University of Granada

Business Intelligence

Practice 2: Visualization and Segmentation

Author:
Mikhail Raudin
Practice Group 1
mraudin@correo.ugr.es

Lecturer:
Dr. Daniel Molina
dmolinac@ugr.es

December 9, 2020

Contents

List of Figures

1 Introduction

The goal of this practice is to analyze the performance of three different clustering techniques and find groups in the fatal traffic accidents dataset of Spain[1], published by the Directorate General of Traffic (DGT). The dataset has more than 30 variables among the years 2008 and 2015. We will focus on the data for the year 2013 (89,519 accidents).

The clustering algorithms applied are

- K-Means
- DBSCAN

We want to understand the dynamics of the traffic accidents in Spain in that year. To do this, based on various attributes that characterize the accident, we intend to find groups of similar accidents and causal relationships that explain types and severity of accidents. Two different case studies are carried out, considering different attributes and a specific value subset for each one of them.

2 Case Study 1: What types of accidents happen on highways on the weekend?

In this section the obtained clusters by K-Means and DBSCAN on the subset containing only accidents on highways will be evaluated, visualized and interpreted. Afterwards, in section 3, the results are compared to the clusters found on conventional roads during the weekend.

2.1 Case Description

On the weekends there is more traffic on the highways. For example, people are driving to their home towns or doing excursions. We want to find out what type of accidents occur during that time and also if this case affects the gravity of the accidents. Therefore, we select the subset of all data that has the value "AUTOVÍA" or "AUTOPISTA" in the TIPO VIA (ROAD TYPE) column and the day 5,6 or 7 in the DIASEMANA (DAY OF THE WEEK) column. The attributes of interest, forming the clusters, are the following:

- total number of victims (TOT VICTIMAS)
- total number of deaths (TOT MUERTOS)
- total number of seriously injured persons (TOT HERIDOS GRAVES)
- total number of minor injured persons (TOT HERIDOS LEVES)

[1] `https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces`

Furthermore, we can observe in fig. 1 that the total number of victims differs on the weekend more then during the week. Hence, we reduce the variable DIASEMANA to the subset covering the weekend (day 5, 6 and 7). Concerning the road type, there is no major difference in the data distribution between highways and conventional roads (fig. 2). All types of roads have outliers with values for total victims not higher then 20, but the conventional road has one above 30 and the highway one above 50. Therefore, it makes sense to compare the gravity of accidents happening on highways vs. conventional roads on the weekend (the time with more variety in the number of victims). In other words, we keep the weekend as the base subset and analyze what different clusters exist considering other variables (road type) during that time.

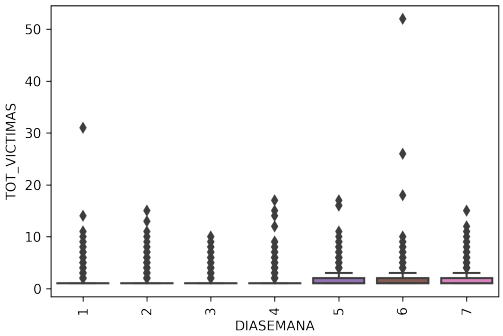

Figure 1: Shows the distribution of the total number of victims over different values for the variable "day of the week" in a box plot. The black points represent the calculated outlier values.

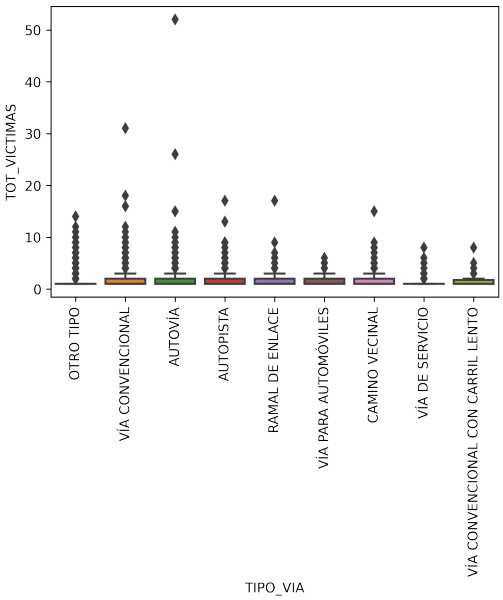

Figure 2: Shows the distribution of the total number of victims over different values for the variable "road type" in a box plot.

Because we are using K-Means, a distance based clustering algorithm, we need to normalize the variables of interest mentioned before, so they all have values between 0-1, to not falsify the real distance between the samples and thus having wrong results.

2.2 Results

The clustering algorithms are evaluated by two metrics. First, the mean Silhouette score for each sample, which measures how similar the sample is to its own cluster as to other clusters. Second, the Calinski-Harabasz Score, which is defined as the ratio between the inter-cluster dispersion and the between-cluster dispersion. Table 3 shows these values for both algorithms. We see that both algorithms created dense and well separated clusters, given the high Silhouette score of 0.86 (the values of the score are distributed between -1 and 1, 1 being the best). However, the K-Means algorithm achieved a better result in the Calinski-Harabasz score. The execution time is really short for both algorithms.

To achieve a good result with the DBSCAN algorithm, different values for the parameters are tried out, because the default configuration leads to a single cluster. So, to increase the number of clusters, the minimum samples parameter is increased from 5 to 50 , having the best result for 20. This produces too many clusters, so the epsilon parameter is decreased from 1 to values between 0.1 and

0.5, creating the optimal result for 0.2. The final configuration for this case study is: epsilon 0.2, minimum samples 20.

Table 1: This table compares the quality of the resulting clusters by K-Means and DBSCAN, using the Silhouette score and the Calinski-Harabasz Score.

algorithm	Silhouette score	Calinski-Harbasz Score	execution time
K-Means (K=3)	0.8645	5225.81	0.1019 s
K-Means (K=12)	0.8674	9719.57	0.1909 s
DBSCAN	0.8637	3730.87	0.5707 s

2.3 Cluster Visualization and Interpretation

We will take a close look at the clusters, by reviewing the sizes, the cluster centers and the distribution over 2 of the attributes of interest at a time using a pair plot.

K-Means

First, we can see in fig. 3 that the clusters are very unbalanced, although well separated (as the scores indicate in table 3). This means that most of the accidents are in cluster 0, some in cluster 1 and almost none in cluster 2.

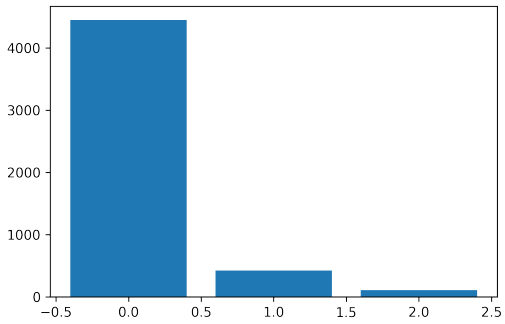

Figure 3: Shows the amount of samples in each cluster found by K-Means (K=3). Cluster 0: 4.447, 1: 425, 2: 109.

To understand what characteristics each cluster has, fig. 11 shows the center of each cluster, and we see that they are semantically well separated. Cluster 0 represents the majority of the accidents, which had on average 1,6 victims and 1,6 minor injured persons involved, with no deaths neither serious injured

persons. Cluster 1 covers the accidents which caused serious injuries, having the highest value of all three clusters in that attribute. Those accidents had also the highest number of total victims (2,224). The last and smallest cluster contains the most fatal accidents, that lead to dead victims. The number of victims and of seriously injured is higher then in cluster 0, although the latter only by 0,24. The number of minor injured is the smallest of all clusters, because victims of these severe accidents mostly end up dead or seriously injured.

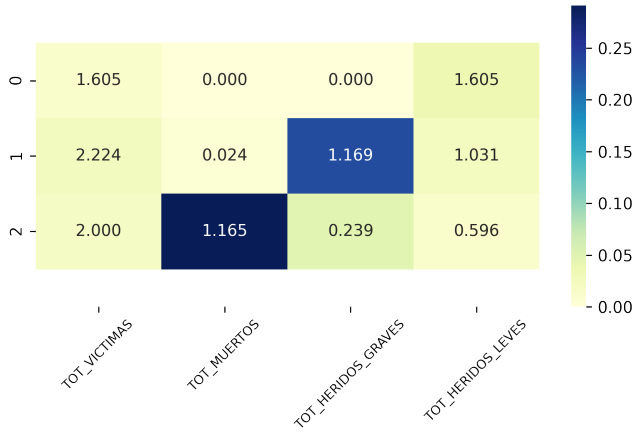

Figure 4: Shows the centers of each clusters, that is, the mean value of all samples within a cluster for each attribute of interest. The clusters are found by K-Means(K=3).

Finally, in fig. 5 we can see the cluster distribution along two variables of interest on the normalized data. We refer to each plot as plot(row, column), plot(0,0) being top left and plot(3,3) bottom right. The figure shows that the number of deaths and seriously injured both determine the cluster membership very well, having almost no intersection between the different clusters. Also the plot(2,1) and plot(1,3) having those two variables on the x and y axis shows clear and separable clusters. The plot(0,1) with the number of deaths on the x axis and the total victims on the y axis shows that for cluster 0 samples, which all lay on x=0 (deaths), there is some interference with cluster 1 samples. Logically, the more victims an accident has, the more likely it is that one of them is seriously injured, thus belonging to cluster 1.

