It's hard to estimate the missing values here. There are 2 options under consideration:
    a. Replace with 0?
    b. Replace with average value of respective column?

    Different crimes have very strict definitions and the distinction between 'απόπειρα/τετελεσμένο' is often hard to make and requires domain expertise, which we don't have. Thus, we don't know if all crimes have the same ratio of 'απόπειρα/τετελεσμένο' and missing values will be set equal to 0.
3. $Επιτυχία = \frac{\text{εξιχνιάσεις}}{\text{απόπειρα+ τετελεσμένο}}$, Ranges between 0 and 1 for most values ('Επιτυχία' for 'Σεξουαλική εκμετάλλευση' is equal to 1.15). We assume that the additional crimes that were solved were committed in past years and, therefore, aren't relevant to our investigation for the year 2016. Therefore, the value 1.15 is set equal to 1.
4. All numeric features are normalized around mean using the $z = \frac{x-\mu}{\sigma}$ transformation
5. The only categorical feature, "Κατηγορία", is transformed to Boolean using one hot encoding. This results in 4 new Boolean features that resemble the four candidate values of the old categorical feature, which is dropped.

After preprocessing, the dataset will look like this:

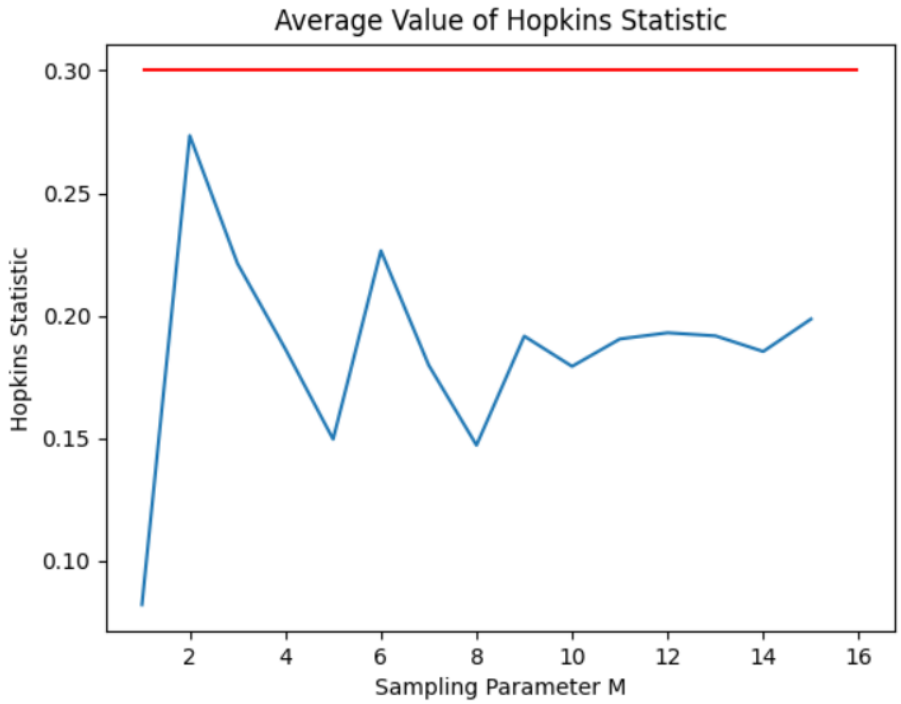
```
(venv) C:\Users\Ires\PycharmProjects\police>python scale.py
```

	Εγκλημα	τελ/να	απόπειρες	εξιχνιάσεις	ημεδαποί	αλλοδαποί	Επιτυχία	Κατηγορία_ΕΠΙΚΡΑΤΕΙΑ	Κατηγορία_ΚΛΟΠΕΣ - ΔΙΑΡΡΗΞΕΙΣ	Κατηγορία_ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ	Κατηγορία_ΛΗΣΤΕΙΕΣ
0	ΑΝΘΡΩΠΟΚΤΟΝΙΕΣ	-0.633800	-0.123451	-0.520089	-0.342632	-0.397086	1.413398	1	0	0	0
1	ΑΠΑΤΕΣ	0.084753	2.110745	0.432013	0.102336	-0.158228	-0.047254	1	0	0	0
2	ΑΡΧΑΙΟΚΑΠΗΛΕΙΑ	-0.619464	-0.362301	-0.524639	-0.346947	-0.511800	1.876523	1	0	0	0
3	ΒΙΑΣΜΟΙ	-0.619464	-0.229137	-0.529696	-0.378769	-0.403371	0.863706	1	0	0	0
4	ΕΚΒΙΑΣΕΙΣ	-0.623338	-0.353846	-0.567618	-0.379848	-0.459943	0.440543	1	0	0	0
5	ΕΠΑΙΤΕΙΑ	-0.046016	-0.402461	0.897698	0.312685	2.304204	1.741407	1	0	0	0
6	ΖΩΟΚΛΟΠΗ	-0.537709	-0.387665	-0.567112	-0.389556	-0.510229	-0.904922	1	0	0	0
7	ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	0.721358	-0.398234	0.226727	0.393589	-0.309085	-0.640486	1	0	0	0
8	ΛΑΘΡΕΜΠΟΡΙΟ	-0.373424	-0.396120	0.025486	-0.045447	0.856916	1.494283	1	0	0	0
9	Ν περί ΝΑΡΚΩΤΙΚΩΝ	1.460640	-0.385552	4.731385	5.626960	4.172635	1.776671	1	0	0	0
10	Ν περί ΟΠΛΩΝ	0.240901	-0.368642	1.462992	1.607679	0.588202	1.515051	1	0	0	0
11	Ν περί ΠΝΕΥΜΑΤΙΚΗΣ ΙΔΙΟΚΤΗΣΙΑΣ	-0.596797	-0.398234	-0.480144	-0.309732	-0.480371	1.732515	1	0	0	0

1.2 Clustering

In order to assess the cluster tendency of our dataset, we will calculate the Hopkins statistic for different values of the sampling parameter m (all features of the dataset will be used). In order to account for the random sampling of the Hopkins statistic, 10 trials will be executed for each value of

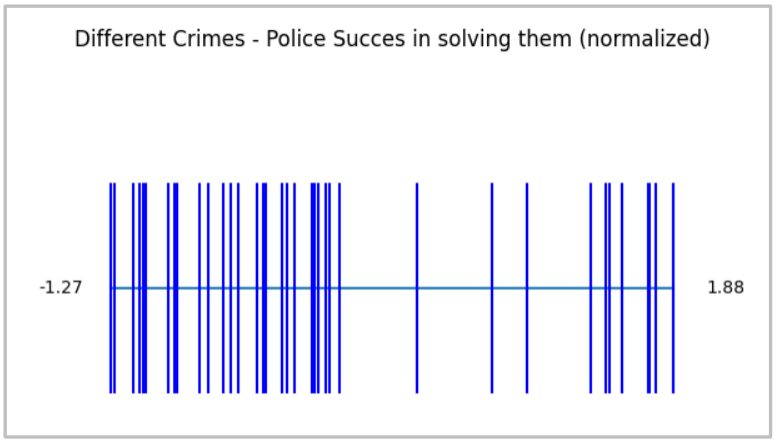
m and their average value will be calculated. Given that the Hopkins library is dependent on the Numpy library, maintaining a constant numpy seed will ensure reproducibility. The results are illustrated in the figure below:



The average value of the Hopkins statistic is lower than 0.3 for every value of the sampling parameter m. This is a strong indication of the existence of clustering tendencies within our data set.

1.2.1 Cluster the types of crimes based on the success of the police in facing/solving them

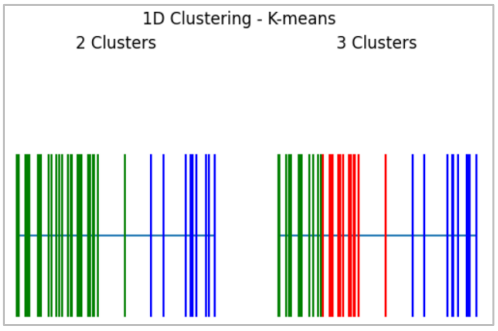
In order to address this task, we will perform a 1D clustering. Only the value of column 'Επιτυχία' (which was normalized) will be kept for each record and the values will be clustered using different algorithms. Given that there is no gold standard at hand to evaluate our results, the silhouette coefficient will be used. We will start by creating a 1D plot of our data:



Each bar stands for a different type of crime and its position along the horizontal line shows the success of the police in solving crimes of this type. Visual inspection of the figure indicates the existence of clusters and that the number of clusters should be either 2 or 3. We will use three

different algorithms to perform the clustering: K – means (proximity based), DBSCAN (density based) and AgglomerativeClustering (hierarchical).