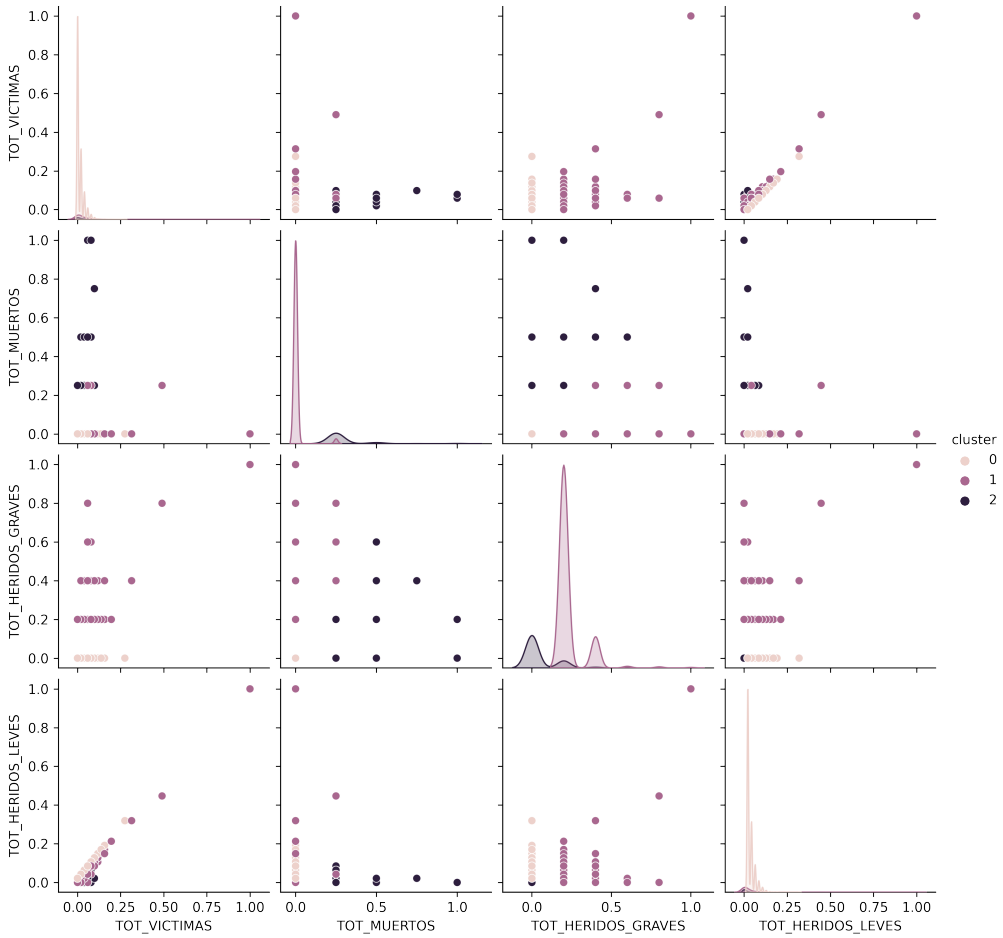

Figure 5: The figure shows a matrix of plots describing the distribution of the three clusters found by K-Means (K=3). Every plot contains one variable of interest on each axis, describing the gravity of an accident, and every point (sample) has the color of the cluster it belongs to (see legend).

DBSCAN

First, we can see in fig. 6 that the five clusters founds are very unbalanced, although well separated (as the scores indicate in table 3). The DBSCAN algorithm finds the same big dominating cluster with 4.447 samples as the K-Means algorithm. The small clusters 1 and 2 of the K-Means algorithm are further divided in totally 4 clusters by the DBSCAN, of which the core one remains big with 363 samples (vs. 425 by K-Means).

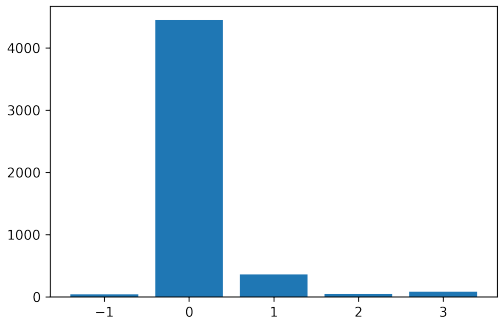

Figure 6: The figure shows the amount of samples in each cluster found by the DBSCAN algorithm. Cluster 0: 4.447, 1: 363, -1: 43, 2: 48, 3: 80.

Fig. 7 shows the center of each cluster, and we see that they are well separated and different from one another. Cluster -1 has 43 elements and covers most of the fatal accidents, having the highest values for every variable, also being higher then the highest center values of the K-Means result. Cluster 3 has one dead person as a mean for the cluster. It seems that DBSCAN divided the cluster of fatal accidents found by K-Means into two clusters(-1 and 3). The clusters 0,1 and 2 have all 0 deaths, but differ in the number of total victims and the seriously injured persons. Cluster 2 has almost double the victims of cluster 1 and double as much seriously injured. Again it seems that DBSCAN divided what was cluster 1 found by K-Means (accidents with heavily injured persons but no deaths) into two clusters(1 and 2), although it filtered out the accidents leading to death (0,239 for cluster 1 in K-Means) into cluster 3. Cluster 0 is the same as cluster 0 from K-Means, describing the minor accidents that had no deaths neither seriously injured persons.

Finally, in fig. 8 we can see the cluster distribution along two variables of interest. The clusters are more intersected then the ones produced by K-Means, and looking just at the number of death or seriously injured people is not enough. Instead, the clusters are better separated by a combination of variables, namely the plots that put the seriously injured persons in relation to the minor injured persons. Also, as with K-Means, the combination of total number of deaths and seriously injured in one plot produces a good separation.

cluster	TOT_VICTIMAS	TOT_MUERTOS	TOT_HERIDOS_GRAVES	TOT_HERIDOS_LEVES
-1	5.488372	1.325581	1.488372	2.674419
0	1.605352	0.000000	0.000000	1.605352
1	1.801653	0.000000	1.000000	0.801653
2	3.166667	0.000000	2.000000	1.166667
3	1.512500	1.000000	0.000000	0.512500

Figure 7: The matrix shows the centers of each clusters, that is, the mean value of all samples within a cluster (y-axis) for each attribute of interest (x-axis). The clusters are found by the DBSCAN algorithm.

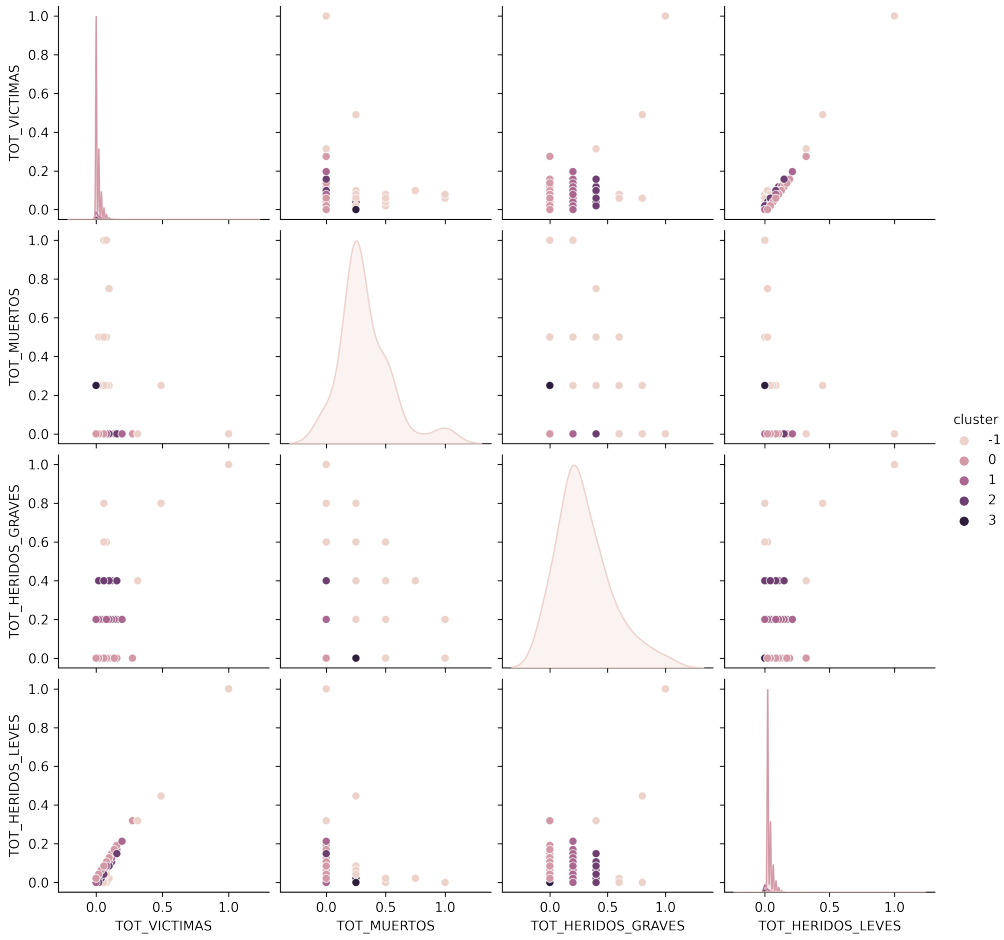

Figure 8: The figure shows a matrix of plots describing the distribution of the 6 clusters (found by DBSCAN algorithm) among the variables of interest.

2.4 K-Means: effect of the parameter k

Using the K-Means algorithm, we define beforehand the number of clusters (parameter k) produced by the algorithm. If we can't tell how many clusters the dataset might have, then it's difficult to guess the right parameter k. The accidents dataset is very large (89,519 accidents) and has many parameters, so we can't assume an appropriate value for k easily. We will use two different methods to explore different outcomes of K-Means when increasing the value of k from 2 until 11.