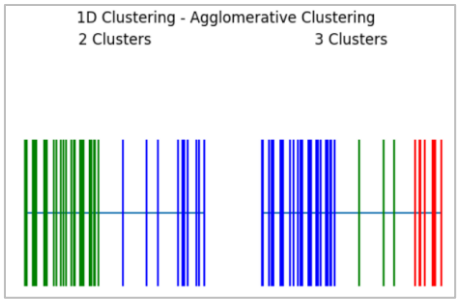
a) K-means clustering:



$Silhouette\ Coefficient = 0.7604$ $Silhouette\ Coefficient = 0.6523$

Both cases result to a sufficient value of the Silhouette coefficient, which indicates the existence of strong clustering tendencies. However, the approach with 2 clusters is superior.

b) Agglomerative Clustering

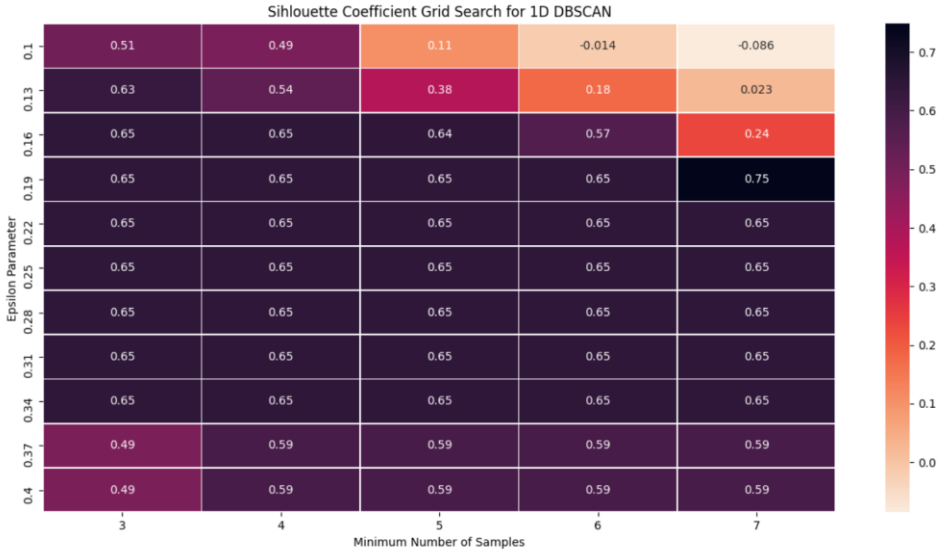


$Sihlouette\ Coefficient = 0.7482$ $Sihlouette\ Coefficient = 0.6472$

Similar with the case of K-means, the Agglomerative Clustering results indicate strong clustering tendency. Splitting the data in 2 clusters rather than 3 is preferred in this case as well.

c) DBSCAN

In the case of DBSCAN, we need to define the Hyperparameters epsilon and minimum sample size. In order to achieve that, we will conduct a grid search and use the silhouette coefficient as an evaluation metric to find the optimal parameter values.

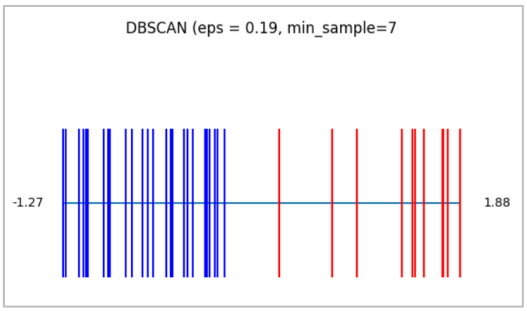


Close examination of the created heatmap dictates that the maximum value of the silhouette coefficient (0.75) is acquired through DBSCAN clustering using the following hyperparameters:

$$epsilon = 0.19$$

$$\text{min\_}samples = 7$$

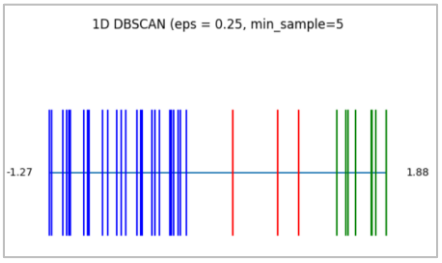
We will now plot the clusters that are created through DBSCAN with epsilon = 0.19 and min_samples=7:



If we have a closer look at the class labels that are stored in the corresponding pandas dataframe, we will notice that all red lines are corresponding to the class with label: -1. This means that the DBSCAN algorithm considered all those crime types outliers, while a single cluster was formed to contain all crimes that are represented in blue color. As a result, the calculation of the silhouette coefficient is not valid, given that there is only 1 cluster and outliers. We will, therefore, run DBSCAN for the hyperparameters of the second greater silhouette coefficient, namely 0.65. There are a set of parameter combinations that perform at a similar level. We will choose a combination of non-extreme values, but rather average ones:

$$epsilon = 0.25$$

$$\text{min\_}samples = 5$$



Based on intuition, this result seems more accurate than the last one. The three crime types in the middle of the graph cannot be clustered, while the rest clearly belong to one of the two clusters that are formed.

Finally, we will compare the results from the 3 clustering algorithms. For the cases of K-means and Agglomerative, only the 2 cluster results will be kept:

Class 0: High Success Rate

Έγκλημα	Επιτυχία	1D_Kmeans_2C	1D_Agglomerative_2C	1D_DBSCAN
ΑΝΘΡΩΠΟΚΤΟΝΙΕΣ	1.413398228	0	0	0
ΑΡΧΑΙΟΚΑΠΗΛΕΙΑ	1.876523034	0	0	0
ΕΠΑΙΤΕΙΑ	1.741406558	0	0	0
ΛΑΘΡΕΜΠΟΡΙΟ	1.494282911	0	0	0
Ν περί ΝΑΡΚΩΤΙΚΩΝ	1.776670777	0	0	0
Ν περί ΟΠΛΩΝ	1.515050709	0	0	0
Ν περί ΠΝΕΥΜΑΤΙΚΗΣ ΙΔΙΟΚΤΗΣΙΑΣ	1.732514883	0	0	0
ΠΛΑΣΤΟΓΡΑΦΙΑ	1.584054129	0	0	0
ΣΕΞΟΥΑΛΙΚΗ ΕΚΜΕΤΑΛΛΕΥΣΗ	1.876523034	0	0	0
Average	1.667824918			

All three algorithms have clustered these crime types together. The normalized average success rate of the police in solving crimes of this class is equal to 1.67 standard deviations above the mean. In that regard, we could argue that police either puts extra effort in solving these crimes, or that it is often somehow easier to do so. By quickly observing the list of the crimes, we could derive the following theories:

- Most crimes in these list have very high impact, such as large monetary losses or even deaths. Moreover, their occurrence significantly disrupts the peace in a society and, therefore, they are often prioritized by the police over other crimes.
- Some of these crimes produce very tangible evidence, which may increase the success rate of the police.

Class 1: Low success rate

Κλοπές - Διαρρήξεις από ιχε αυτ/τα	-1.086023637	1	1	1
Κλοπές - Διαρρήξεις ιερών ναών	0.012580319	1	1	1
Κλοπές - Διαρρήξεις καταστημάτων	-0.280433282	1	1	1
Κλοπές - Διαρρήξεις λοιπές	-0.555137785	1	1	1
Κλοπές - Διαρρήξεις οικιών	-0.770162015	1	1	1
Κλοπές - Διαρρήξεις σε συγκοινωνιακά μέσα	-1.244509469	1	1	1
Κλοπές με αρπαγές τσαντών	-1.068586759	1	1	1
Κλοπές σε δημόσιο χώρο-μικροκλοπες	-1.102764476	1	1	1
Κλοπές Τροχοφόρων ΙΧΕ αυτ/των	-0.415891883	1	1	1
Κλοπές Τροχοφόρων ΙΧΦ-Λεωφορείων	-0.129129175	1	1	1
Κλοπές Τροχοφόρων Λοιπών οχημάτων	-0.908639447	1	1	1
Κλοπές Τροχοφόρων Μοτοποδηλάτων	-0.140195506	1	1	1
Κλοπές Τροχοφόρων Μοτοσυκλετών	-0.40161809	1	1	1
Ληστείες εντός καταστημάτων	-0.107976961	1	1	1
Ληστείες εντός οικιών	-0.241229763	1	1	1
Ληστείες κινητών τηλεφώνων-μικροποσών	-0.594428964	1	1	1
Ληστείες λοιπές	-0.066311288	1	1	1
Ληστείες με αρπαγή τσάντας	-0.410647018	1	1	1
Ληστείες οδηγών ταξί	-0.309578729	1	1	1
Ληστείες πρατηρίων υγρών καυσίμων	-0.452608803	1	1	1
Ληστείες σε ΕΛ.ΤΑ.	-0.892475806	1	1	1
Ληστείες σε Μίνι Μάρκετ-κατ/τα ψιλικών	-0.40586498	1	1	1
Ληστείες σε περίπτερα	-0.719413378	1	1	1
Ληστείες σε πρακτορεία ΟΠΑΠ	-1.137647578	1	1	1
Ληστείες σούπερ μάρκετ	-0.137172343	1	1	1
Ληστείες ταχυδρομικών διανομέων	-0.941922214	1	1	1
Ληστείες χρηματαποστολών	-1.268698474	1	1	1
ΑΠΑΤΕΣ	-0.047254065	1	1	1
ΖΩΟΚΛΟΠΗ	-0.904922076	1	1	1
ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	-0.640486349	1	1	1
Average	-0.578971666			

All three algorithms have clustered these crime types together. The normalized average success rate of the police in solving crimes of this class is equal to 0.58 standard deviations below the mean. In that regard, we could argue that police doesn't always prioritize these crimes, or that it is often somehow hard to solve them. Most of these crimes are somehow related to theft. When something is stolen, unless It happens under obvious and often violent conditions, it might not be directly noticeable. Unless the victim needs to access the stolen item in a short period of time after the theft, the crime might not be reported for a very long time, or maybe not at all. When the investigation is delayed, the

probability of police success is reduced significantly. As a result, we could argue that the low success rate of the police when dealing with these crimes might be correlated with the reduced likelihood of observing the commitment of such crimes within a reasonable timeframe to solve the case.

Finally, there are few crime types that are clustered differently from the various algorithms:

ΒΙΑΣΜΟΙ	0.863706212	0	0	-1
ΕΚΒΙΑΣΕΙΣ	0.440543019	1	0	-1
Ληστείες τραπεζών,ταχ/κών ταμιευτηρίων	1.054476503	0	0	-1
Average	0.786241911			

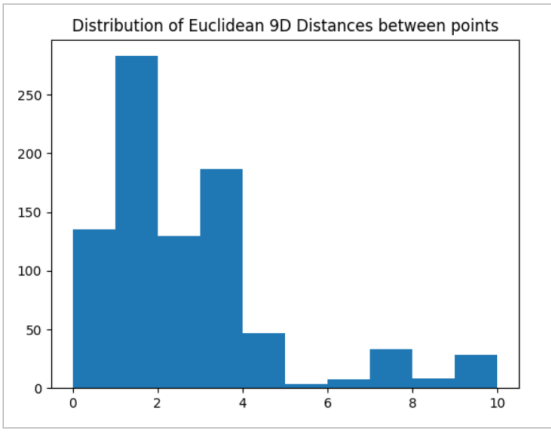
The success rate of the police in solving them lies between the other 2 clusters. While K-means and agglomerative are putting these crimes in either of the 2 clusters, DBSCAN considers them outliers. There is no direct explanation of why these crimes are harder to cluster than the rest, but we will take them into consideration when examining potential outliers.

This first step of the analysis gave us valuable insights on potential crime clusters. However, we expect the next step, which will involve all data set features to give us a more detailed view on the matter.

1.2.2 Cluster the types of crimes and explain what each cluster represents

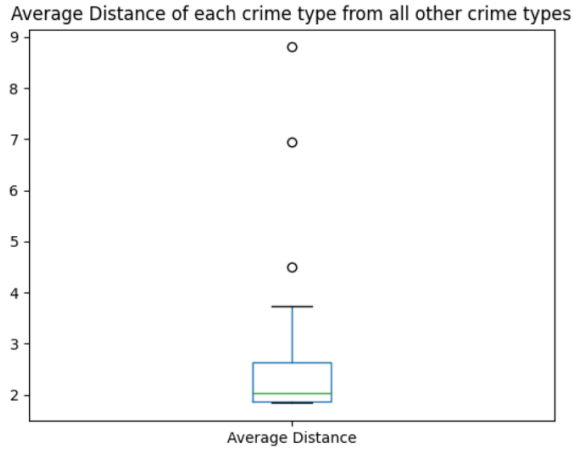
In this section we will repeat what we did in the last section, but we will include all data set features. This will allow us to further elaborate on the different types of crimes. We will still use the silhouette coefficient as the evaluation metric and experiment with the different clustering algorithms. However, we now have 9 features, 5 numerical ('Επιτυχία' feature will be dropped, as it is dependent on the other features) and 4 boolean (created from the one hot encoding of the super class). Therefore, we will not be able to visualize the results and consult our intuition, but rather rely on a grid search and our evaluation metric.

An important decision when clustering in high dimensional spaces is the proper choice of distance measure. The most commonly used measure is the Euclidean distance. However, it is known to deteriorate in higher dimensions. Therefore, in order to examine the suitability of the Euclidean distance for our case study, we will plot the distribution of the distances between points in our 9D space using the Euclidean distance metric. In case the distance distribution is approximating a Pareto distribution where most distances are stacked in a region of extremely high values, then alternative distance metrics will be examined.



$$E(distance) = 2.65, \qquad STD(Distance) = 2.16$$

The distance distribution indicates that the Euclidean distance is still functional in our 9D space. Consequently, it will be the distance metric of choice in our clustering analysis. Moreover, we will create a box plot to represent the average Euclidean distance from each point to all other points:



The above box plot allows for the identification of 3 potential outliers. This values correspond to the following crime types:

Ν περί ΝΑΡΚΩΤΙΚΩΝ: 8.80

Κλοπές - Διαρρήξεις οικιών: 6.94

Κλοπές - Διαρρήξεις από ιχε αυτ/τα: 4.50

We will examine the distances between these potential outliers, in order to get a better understanding of their relative position compared to the rest of the data points.

Distances	Ν περί ΝΑΡΚΩΤΙΚΩΝ	Κλοπές - Διαρρήξεις οικιών	Κλοπές - Διαρρήξεις από ιχε αυτ/τα
Ν περί ΝΑΡΚΩΤΙΚΩΝ	0	8.42	8.94
Κλοπές - Διαρρήξεις οικιών	8.42	0	3.15
Κλοπές - Διαρρήξεις από ιχε αυτ/τα	8.94	3.15	0

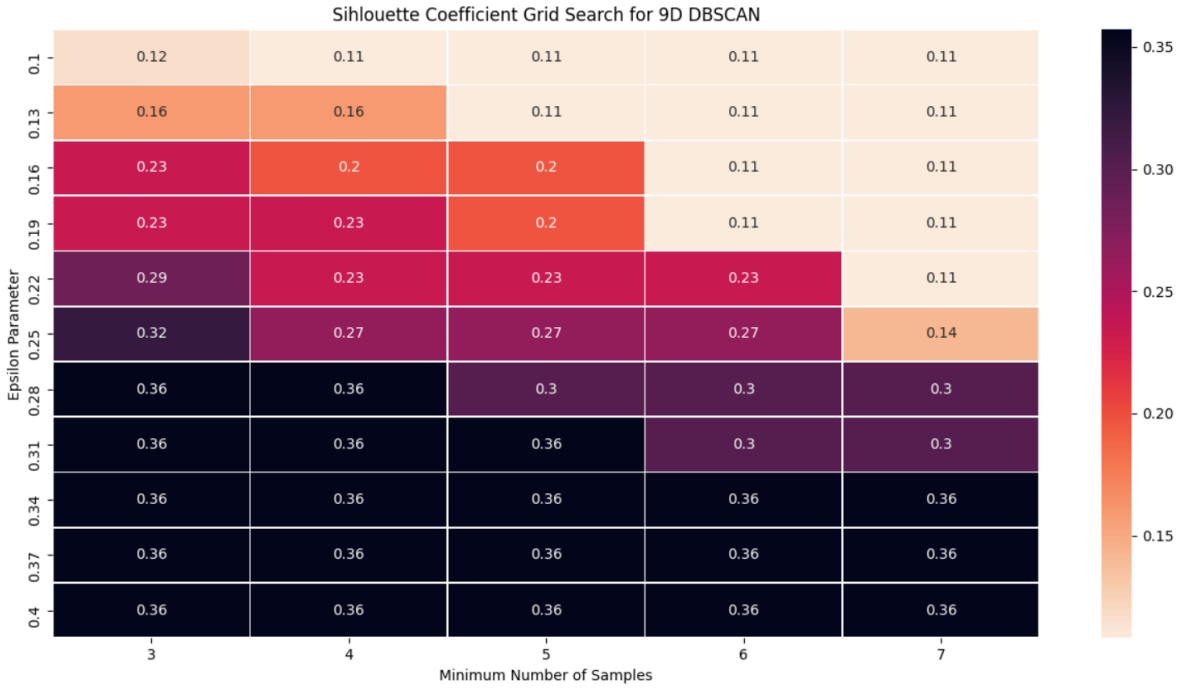
Apparently, the two outliers are located closer to each other, while the 'Ν περί ΝΑΡΚΩΤΙΚΩΝ' outlier is diverging significantly. After determining the Euclidean distance as a reasonable distance metric and briefly exploring the distribution of the points in the 9D space, we will examine different clustering algorithms and evaluate their results based on the value of the Silhouette Coefficient.