Elbow Method

In each iteration, we calculate the inertia, which is the sum of squared distances of samples to their closest cluster center[2]. It tells us how internally coherent the clusters are. In fig. 9 we can see that the inertia decreases when k increases, because the distance within a cluster gets smaller as we create more clusters, and thus more cluster centers. The point which indicates the best trade-off between error and number of clusters is called elbow point. It is where the inertia starts decreasing in a linear way. For our graph we locate the elbow point at k=6.

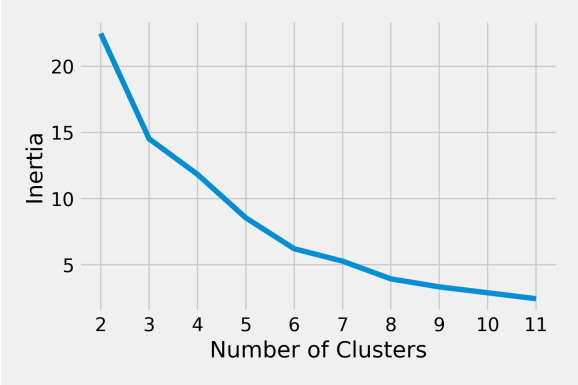

Figure 9: The graph plots the inertia of the clusters found by K-Means for different values of k. The elbow point is located at k=6.

Silhouette Coefficient

The Silhouette Coefficient is also one of our two evaluation metrics, so it makes sense to optimize the value for k by this measure. We can observe in fig.10 that the Silhouette Coefficient is high for k=3, but even higher for k=9. Trying to keep the number of clusters small, and given a higher importance to the Silhouette Coefficient then inertia, k=3 is chosen for the evaluation.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

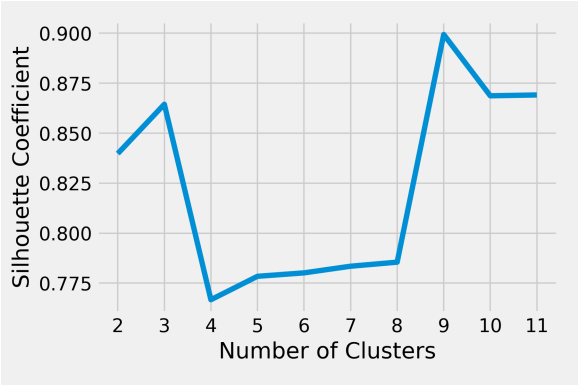

Figure 10: The graph plots the Silhouette Coefficient of the clusters found by K-Means for different values of k. The highest scores are found for k=3 and k=9.

Out of curiosity, we discuss the results for k=9 to gain interesting knowledge about the case study. The table 2 shows that for k=9 our evaluation metrics are better, and therefore the quality of the clusters. Fig.**??** shows that the clusters are still very different from one another, but their characteristics are similar to the ones obtained with k=3. The result is more refined, cluster 1,3 and 8 corresponding to cluster 0 of k=3 (minor accidents), cluster 0 and 4 to cluster 1 (accidents with seriously injured persons). Cluster 6 has only two samples, it averages two accidents that had a lot of victims involved, way more then the usual number. These are two outlier points representing extreme accidents with a lot of cars involved. Cluster 7, with 3 samples, describes the accidents with more then double the average of death people (compared to cluster 2 from k=3), being fatal car crashes. To conclude, K-Means with k=9 clusters provides us with knowledge about the dynamics of car crashes on the highways during the weekend, but it creates some very small clusters due to the outlying big crashes. In effect, we prefer the result with k=3, because the clusters are bigger and it is a more understandable and interpretable result.

Table 2: This table compares the quality of the resulting clusters by K-Means for the best values of k, using the Silhouette score and the Calinski-Harabasz Score.

algorithm	Silhouette score	Calinski-Harbasz score
K-Means (K=3)	0.8645	5225.8144
K-Means (K=9)	0.8994	7776.5024

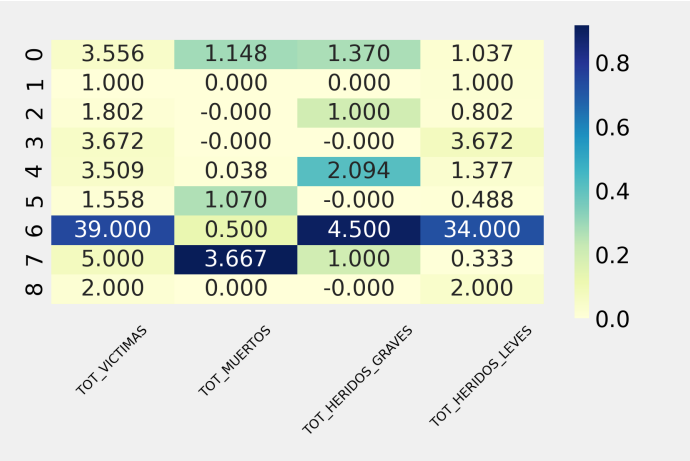

Figure 11: Shows the centers of each clusters found by K-Means (K=9).

2.5 Conclusion

We could see that there are three types of accidents happening on highways on the weekend. Most of all, minor accidents with a few minor injured persons happen. Moreover, accidents with seriously injured persons or deaths happen a lot, 534 times in 2013, more then one accident each day (1,46 per day). But compared to the first group (4447 samples), it is the clear minority in this case study.

Further, observing the results of k=9 we can say that k=3 loses a lot of information about the underlying car crashes in the three clusters, due to the outliers far from the cluster center. We could observe with k=9, that 2 huge accidents happened in 2013, with a lot of cars involved. The DBSCAN results are more subtle to the different accidents and the algorithm creates more clusters, but having in the end less quality then the K-Means k=3 results, due to the lower Calinski-Harabasz Score.

Both algorithms find semantically similar clusters. Also, both algorithms find a huge cluster and some other clusters that are all very small. Therefore the selected data for this case study does not cluster well, or, we can assume that many accidents are very similar.

3 Case Study 1.2: What types of accidents happen on conventional roads on the weekend?

In this section the case study 1 is continued and the obtained clusters by K-Means and DBSCAN on the contrary subset of case study 1 will be evaluated, visualized and interpreted. Thus, the clusters of accidents on conventional roads

during the weekend are compared to the clusters of accidents on highways during the weekend (Section 2). The case and decision to compare accidents on conventional roads vs. highways is described in detail in Section 2.1.

3.1 Results

Table 3 shows the score values for both algorithms. We see that the K-Means algorithm created dense and well separated clusters, given the high Silhouette score of 0.8459 and very high Calinski-Harabasz score. The result of the DBSCAN algorithm is clearly worse, as shown in the table. The final configuration of DBSCAN for this subset is: epsilon 0.1, minimum samples 80.

Table 3: This table compares the quality of the resulting clusters by K-Means and DBSCAN, using the Silhouette score and the Calinski-Harabasz Score.

algorithm	Silhouette score	Calinski-Harbasz score
K-Means (K=5)	0.8459	10369.97
DBSCAN	0.6828	2922.93

3.2 Cluster Visualization and Interpretation

We will take a close look at the clusters, by reviewing the sizes, the cluster centers and the distribution over 2 of the attributes of interest at a time using a pair plot.

K-Means

First, we can see in fig.12 that the clusters are unbalanced, although well separated. This means that most of the accidents are in cluster 1 (6878), as well as cluster 0 (2638) and 2 (1673), and the minority in cluster 3 and 4. The clusters are better balanced then the result from the highway subset, having 3 clusters larger then 1.500 samples.

Fig.13 shows the center of each cluster, and we see that they are semantically well separated. Cluster 1 represents the majority of the accidents, which had on average 1 victim and 1 minor injured person involved, with no deaths neither serious injured persons. Cluster 0 represents accidents with similar consequences as cluster 1, that is, only minor injured persons. It describes the accidents with more victims (2,252) - of which a lot receive minor injuries (2,252 as well). Cluster 2 covers the accidents which caused serious injuries, having the highest value of all three clusters in that attribute. Those accidents have a small number of total victims (1,524). Cluster 3 describes accidents where the highest amount of persons suffer. Almost 5 victims in general (4,827) and minor injured persons, but in contrast to cluster 0 and 1 it has slightly higher values for the serious

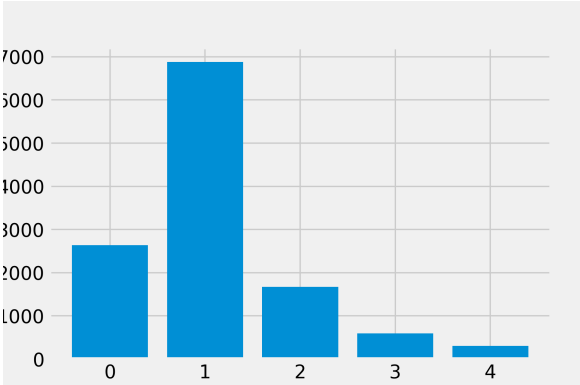

Figure 12: Shows the amount of samples in each cluster found by K-Means (K=5). Cluster 0: 2638, 1: 6878, 2: 1673, 3: 694 4: 302

injured and death persons, meaning the accidents in cluster 3 are more severe. The last and smallest cluster contains the most fatal accidents, that lead to dead victims.

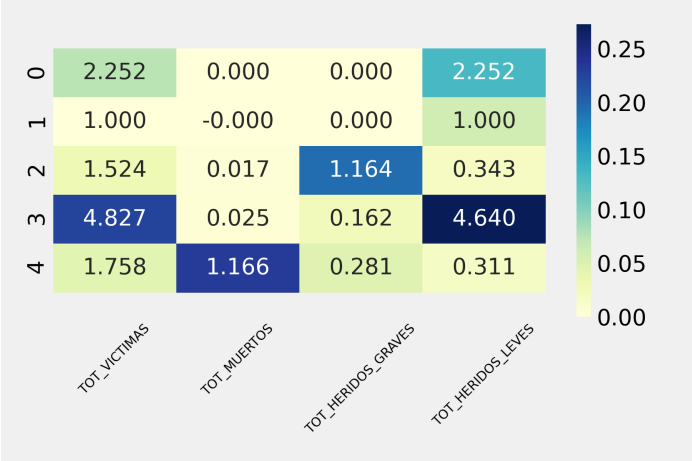

Figure 13: Shows the centers of each cluster, that is, the mean value of all samples within a cluster for each attribute of interest. The clusters are found by K-Means (K=5).

Finally, in fig. 14 we can see the cluster distribution along two variables of interest on the normalized data. The figure shows that no variable separates the clusters on its own (diagonal axis). Good separations can be observed with the variables "total seriously injured" and "total deaths" (1,2), "total seriously

injured" with "total minor injured" (3,2). Interestingly, none of the accidents with more then 10 victims (see first row in plot), belongs to the deadly cluster 4.

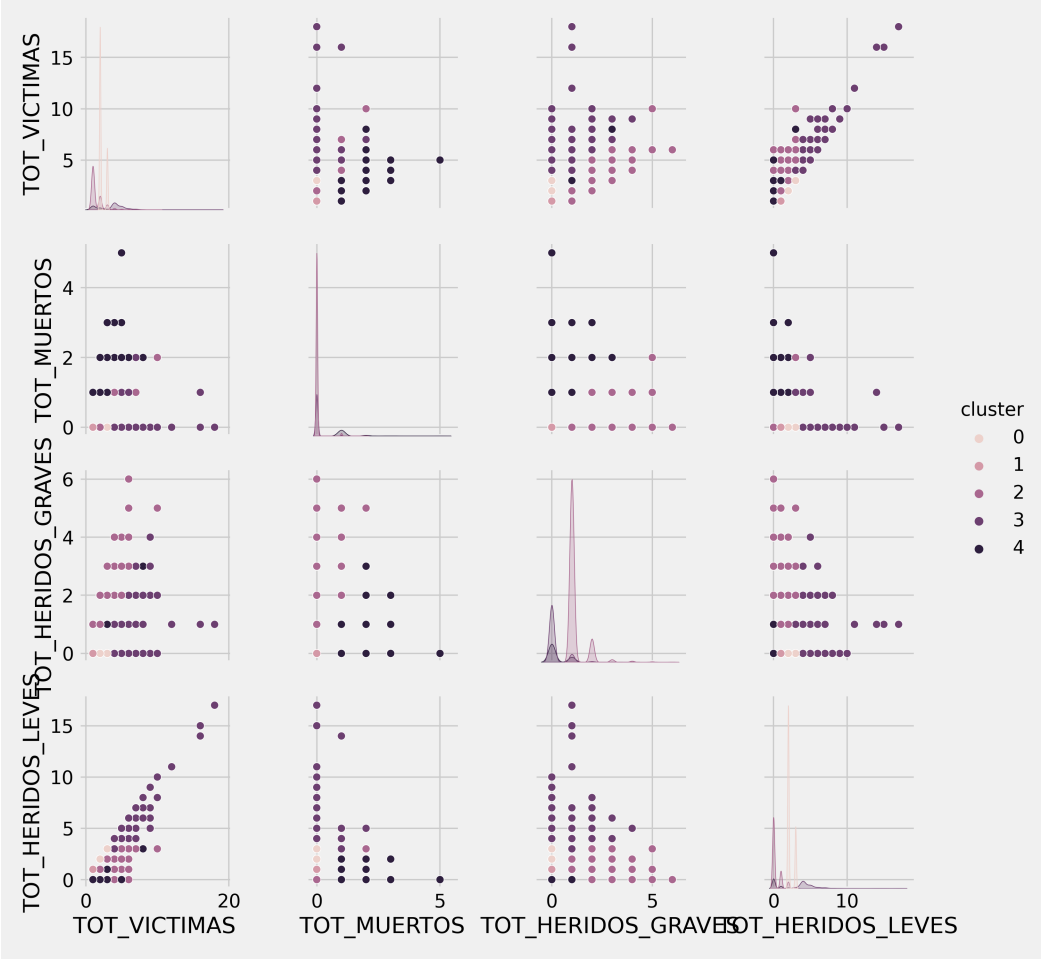

Figure 14: The figure shows a matrix of plots describing the distribution of the five clusters found by K-Means (K=5) on the conventional roads subset.

DBSCAN

First, we can see in fig. 15 that the five clusters found are very unbalanced. The DBSCAN algorithm finds a big cluster with 10.008 samples, one with 1508 and three small ones. Having the same number of clusters as with K-Means, but different sizes, we conclude that the clusters must be semantically different

from the K-Means clusters.

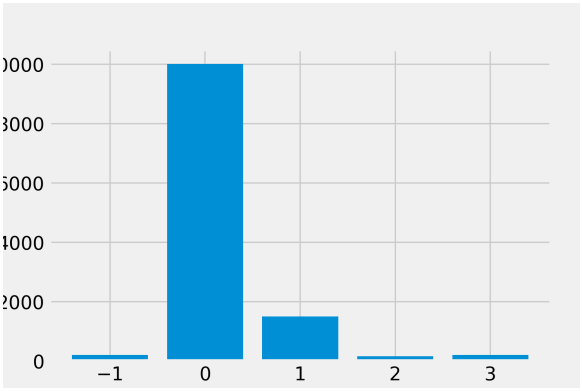

Figure 15: The figure shows the amount of samples in each cluster found by the DBSCAN algorithm. Cluster 0: 10.008, 1: 1508, -1: 204, 2: 157, 3: 208.

Fig. 16 shows the center of each cluster, and we see that they are different from one another. The first observation is that in contrast to K-Means, DBSCAN finds two different cluster with deadly accidents. The first, cluster -1, has the most victims involved (mean value 4,5), but has more seriously injured victims and less minor injured victims then the cluster with the most victims involved (4,8) of K-Means. So cluster -1 of DBSCAN conveys a different characteristic: accidents with most victims have more severe consequences (deaths and serious injuries) then the cluster of K-Means lets us believe. Cluster 1 with serious injuries is similar to cluster 2 of K-Means covering the same accidents. Cluster 2 with 2,5 victims and 2 seriously injured does not appear in the K-Means result.

cluster	TOT_VICTIMAS	TOT_MUERTOS	TOT_HERIDOS_GRAVES	TOT_HERIDOS_LEVES
-1	4.549020	0.916667	1.504902	2.127451
0	1.507094	0.000000	0.000000	1.507094
1	1.396552	0.000000	1.000000	0.396552
2	2.515924	0.000000	2.000000	0.515924
3	1.230769	1.000000	0.000000	0.230769

Figure 16: The matrix shows the centers of each cluster found by the DBSCAN algorithm.

Finally, in fig. 17 we can see the cluster distribution along two variables of interest. The clusters seem to be a bit better separable by most variable combinations then with K-Means. Also DBSCAN puts a lot of samples in the same cluster (-1), which have been in different clusters in the K-Means result.

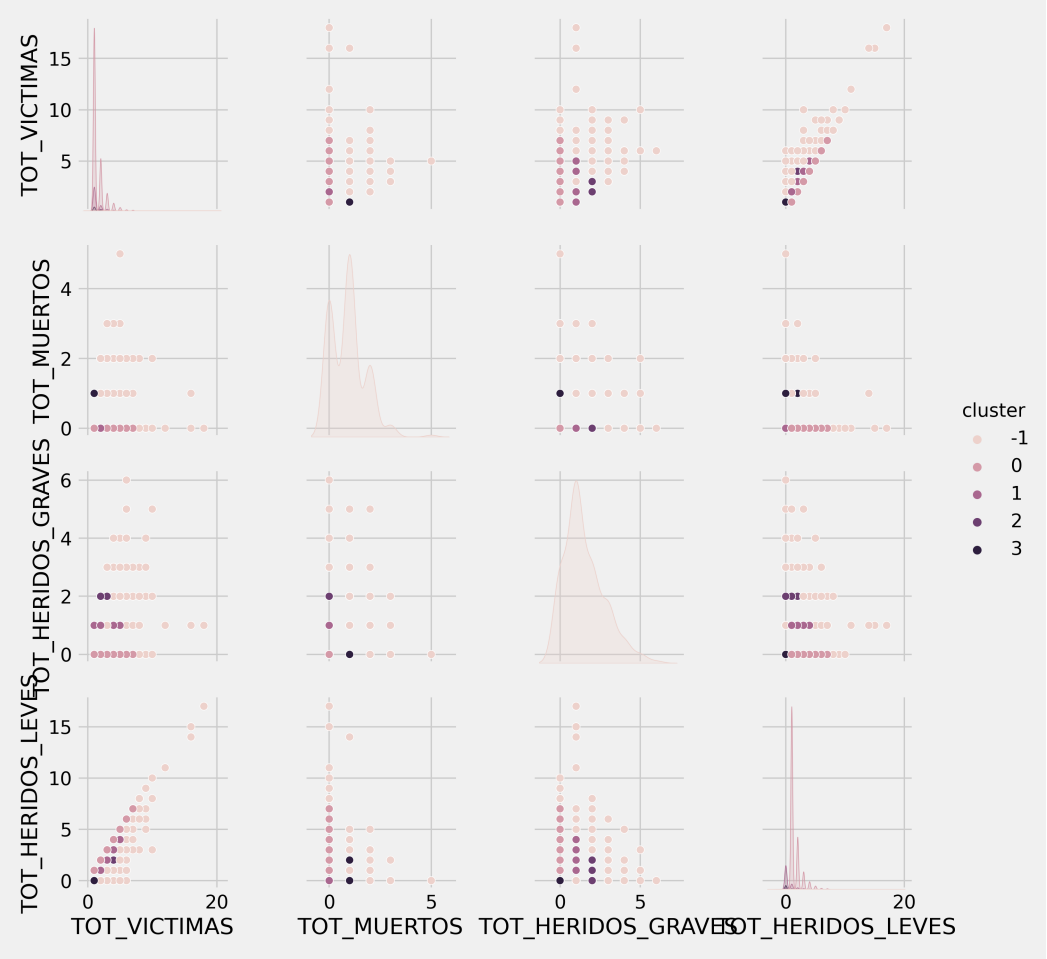

Figure 17: The figure shows a matrix of plots describing the distribution of the 6 clusters (found by DBSCAN algorithm) among the variables of interest.