K-Means	
Number of Clusters	Average Silhouette Coefficient
2	0.5303
3	0.5093
4	0.4648
5	0.4810
6	0.4403

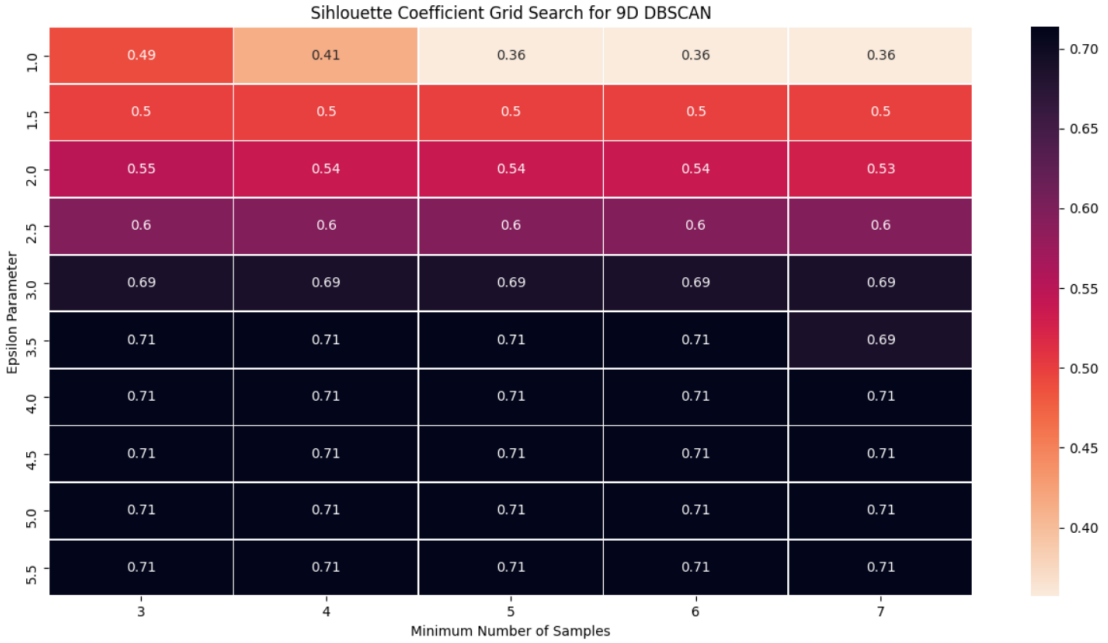
Agglomerative			
Number of Clusters	Average Silhouette Coefficient		
	'complete linkage'	'average linkage'	'single linkage'
2	0.7135	0.7135	0.7135
3	0.6175	0.6434	0.6434
4	0.4526	0.3048	0.3048
5	0.5075	0.3332	0.3332
6	0.4770	0.3750	0.3750

The agglomerative clustering seems to be performing better than the K-means algorithm. Moreover, the ideal number of cluster in both cases is considered to be 2. We will only keep the agglomerative results for 2 clusters and 'complete linkage'.

In order to receive the DBSCAN results in 9 dimensions, we will conduct a new grid search.



For this grid search, we used the same parameter intervals that were used in the 1D space. Given that the silhouette coefficient values are significantly lower, as well as that the values are gradually increasing when the epsilon parameter increases, we suspect that this parameter interval isn't fitting. Distances are gradually increasing in high dimensional spaces and, thus, we will need a higher tolerance for points of the same cluster to be located further away from each other compared to the ones of low dimensional spaces. The small values of epsilon result to the creation of many small clusters that aren't effectively handing the problem. Concluding, we will explore higher values of epsilon:



The silhouette coefficient is performing way better for these values of epsilon. If epsilon becomes larger than 6, all points are put in the same cluster and the silhouette coefficient can no longer be calculated. We will print the clusters that are produced when using the following parameters:

$$epsilon = 4$$

$$minimum\ number\ of\ samples = 4$$

Similarly to the 1D clustering, the set of parameters that corresponds to the maximum value of the silhouette coefficient is deceptive. Almost all points are grouped in a single cluster and a small number are considered outliers. Similar behavior was observed when using DBSCAN with other sets of parameters that correspond to relatively high values of silhouette coefficient, in the range of 0.6 to 0.71. Consequently, selecting the algorithm with the maximum value of silhouette coefficient is ill-advised. We will follow a different approach by comparing different types of algorithms that result to a clustering with similar value of silhouette coefficient. In that regard, we will be able to compare different approaches that observe a similar outcome and draw conclusions about the appropriate form of the clusters. We consider this approach similar to observing low dimensional projections of a high dimensional item and trying to infer information about the item. We used nine different clustering settings, executed in three rounds each consisting of three different algorithms:

Round 1		
KMeans(n_clusters=2)	Agglomerative (n_clusters=3, linkage='average')	DBSCAN(eps=2.5, min_samples=4)

This round resulted to larger clusters with a few outliers.

Round 2		
KMeans(n_clusters=3)	Agglomerative (n_clusters=3, linkage='average')	DBSCAN(eps=1.5, min_samples=4)

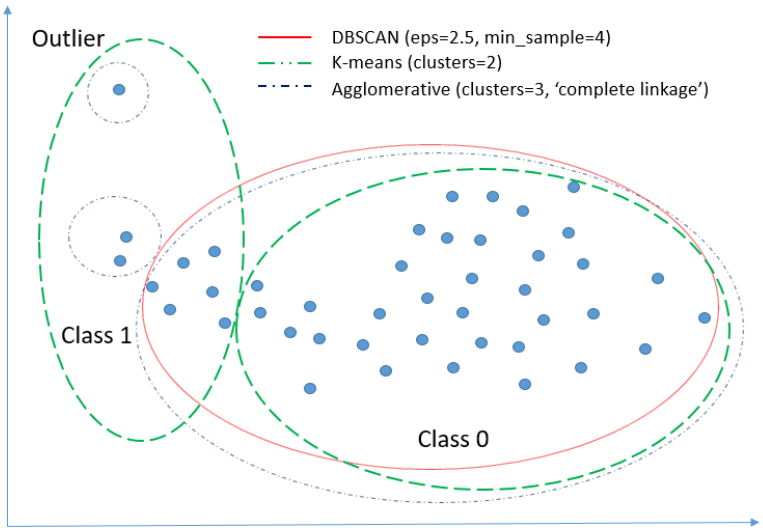
The second round algorithms created more fine clusters than round one.

Round 3		
KMeans(n_clusters=4)	Agglomerative (n_clusters=4, linkage='complete')	DBSCAN(eps=1 , min_samples=4)

The third round algorithms created very fine clusters. It looks like a compromise has to be made between round 2 and round 3.

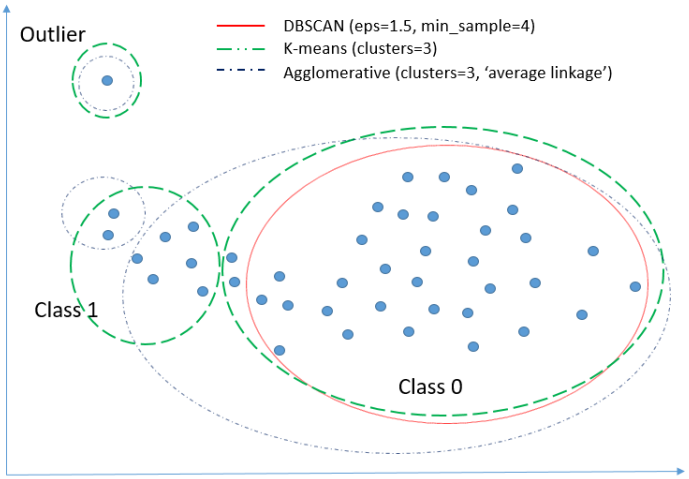
Through observation of results of each algorithm, an intuition regarding the final form of the clusters was built. In order to describe this process, a 2D representation of the data will be used. It's worth noting that this representation wasn't created through dimensionality reduction and doesn't represent the original data. It only serves as a graphic tool to illustrate the insights that were drawn through the clustering iterations. Specifically, we concluded that the data is grouped two clusters. The first one is larger and denser, while the second one is smaller and sparser. Finally, there is a global outlier that was commonly identified by all algorithms. The rationale that led to this conclusion is illustrated through interpretation of the following figures:

Round 1 Algorithms:



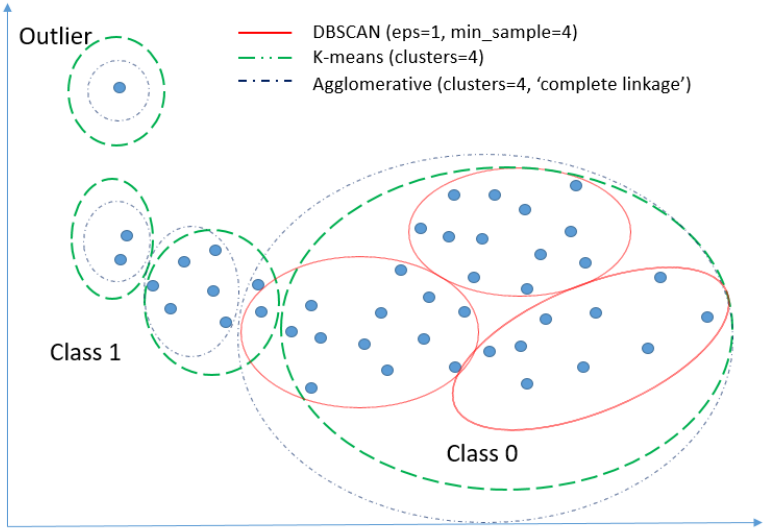
Most of the data points were clustered together by all three algorithms in a single big cluster. We will call this area 'class 0'. DBSCAN (red) considered the rest of the points outliers, while the Agglomerative (dark blue) created 2 small clusters around the area of 'class 1' to handle them. One of these clusters contains two out of the three identified outliers, while the other one contains the seemingly global outlier 'Ν περί ΝΑΡΚΩΤΙΚΩΝ'. Finally, K-means created a different partition, where most of the 'class 0' instances were clustered together, while the rest were grouped in a second cluster together with the 3 potential outliers. At this stage, the global outlier at the top left is put in its own cluster by Agglomerative, while DBSCAN also considers it an outlier along with all other points not contained in 'class 0'.

Round 2 Algorithms:



In this round, the algorithms tend to create more and finer clusters. Most of the instances that were part of 'class 0' in the previous round were still clustered together. K – means created the same clusters as before, but separated the global outlier and put it in its own cluster. This version of DBSCAN with a lower value of epsilon still groups all points in a single cluster in the area of 'class 0' and considers the rest outliers. However, even more points were left out of the original cluster. In that regards, the results of DBSCAN and K-means started to converge. The same version of Agglomerative was used in this iteration for comparison purposes.

Round 3 Algorithms:



In the third and final round, even finer algorithms were used and more clusters were created. K-means still created very similar clusters with the last iteration, but separated the 2 outliers from the rest of the instances in the area of 'class 1' and grouped them in their own cluster. The new version of

Agglomerative gave very similar results to the ones of K-means. The main difference is the grouping of a small number of 'intermediate' points to the cluster of 'class 0' or to the cluster of 'class 1'. Finally, the finer version of DBSCAN identified 3 sub – clusters within the area of 'class 0' and considered the rest of them outliers. After observing the behavior of different algorithms, the following conclusions can be drawn:

- The vast majority of crime types are clustered together by almost all algorithms. We will consider this cluster 'class 0'.
- Most points that aren't clustered in 'class 0' can be clustered together in a different cluster that we will call 'class 1'. There is a number of points that fall in-between can be included in either of the two clusters. These will be examined separately.
- There are 2 outliers, 'Κλοπές - Διαρρήξεις οικιών' and 'Κλοπές - Διαρρήξεις από ιχε αυτ/τα', that are located close to the area of 'class 1' and are eventually clustered together with 'class 1' by some algorithms. We could either consider these collective outliers or contextual outliers of 'class 1'.
- There is one global outlier, 'Ν περί ΝΑΡΚΩΤΙΚΩΝ', that is identified as such by all algorithms. This confirms the exploratory analysis that was conducted earlier.

In order to interpret the meaning of each cluster and the nature of each outlier, we will firstly have a look at the clusters that were formed. We will begin by examining the instances that could be either placed in 'class 0' or 'class 1'. We will apply a voting scheme to decide their cluster. It's worth noting that the 'class -1' assigned by DBSCAN indicates a data point that was identified as an outlier and is not part of 'class 0'. Moreover, the 'class 1' predicted by DBSCAN in the case of 'Κλοπές σε δημόσιο χώρο-μικροκλοπες' indicates a sub cluster of 'class 0'. In that regard, we consider it equivalent to 'class 0' for our voting scheme purposes:

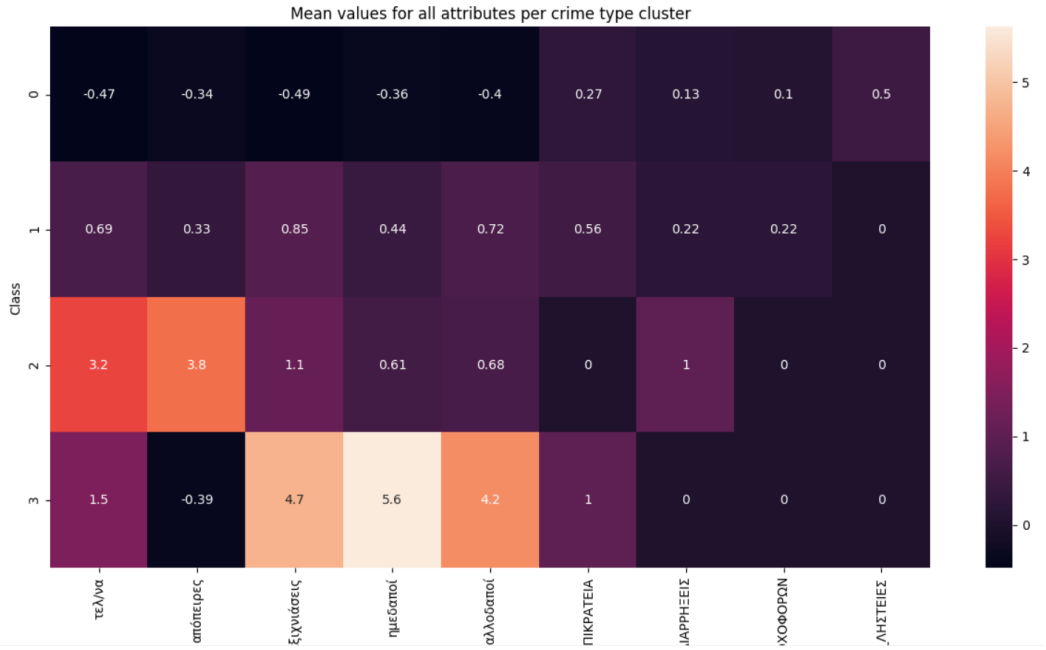
Crime Type	K-Mean_3C	Agglo_3C	DBSCAN_eps1.5	K-Mean_4C	Agglo_4C	DBSCAN_eps1	Voting
ΕΠΑΙΤΕΙΑ	1	0	-1	1	0	-1	1
ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	0	0	-1	1	0	-1	?
ΛΑΘΡΕΜΠΟΡΙΟ	0	0	0	1	0	-1	0
Ν περί ΟΠΛΩΝ	1	0	-1	1	1	-1	1
ΠΛΑΣΤΟΓΡΑΦΙΑ	1	0	-1	1	0	-1	1
Κλοπές σε δημόσιο χώρο-μικροκλοπες	0	0	0	1	0	1 (indicates 0)	0
Κλοπές Τροχοφόρων ΙΧΕ αυτ/των	1	0	-1	1	1	-1	1
Κλοπές Τροχοφόρων Μοτοσυκλετών	1	0	-1	1	1	-1	1

All cases are handled by the voting scheme, except the case of 'ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ', where it results in a draw. We will consider the predictions of the more 'fine' algorithms more important and, therefore, we will assign it to 'class 1'.