3.3 K-Means: effect of the parameter k

We will use the same two methods as in case study 1 (Section 2.4) to explore different outcomes of K-Means when increasing the value of k from 2 until 12. The graph for the elbow method, fig. 18, shows that the optimal value is k=5.

The Silhouette Coefficient graph, fig. 19, does not provide a clear answer. The Silhouette Coefficient for k=5 is 0.8459 and good, but slightly smaller then the best result of the first subset (highways). But if k increases more, the Silhouette score increases a lot and gets better. Table 4 shows that the Calinski-Harbasz score increases as well, from 10369.97 (k=5) to around 13000 for k=7,8 and 9. So another good value for k would be 7, increasing the Silhouette score by 0.0563 from 0.8459 to 0.9022 and the Calinski-Harabasz score by 2556. This configuration is slightly less interpretable then k=5, although it provides a better cluster quality.

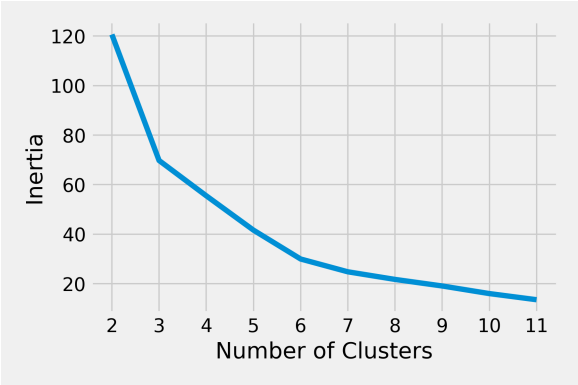

Figure 18: The graph plots the inertia of the clusters found by K-Means for different values of k. The elbow point is located at k=5.

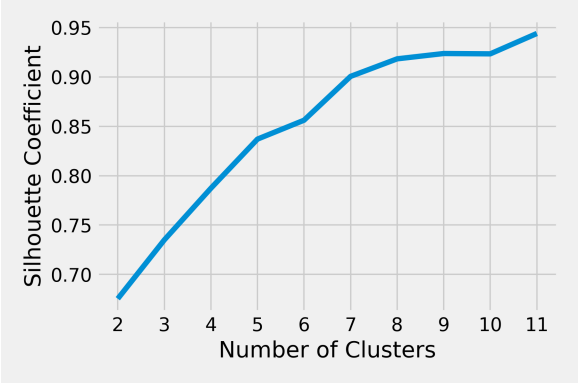

Figure 19: The graph plots the Silhouette Coefficient of the clusters found by K-Means for different values of k. The highest scores are found for k=8 and k=9.

Table 4: This table compares the quality of the resulting clusters by K-Means for the best values of k, using the Silhouette Coefficient and the Calinski-Harabasz Score.

algorithm	Silhouette Coefficient	Calinski-Harbasz score
K-Means (K=5)	0.8459	10369.97
K-Means (K=6)	0.8523	12460.55
K-Means (K=7)	0.9022	12963.95
K-Means (K=8)	0.9156	12926.04
K-Means (K=9)	0.9230	13086.66

3.4 Conclusion

There are five types of accidents on conventional roads during the weekend. The two biggest groups are minor accidents with no deaths or seriously injured people. One group has all accidents with 1 victim as a group center, the other 2,25 as its center. Those follows a still big group, cluster 2 with 1673 accidents, that can be described as having more seriously injured victims then minor injured. The smallest cluster describes the deadliest accidents, happened 303 times on weekends in 2013.

The algorithms find semantically different clusters. The results of the DBSCAN have a lower quality by the evaluation metrics, and they group the samples in a different way, obtaining different cluster centers then the K-Means algorithm.

3.5 Conclusion Case Study 1: Accidents on highways vs. conventional roads during weekends in Spain

For both subsets of the data, the accidents on highways and the accidents on conventional roads, the K-Means algorithm produces the better result. In the first case, we have 3 clusters, being highly unbalanced. In the second, we have 5 clusters, being more balanced. There were 4981 accidents on highways and 12.085 on conventional roads during the weekends of the year 2013 in Spain. This is important for the interpretation of the clusters and their individual size. In both cases, the largest clusters are the ones containing the minor accidents. On highways those are 4447 accidents being 89,28% of all accidents on highways. For conventional roads, cluster 0 and 1 make up 9516 of all samples, being 78,54% of all accidents on that roads.

Next, we have cluster 1 for highways and cluster 2 for conventional roads both describing the accidents with seriously injured victims. This type of accidents turn out to be more severe on highways, because the mean value of total victims is 2,224 (0,7 higher then on conventional roads) and the mean value of minor injured 1,031 compared to 0,343 only. On highways they happened 425 times

(8,56% of all accidents) and on conventional roads 1673 times (13,84%), so they occur less often on highways but are more severe.

Equally important is a comparison between the clusters representing the most fatal accidents, involving dead victims. Surprisingly the mean value of death victims is the same (1,16), the mean of seriously injured very similar (0,239 for highways and 0,281 for conv. roads). However on highways, those accidents tend to have more victims (2 vs 1,75) and thus more of them being minor injured. They make up 2,18% of all highway accidents and 2,5% of all conventional road accidents according to the clusters found.

Additionally, cluster 3 of conv. roads is a unique group having almost 5 victims as a mean, a values higher then 0 for deaths and serious injuries, with most victims having minor injuries (4,64). This cluster has 694 samples, 5,74% of all samples of the subset.

In both cases all three variables are used to distinguish between clusters, only the total dead victims is not considered by the K-Means algorithm, as shown in the cluster center matrix. All values for this variable are 0 or almost 0 (smaller then 0,025) besides in one cluster.

To conclude, more minor accidents happen on highways then on conventional roads, which is a surprising result. On the contrary, accidents on highways are more severe when they lead to serious injuries, but occur less often then on conventional roads. Likewise, the deadly accidents include more people on the highways, leading to more minor injures, but not raising the mean value of actual deaths and serious injuries. Conventional roads have a small portion of accidents (5,74%) that have double the victims as a mean value then any cluster of the highways accidents. Looking at the attributes of interest, total deaths and total serious injuries, the accidents on highways and on conventional roads tend to be similar fatal, no mean value is significantly higher in one subset or the other.

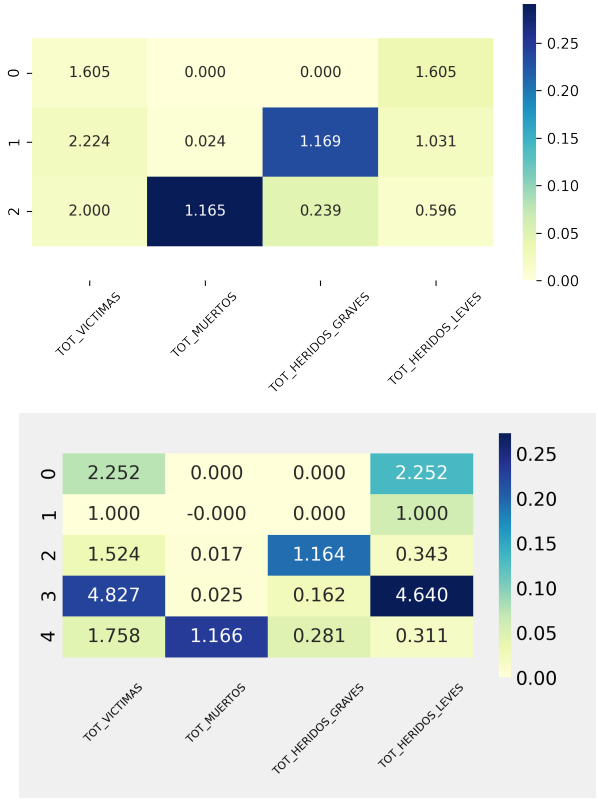

Figure 20: The centers of each cluster for the attributes of interest characterizing the gravity of car accidents. The top matrix shows the centers for highways accidents and the bottom matrix for conventional road accidents.

4 Case Study 2: What clusters exist during bad weather?

In this section the obtained clusters by K-Means and DBSCAN on the subset that contains accidents that occurred during bad weather will be evaluated, visualized and interpreted.

4.1 Case Description

This time only accidents that happened during bad weather conditions will be clustered. We assume that rain and wet streets can cause more accidents, and we want to learn more about the nature of the accidents with this weather condition.