We are finally able to represent the different clusters. We have 2 main clusters, 'class 0' and 'class 1', a smaller one that contains the 2 conditional outliers, named 'class 3' and the single global outlier:

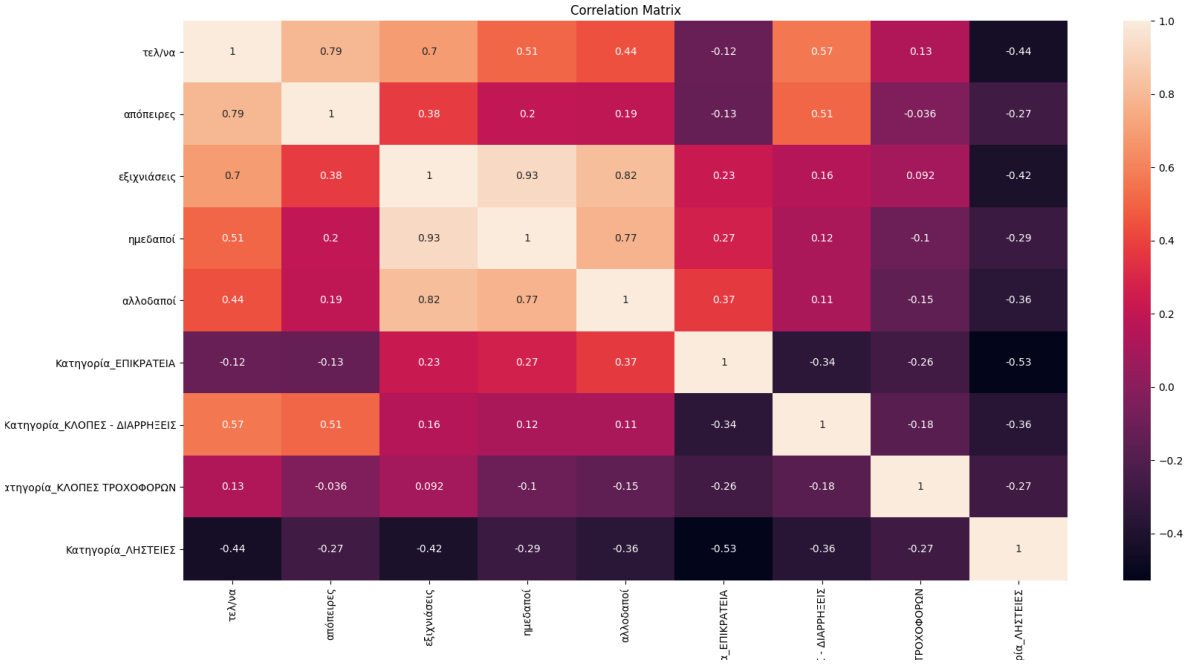
'Class 0'	
ΑΝΘΡΩΠΟΚΤΟΝΙΕΣ	Ληστείες εντός καταστημάτων
ΑΡΧΑΙΟΚΑΠΗΛΕΙΑ	Ληστείες εντός οικιών
ΒΙΑΣΜΟΙ	Ληστείες κινητών τηλεφώνων-μικροποσών
ΕΚΒΙΑΣΕΙΣ	Ληστείες λοιπές
ΖΩΟΚΛΟΠΗ	Ληστείες με αρπαγή τσάντας
Ν περί ΠΝΕΥΜΑΤΙΚΗΣ ΙΔΙΟΚΤΗΣΙΑΣ	Ληστείες οδηγών ταξί
ΣΕΞΟΥΑΛΙΚΗ ΕΚΜΕΤΑΛΛΕΥΣΗ	Ληστείες πρατηρίων υγρών καυσίμων
Κλοπές - Διαρρήξεις ιερών ναών	Ληστείες σε ΕΛ.ΤΑ.
Κλοπές - Διαρρήξεις σε συγκοινωνιακά μέσα	Ληστείες σε Μίνι Μάρκετ-κατ/τα ψιλικών
Κλοπές με αρπαγές τσαντών	Ληστείες σε περίπτερα
Κλοπές σε δημόσιο χώρο-μικροκλοπες	Ληστείες σε πρακτορεία ΟΠΑΠ
Κλοπές Τροχοφόρων ΙΧΦ-Λεωφορείων	Ληστείες σούπερ μάρκετ
Κλοπές Τροχοφόρων Λοιπών οχημάτων	Ληστείες ταχυδρομικών διανομέων
Κλοπές Τροχοφόρων Μοτοποδηλάτων	Ληστείες τραπεζών,ταχ/κών ταμιευτηρίων
ΛΑΘΡΕΜΠΟΡΙΟ	Ληστείες χρηματαποστολών
'Class 1'	
ΑΠΑΤΕΣ	Κλοπές - Διαρρήξεις καταστημάτων
ΕΠΑΙΤΕΙΑ	Κλοπές - Διαρρήξεις λοιπές
ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	Κλοπές Τροχοφόρων ΙΧΕ αυτ/των
Ν περί ΟΠΛΩΝ	Κλοπές Τροχοφόρων Μοτοσυκλετών
ΠΛΑΣΤΟΓΡΑΦΙΑ	
'Class 2'	
Κλοπές - Διαρρήξεις από ιχε αυτ/τα	Κλοπές - Διαρρήξεις οικιών
Global Outlier: 'Class 3'	
Ν περί ΝΑΡΚΩΤΙΚΩΝ	

In order to identify the differences between crimes in each cluster, we will plot the mean value of each attribute for all crimes of the same class in a heatmap:



The crime types of 'class 0' are mainly of type 'ΛΗΣΤΕΙΕΣ' and correspond to crimes, which occur rarely compared to 'class 1'. Moreover, they have a lower number of solved cases, as well as lower number of criminals, both local and foreigners. On the other hand, the instances of 'class 1' correspond primarily to the category 'ΕΠΙΚΡΑΤΕΙΑ' and are characterized by greater mean values than 'class 0' for all attributes. This pattern indicates the existence of a correlation between the features that has to be explored. Finally, the outliers clustered in 'class 2' have the highest values for crime occurrence, while the global outlier is subject to extreme values of solved cases and arrested criminals (which again indicates a correlation).

In order to examine the potential correlations, we will plot the correlation matrix of our data set:



The correlation matrix indicates the existence of some very strong correlations between the features. Intuitively, this does make sense. The attributes 'τελ/να' and 'απόπειρες' are both indicative of the frequency of occurrence of the specific crime type and are thus correlated. Moreover, no matter what the success rate of the police is at solving a specific crime type, more crimes occurred means more cases will be solved. Consequently, the 'εξιχνιάσεις' feature is correlated with the 'τελ/να' feature. Finally, as a result of more cases solved, more criminals, both local and foreigners will be arrested and the respective features are extremely correlated with the 'εξιχνιάσεις' feature. Ideally, this investigation should have been executed at the beginning of the analysis, given that the identified correlations have a huge impact on all clustering results of the 9D clustering that was conducted. However, due to time restrictions at the current time, it is not feasible to go back, create new features, repeat all calculations and alter the report. Although the results of clustering will be based on the correlated data set, we will conduct a feature engineering to handle the intrinsic correlations of the data set in order to improve the results of the classification task.