In fig. 21 it is shown that some road surface values have a wider distribution

over the total victims then others. Those are wet (mojada), dry and clean (seca y limpia), iced (helada), snow (nevada) and shady (umbria).

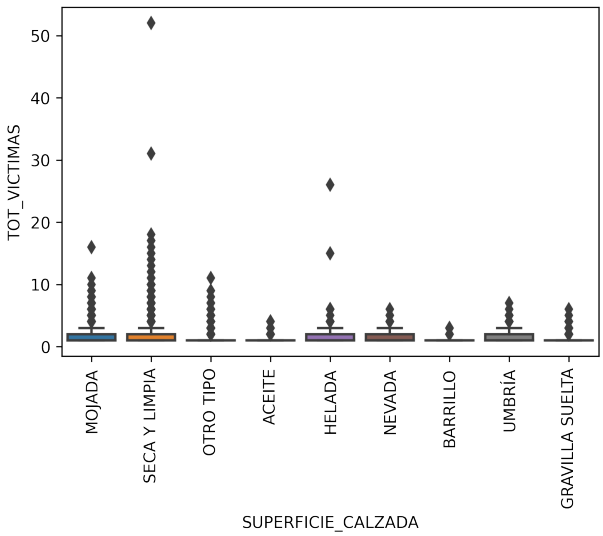

Figure 21: Shows the distribution of the total number of victims over different values for the variable "road surface".

We select all samples that have "strong rain" and "raining" as a "FACTORES ATMOSFERICOS" (ATMOSPHERIC FACTORS) value and "wet" as a "SUPERFICIE CALZADA" (ROAD SURFACE) value out of the interesting ones from fig. 21, to align with the rain. This selections gives a subset with 9.325 accidents. The variables of interest are staying the same as in case study 1, namely

- total number of victims (TOT VICTIMAS)
- total number of deaths (TOT MUERTOS)
- total number of seriously injured persons (TOT HERIDOS GRAVES)
- total number of minor injured persons (TOT HERIDOS LEVES)

For the K-Means algorithm all samples of the variables of interest are normalized again.

4.2 Results

The clustering algorithms are evaluated by the same two metrics as in the first case study. Table 5 shows these values for both algorithms. We see that the K-Means algorithm creates dense and well separated clusters, given the high

Silhouette score of 0.94 and a very high Calinski-Harabasz Score. The DBSCAN algorithm performs significantly worse on this selected subset of samples. The execution time is really short for both algorithms.

The final configuration of DBSCAN for this case study is: epsilon 0.1, minimum samples 70. Trying other algorithms such as MeanShift and Affility Propagation produces only a single cluster on the default setting.

Table 5: This table compares the quality of the resulting clusters by K-Means and DBSCAN, using the Silhouette score and the Calinski-Harabasz Score.

algorithm	Silhouette score	Calinski-Harbasz score	execution time
K-Means (K=6)	0.9468	12813.9	0.23 s
DBSCAN	0.7037	1587.83	1.6 s

4.3 Cluster Visualization

We will take a close look at the clusters, by reviewing the sizes, the cluster centers and the distribution over 2 of the attributes of interest at a time using a pair plot.

K-Means

First, we can see in fig. 22 that the clusters are very unbalanced, although well separated as the scores indicated. Most of the accidents are in cluster 0 (6341 samples), followed by cluster 4 (1527), some in cluster 2 (578), 5 (471) and cluster 1 (318). Cluster 3 is the smallest (90 samples).

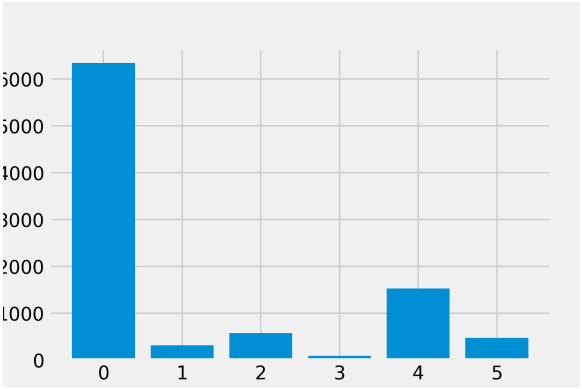

Figure 22: Shows the amount of samples in each cluster found by K-Means (K=6). Cluster 0: 6341, 1: 318, 2: 578, 3: 90, 4: 1527, 5:471.

To understand what characteristics each cluster has, fig. 23 shows the center of each cluster. Cluster 0, 1, 4 and 5 represent minor accidents (only minor injuries), distinguished by the mean number of victims. That number correlates with the number of injured, meaning every victim got also injured. Cluster 0 has 1 victim, cluster 4 has 2, cluster 5 has 3 and cluster 1 has 4.63. Cluster 1 and 5 differ for the higher numbers of victims a bit more in this group, having also more then 0 seriously injured (but still a small mean: 0.044).

Cluster 2 is the third largest cluster and it can be described as the one with seriously injured victims, having a much higher value then the rest of the clusters. Finally, cluster 3 characterizes the deadly accidents, being the single cluster with a significant death rate (1,144 for 1,7 victims). It has more minor injured cases then the cluster 2 with the seriously injured victims.

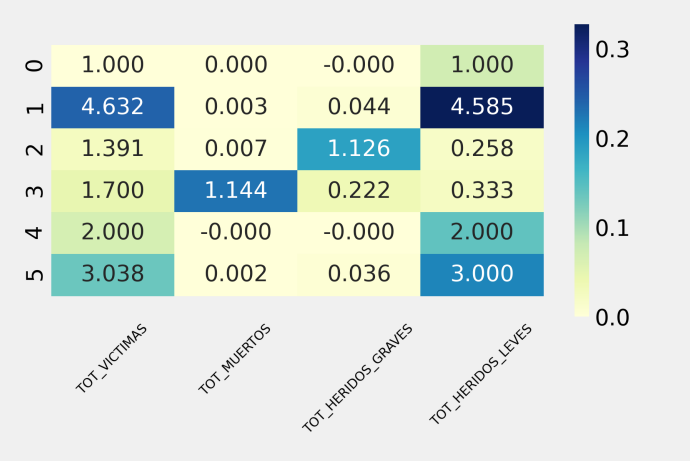

Figure 23: Shows the centers of each cluster found by K-Means(K=6).

Finally, in fig. 24 we can see the cluster distribution along two variables of interest on the normalized data.

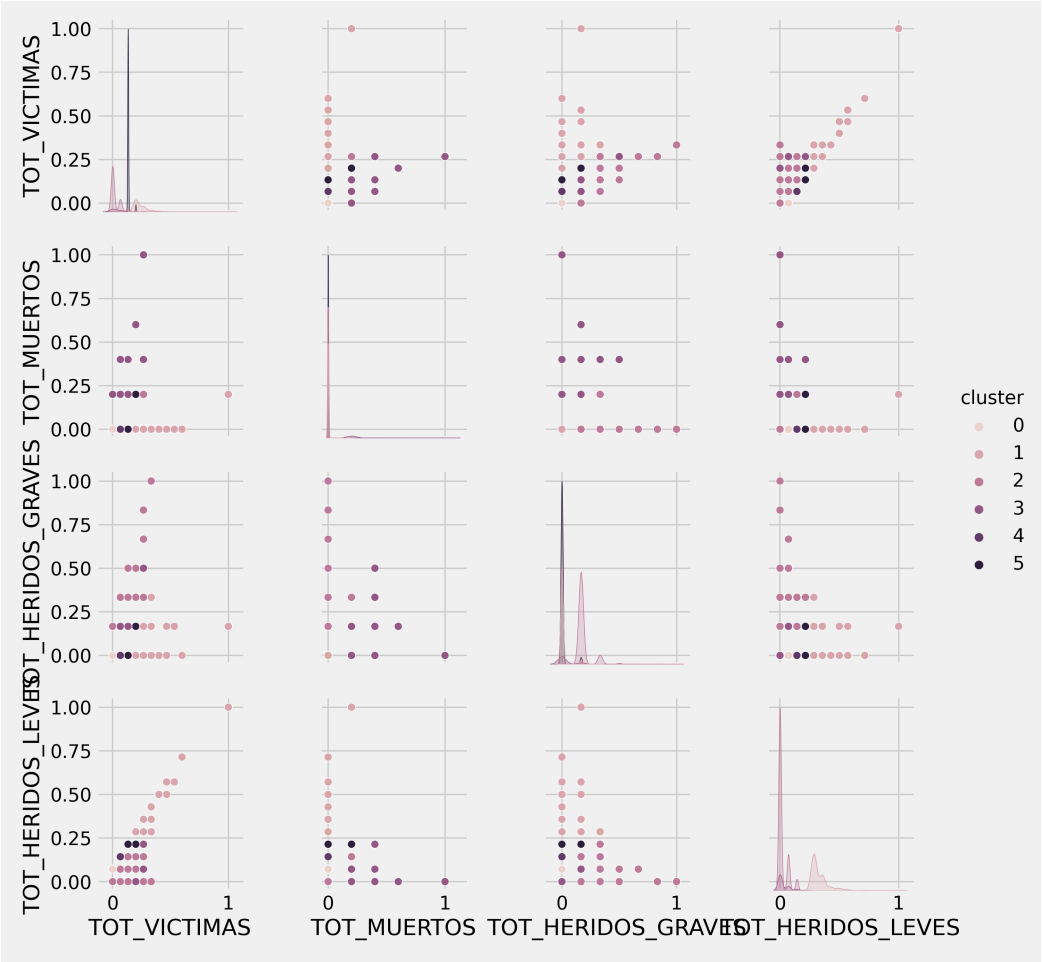

Figure 24: The figure shows a matrix of plots describing the distribution of the three clusters found by K-Means (K=6).

We observe on the diagonal axis, that some clusters are separable, whereas others being part or a subset of another cluster. Looking at the different colored points in column 2 and 3 (minor and seriously injured people), we see the same more separability of certain colors like on the diagonal axis, e.g. cluster 1 and 2 or 1 and 3.

DBSCAN

First, we can see in fig. 25 that the three clusters founds are very unbalanced. The DBSCAN algorithm finds a huge dominating cluster with 8.622 samples, 2.281 more then the biggest cluster of the K-Means algorithm. Cluster 1 has 536 samples and cluster -1 167.

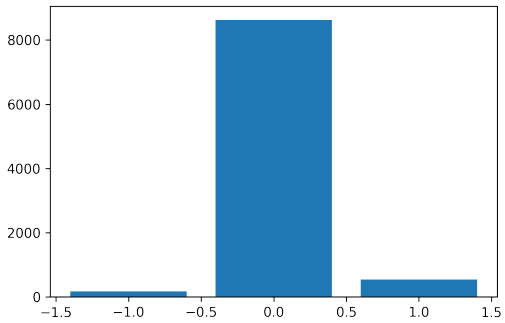

Figure 25: The figure shows the amount of samples in each cluster found by the DBSCAN algorithm. Cluster -1: 167, 0: 8622, 1:536

Fig. 26 shows the center of each cluster, and we see that they are well separated and different from one another. Cluster -1 has the most victims, all the dead accidents, but also accidents with seriously injured people and minor injured people. It seems that this cluster covers a lot of different type of consequences of accidents and/or accidents that had different types of damage for each victim. Probably the better separated clusters found by K-Means are put together in this cluster.

Cluster 0 contains all the minor accidents, summing um cluster 0 and 4 of K-Means which contain the same type of accidents. Cluster 1 finally contains accidents with seriously injured people, but no dead victims.

Finally, in fig. 27 we can see the cluster distribution along two variables of interest. The number of total minor injuries and total victims provides a good separation of the three clusters. In addition, the total death victims and total minor injured or total victims provide a separation. No variable itself separates the clusters well enough.

cluster	TOT_VICTIMAS	TOT_MUERTOS	TOT_HERIDOS_GRAVES	TOT_HERIDOS_LEVES
-1	2.622754	0.652695	0.994012	0.976048
0	1.403967	0.000000	0.000000	1.403967
1	1.330224	0.000000	1.000000	0.330224

Figure 26: The matrix shows the centers of each clusters found by the DBSCAN algorithm.

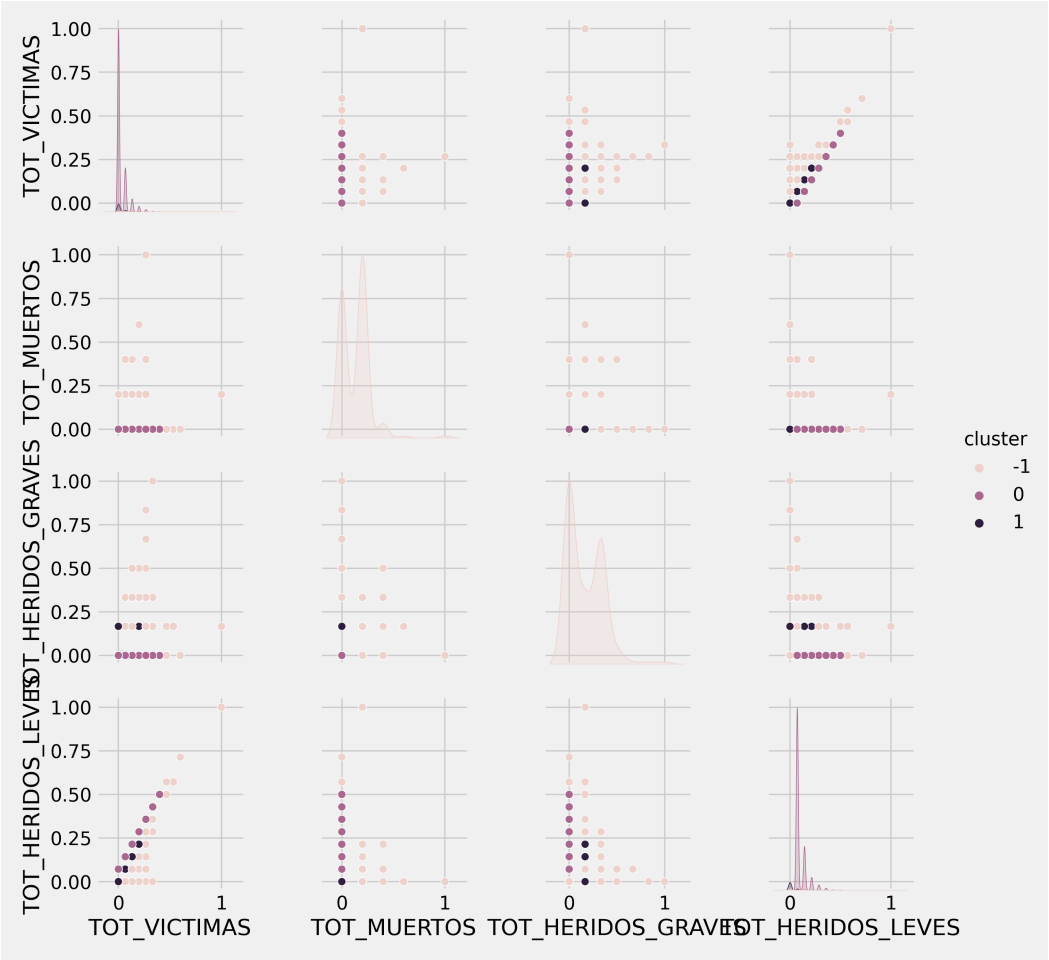

Figure 27: The figure shows a matrix of plots describing the distribution of the 3 clusters found by DBSCAN algorithm.

4.4 K-Means: effect of the parameter k

We will use the same two methods as in case study 1 (Section 2.4) to explore different outcomes of K-Means when increasing the value of k from 2 until 12. The graph for the elbow method, fig. 28, shows that the optimal value is k=4. The Silhouette Coefficient graph, fig. 29, does not provide a clear answer. The Silhouette Coefficient for k=4, the optimum by the elbow method, is good and like the best results of case study 1, 0.86. But if k increases more, the Silhouette score increases too and gets even better. Table 6 shows that the Calinski-Harbasz score increases surprisingly as well, tripling from k=4 (5225.8144) to k=7 (14403.4013). We can conclude that more clusters return better quality clusters, as indicated by these two evaluation scores.

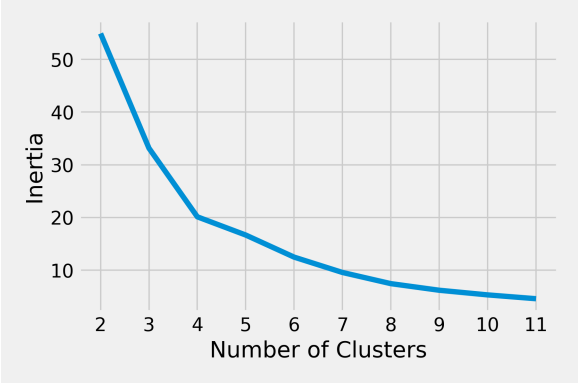

Figure 28: The graph plots the inertia of the clusters found by K-Means for different values of k. The elbow point is located at k=4.

We see in fig. 30 the cluster distribution along two variables of interest on the normalized data for k=4. We can see that the plots with total minor injuries and any other variable (plots (3,1),(3,2)(3,3) or plots in column 3) separate very well between cluster 3 and the rest. The plots with total seriously injured and any other variable distinguish very good between all clusters. To conclude, choosing k=4 for the K-Means algorithm leads to graphically more interpretable and well separable clusters, but the evaluation metrics on the inner and outer qualities of the clusters are worse compared to higher values for k.

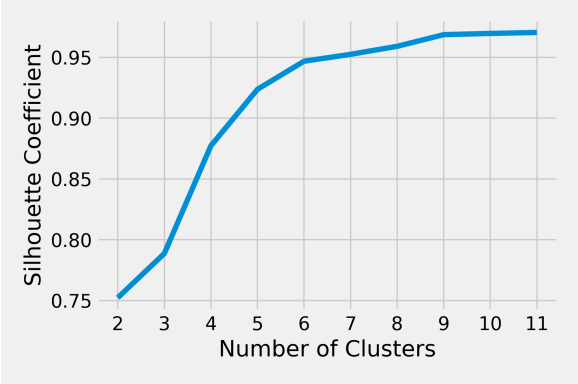

Figure 29: The graph plots the Silhouette Coefficient of the clusters found by K-Means for different values of k. The score increase with k.