1.3 Classification

1.3.1 Feature engineering

Firstly, we have to only include one feature that describes the frequency of occurrence of each crime type, namely one of 'τελ/να' and 'απόπειρες'. We will keep 'τελ/να'. The 'απόπειρες' feature will be replaced with the ratio of 'απόπειρες' per 'τελ/να'.

$$\text{'απόπειρες/τελ'} = \frac{\text{'απόπειρες'}}{\text{'τελ/να'}}$$

In that regard, the new feature 'απόπειρες/τελ' will describe the tendency that exists for each crime type to result to an unsuccessful attempt and not to a fully committed crime. Moreover, the 'Επιτυχία' feature will be reintroduced in order to replace 'Εξιχνιάσεις', which only describes the absolute number of solved cases, rather than the percentage of crimes that are actually being solved.

$$Επιτυχία = \frac{\text{εξιχνιάσεις}}{\text{απόπειρα} + \text{ τετελεσμένο}}$$

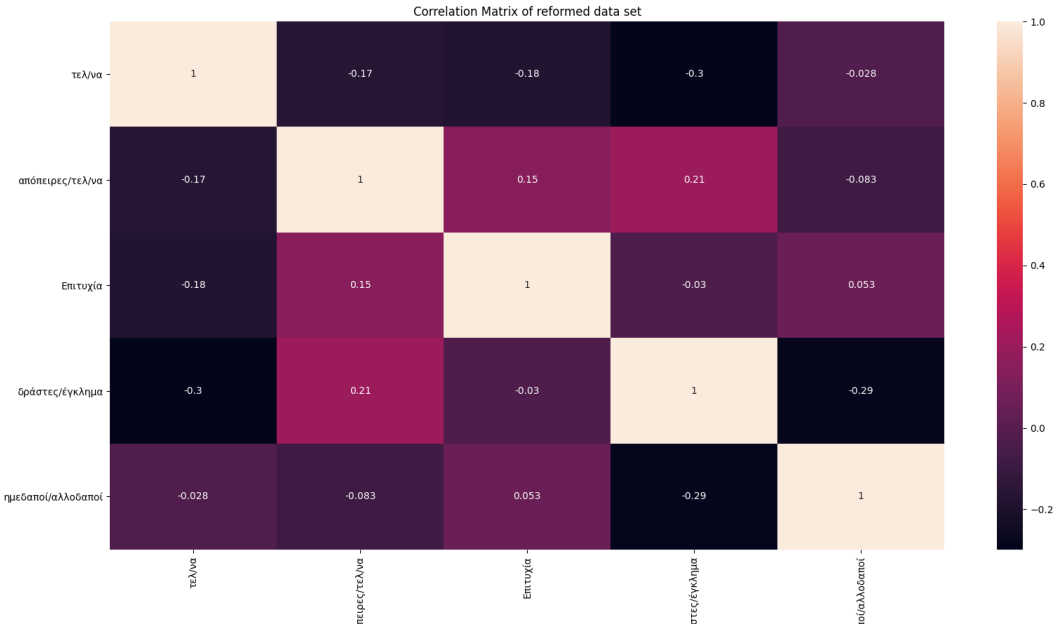
Finally, the attributes 'ημεδαποί' and 'αλλοδαποί' have to be handled appropriately. Instead of absolute values, they will be replaced by 2 appropriate percentage features. We will introduce the 'δράστες/έγκλημα' feature:

$$\text{'δράστες/έγκλημα'} = \frac{\text{'ημεδαποί'} + \text{'αλλοδαποί'}}{\text{'Εξιχνιάσεις'}}$$

This feature describes the average number of criminals involved in each solved case for the specific crime type. The final feature will account for the ratio between local and foreigner criminals:

$$\text{'ημεδαποί/αλλοδαποί'} = \frac{\text{'ημεδαποί'}}{\text{'αλλοδαποί'}}$$

All new features are normalized using Z-normalization and the correlation matrix of the new data set is plotted:



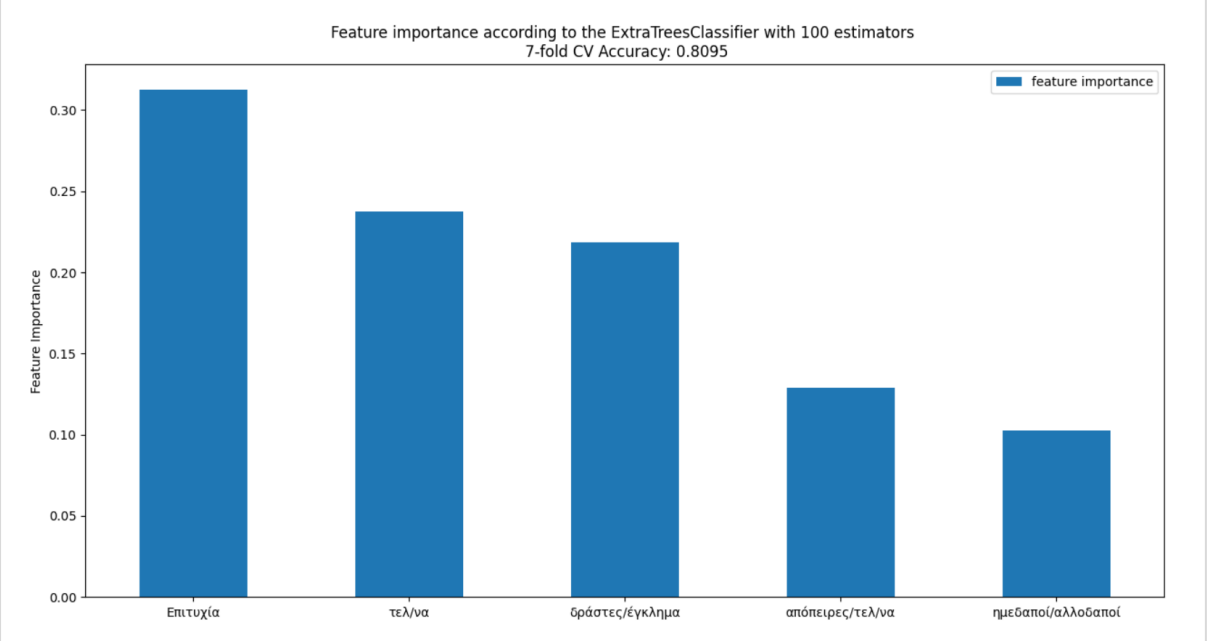
The transformed data set has no intrinsic correlations between its features. This form of the data set will be used for classification purposes. We will also need to store the summary statistics per column of the transformed data set in order to reverse the Z score transformation and interpret the results of our classification algorithms later on:

	τελ/να	απόπειρες/τελ/να	Επιτυχία	δράστες/έγκλημα	ημεδαποί/αλλοδαποί
mean	3352.523	0.127	0.429	1.112	3.240
std	5224.331	0.278	0.308	0.569	2.526

1.3.2 Predict the Superclass

In our classification task, we want to be able to provide an explanation along the results of our classification algorithm. In that regard, there is no need to conduct an extensive search over different Machine Learning algorithms, since most of them operate in a black – box fashion. We will be using the ExtraTree algorithm along with a simple Decision Tree Classifier. Firstly, the ExtraTree algorithm will be trained to give us an insight on the importance of each specific feature. Then, a grid search will be conducted to identify the best hyperparameters for our Decision Tree. Both classifiers will be trained using a cross validation schema. However, our data set size is very small (42 instances) and, therefore, we won't be able to split our data in training and test set, while doing corss validation only on the training set. We will use no test set at all. We will execute a 7 – fold cross – validation, where each time the model will train on 36 instances and validate its accuracy on the remaining 6 (42 instances = 7 folds * 6 instances). This might create some problems with predicting the class3 ('ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ'), which is represented only by five instances, a number smaller than the number of folds.

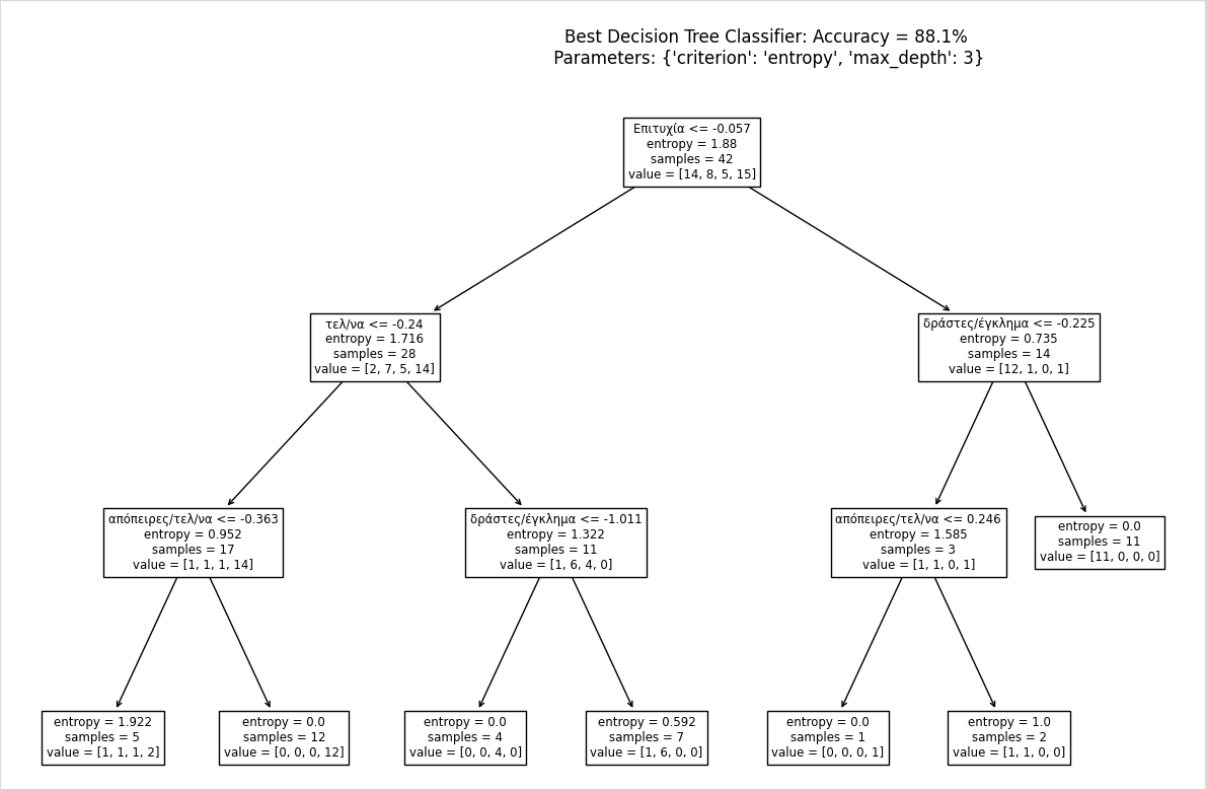
The average accuracy of the 7 – fold cross validation using an ExtraTrees classifier with 100 estimators was roughly 81%, while the feature importance is illustrated in the figure below:



A quick interpretation of our results is that the most important feature to predict the superclass of the crime is the success rate of the police in solving it ('Επιτυχία'). Apparently, the results of the 1D

clustering of the crimes based on this single feature might be representing different superclasses, at least to an extent. Moreover, the second most important aspect is the frequency of occurrence of each crime type (indicated by the 'τελ/να' feature) followed by the average number of criminals that were arrested in a solved case (δράστες/έγκλημα).

As a next step, a grid search was conducted to identify the best performing decision tree classifier for our dataset using the same 7 – fold cross validation scheme. Both the 'entropy' and the 'Gini index' were investigated, as well as potential maximum tree depth in the range [2,15]. It turned out that the best performing classifier, with an accuracy over 88%, is a decision tree using the 'entropy' split criterion with maximum_depth = 3. A graphical representation of the tree was created:



In order to extract some rules, we will focus on the leaves with the highest purity and a significant number of predictions. Moreover, in order for our rules to be interpretable, we will reverse the Z score transformation that was introduced earlier using the formula:

$$x = z * \sigma + \mu$$

Rule1:

If ('Επιτυχία > 41.14%) AND ('δράστες/έγκλημα' > 0.98) => superclass = 'ΕΠΙΚΡΑΤΕΙΑ'

Rule2:

If ('Επιτυχία <= 41.14%)

AND ('τελ/να'> 2099)

AND ('δράστες/έγκλημα' > 0.54) => superclass = mostly 'ΚΛΟΠΕΣ – ΔΙΑΡΡΗΞΕΙΣ'

Rule3:

If ('Επιτυχία <= 41.14%)

AND ('τελ/να'> 2099)

AND ('δράστες/έγκλημα' <= 0.54) => superclass = 'ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ'

Rule4:

If ('Επιτυχία <= 41.14%)

AND ('τελ/να'<= 2099)

AND ('απόπειρες/τελ/να' > 2.6%) => superclass = 'ΛΗΣΤΕΙΕΣ'

We can summarize our results as follows. If the success rate of the police in dealing with a specific crime type is higher than 41.14%, then it's probably not associated with theft, but it's a rather more generic crime type and is classified as 'ΕΠΙΚΡΑΤΕΙΑ'. On the contrary, most crimes that are harder to solve (success < 44.14%) are associated with theft and are members of the other 3 superclasses. This insight is in line with the results of 1D clustering, as well as the ExtraTrees classifier that identified the success rate of the police as the most dominant feature. We will try to describe the characteristics of each crime superclass based on the extracted rules:

Superclass1 - 'ΕΠΙΚΡΑΤΕΙΑ':

Crime types in this super class fall in the cluster of high police success rate in solving them. Moreover, the average number of criminals per solved case is equal or greater than one, specifically the average for all crime types of this superclass equals to 1.21 criminals per solved case. This indicates that different crimes are committed by different individual criminals or groups of criminals. It's rarer that the same criminal is involved in more than 1 cases compared to the other superclasses.

Superclasses2,3,4: - 'ΚΛΟΠΕΣ – ΔΙΑΡΡΗΞΕΙΣ', 'ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ', 'ΛΗΣΤΕΙΕΣ':

Although all these superclasses fall in the low police success rate cluster, the decision tree classifier was able to spot differences between them using the rest of the features. 'ΛΗΣΤΕΙΕΣ' are committed less often than the other two superclasses (lower 'τελ/να') and the decision boundary is 2100 committed crimes per year. Moreover, 'ΛΗΣΤΕΙΕΣ' tend to result to an 'attempted crime' instead of a 'fully committed' one more often than the other two superclasses. Specifically, 12.12% of 'ΛΗΣΤΕΙΕΣ' end up as 'attempts', while only 7.5% of 'ΚΛΟΠΕΣ – ΔΙΑΡΡΗΞΕΙΣ' and 2.27% of 'ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ'. Of course, this insight is related to the interpretation of the 1D clustering results, where it was noted that 'attempts' of thefts that do not involve violence might not be noticed and recorded at all. Finally, 'ΚΛΟΠΕΣ – ΔΙΑΡΡΗΞΕΙΣ' are separated from 'ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ' due to their higher values in the feature 'δράστες/έγκλημα', while the decision boundary is the value 0.54. When having a closer look, an average of 1.37 criminals are involved in crimes of the 'ΚΛΟΠΕΣ – ΔΙΑΡΡΗΞΕΙΣ' superclass, indicating that these types of crimes are often committed by groups of criminals. On the other hand, the average number of criminals per solved case for the 'ΚΛΟΠΕΣ ΤΡΟΧΟΦΟΡΩΝ' superclass is equal to 0.31. This means that the average criminal arrested for crimes of this class is involved in 3.23 different cases. It's worth noting that the 'Κλοπές Τροχοφόρων ΙΧΕ

αυτ/των' crime type is characterised by 0.155 criminals per case meaning that each criminal is involved in 6.43 different cases on average!

1.4 Code

All calculations that were described in this section were implemented using python scripts in the PyCharm environment. Specifically, the 3 different scripts that were created and their functionality will be briefly explained. It's important to note that the original excel file was altered slightly manually. In order to reproduce the results, all python files should read the excel file that can be found in the 'code/police' folder of the zip file.

- 1D_clustering.py: Imports the dataset, executes some basic preprocessing, introduces the 'Police Success Rate' column and plots the Hopkings Statistic graph. Moreover, it conducts 1D clustering with K-means and Agglomerative clustering and plots the results for different parameters. Finally, it executes a grid search for the parameters of DBscan and plots the clustering results of two different parameter setting for DBscan.
- 9D_clustering.py: Imports the dataset, executes some basic preprocessing and plots the distribution of Euclidian distances between the instances of the data set in the 9D space. Then, a similar approach with the 1D clustering is followed, the different parameter settings of different clustering algorithms are calculated and printed, along with a grid search for the silhouette coefficient for DBScan. Finally, a graph of the mean values of all features per cluster is printed along with the correlation matrix of the data set.
- Classification.py: The data set is imported once again. However, it is used to create new, uncorrelated features. After some preprocessing, the correlation matrix of the reformed data set is printed. Furthermore, an ExtraTrees classifier is trained and used to estimate the importance of the different features, which is illustrated using a bar chart. Finally, a grid search is conducted to identify the best performing decision tree classifier and its parameters are printed along with a graphical representation of the tree nodes.

All three python files, as well as the excel file with the data and a 'README.txt' file with the required python libraries can be found in the 'code/police' folder of the zip file. For reproducibility of results, it's recommended to create a virtual environment, install the necessary libraries and make sure that the excel file is present (as noted earlier, it's slightly altered compared to the original data set) before running the scripts.