Table 6: This table compares the quality of the resulting clusters by K-Means for the best values of k, using the Silhouette Coefficient and the Calinski-Harabasz Score.

algorithm	Silhouette Coefficient	Calinski-Harbasz score
K-Means (K=4)	0.8645	5225.8144
K-Means (K=5)	0.8994	7776.5024
K-Means (K=6)	0.9468	12813.8965
K-Means (K=7)	0.9525	14403.4013
K-Means (K=8)	0.9590	16232.4460

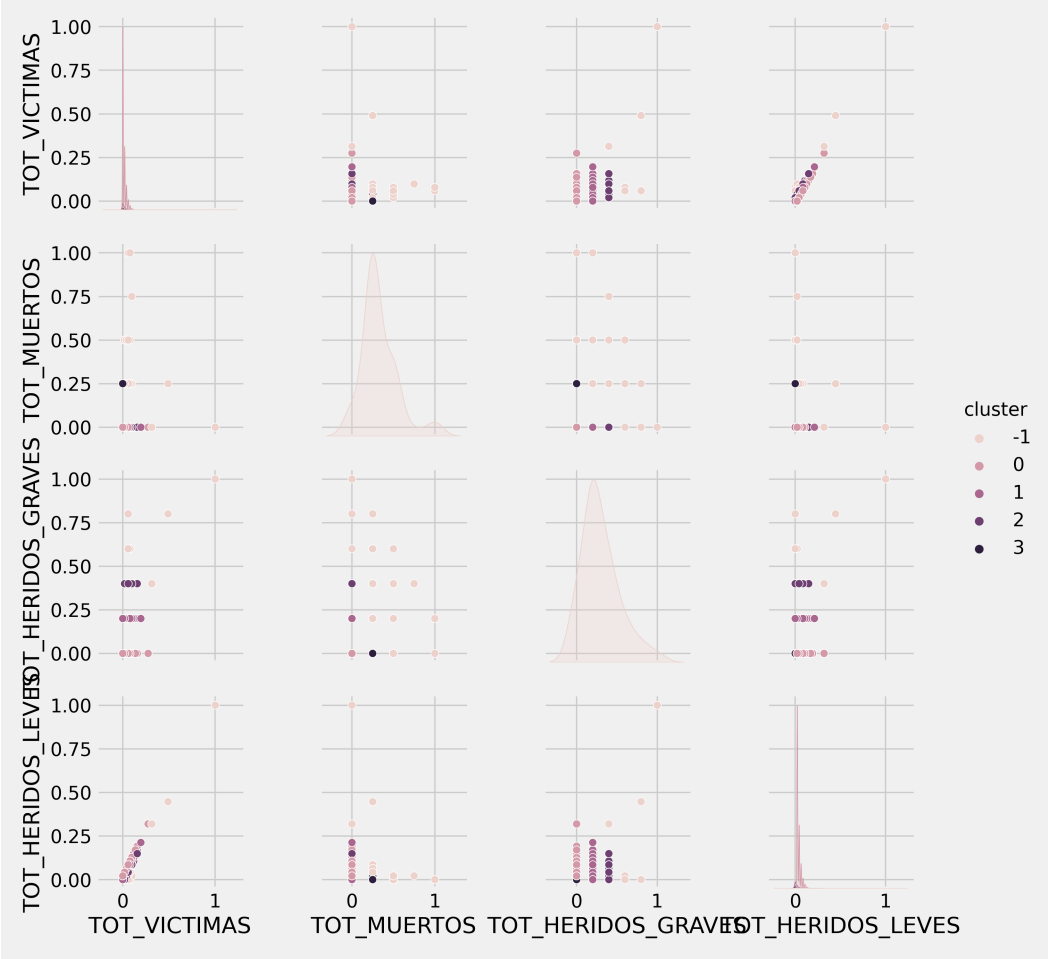

Figure 30: The figure shows a matrix of plots describing the distribution of the three clusters found by K-Means (K=4) for every pair of variables from the set of variables of interest (describing the gravity of an accident).

4.5 Conclusion

The K-Means algorithm finds out that we have 5 types of different accidents during rainy weather and wet roads. The total victims and total minor injured variable are used to separate between the clusters, so we have accidents with 1, 2 or 3 victims during rainy weather. Moreover accidents with deaths and with seriously injured occur in two cluster groups.

The DBSCAN algorithm clusters the accidents significantly worse, but it still tries to use the total victims and minor injured values to separate between clusters, like K-Means.

5 Visualizations of Practice 1 (Mammography Dataset)

5.1 Data Preprocessing

We can observe the results of each missing value replacement process. We applied the replacement technique my the mean value, the median value, the most frequent value and by a constant (0) on the variables of the dataset. Surprisingly, none of the replacement techniques improves the model performance, only the replacement by a constant comes somewhat close to the results of the dataset without missing values. The KNN result shows fig.31, the Decision Tree result shows fig.32, the kernel SVM results shows fig.33, the Logistic Regression results shows fig.34 and the Random Forrest results shows fig.35.

Evaluation Measure	mean	median	most_frequent	constant
accuracy	0.7378	0.7378	0.7378	0.7940
sensitivity	0.7528	0.7528	0.7528	0.8045
specificity	0.7248	0.7248	0.7248	0.7849
FPR	0.2752	0.2752	0.2752	0.2151
preciscion	0.7023	0.7023	0.7023	0.7633
AUC	0.7733	0.7733	0.7733	0.8568

Figure 31: The Table shows the effect of each preprocessing technique on the evaluation metrics of the KNN model.

Evaluation Measure	mean	median	most_frequent	constant
accuracy	0.7825	0.7825	0.7825	0.8137
sensitivity	0.8090	0.8090	0.8090	0.7775
specificity	0.7597	0.7597	0.7597	0.8450
FPR	0.2403	0.2403	0.2403	0.1550
preciscion	0.7438	0.7438	0.7438	0.8122
AUC	0.7901	0.7901	0.7901	0.8772

Figure 32: The Table shows the effect of each preprocessing technique on the evaluation metrics of the Decision Tree model.

Evaluation Measure	mean	median	most_frequent	constant
accuracy	0.7586	0.7586	0.7586	0.7981
sensitivity	0.7708	0.7708	0.7708	0.8337
specificity	0.7481	0.7481	0.7481	0.7674
FPR	0.2519	0.2519	0.2519	0.2326
preciscion	0.7252	0.7252	0.7252	0.7556
AUC	0.8014	0.8009	0.8012	0.8715

Figure 33: The Table shows the effect of each preprocessing technique on the evaluation metrics of the kernel SVM model

Evaluation Measure	mean	median	most_frequent	constant
accuracy	0.7742	0.7742	0.7742	0.8169
sensitivity	0.7978	0.7978	0.7978	0.8135
specificity	0.7539	0.7539	0.7539	0.8198
FPR	0.2461	0.2461	0.2461	0.1802
preciscion	0.7365	0.7365	0.7365	0.7956
AUC	0.8098	0.8098	0.8098	0.8912

Figure 34: The Table shows the effect of each preprocessing technique on the Logistic Regression model.

Evaluation Measure	mean	median	most_frequent	constant
accuracy	0.7617	0.7617	0.7617	0.8252
sensitivity	0.7596	0.7596	0.7596	0.8000
specificity	0.7636	0.7636	0.7636	0.8469
FPR	0.2364	0.2364	0.2364	0.1531
preciscion	0.7348	0.7348	0.7348	0.8184
AUC	0.7876	0.7876	0.7876	0.8967

Figure 35: The Table shows the effect of each preprocessing technique on the Random Forrest model.

5.2 ROC Curve

The ROC Curve with all classifiers can be seen in fig. 36.

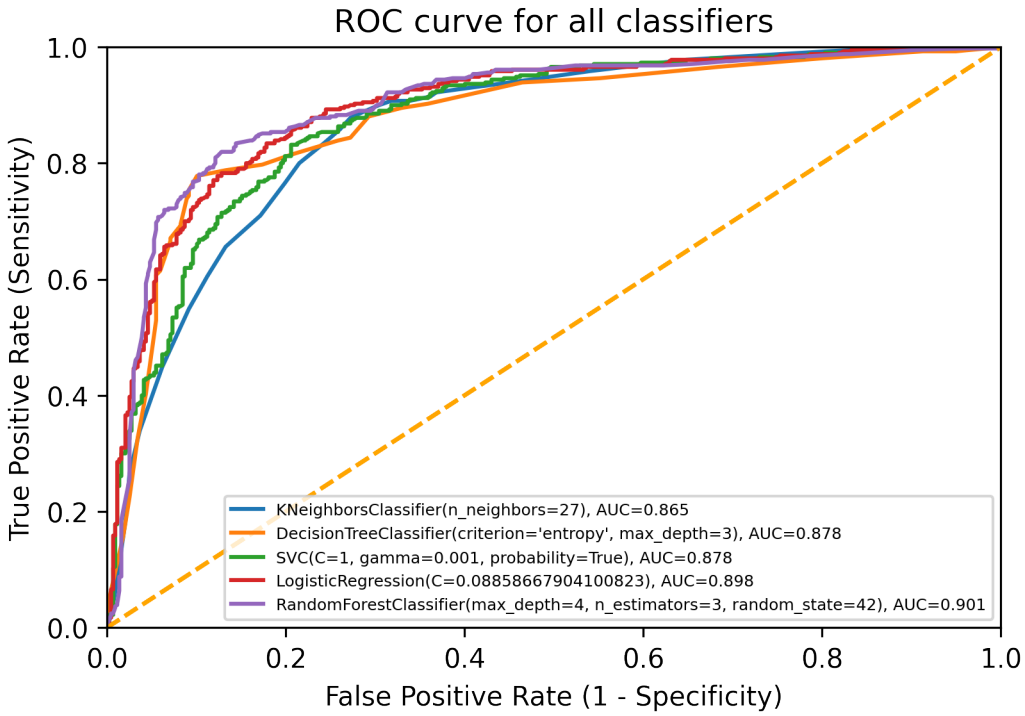

Figure 36: ROC curve with all classifiers

5.3 Visualization of feature impact on severity

Fig. 37 shows the impact of the feature BI-RADS on the severity of the cancer. We notice two very important observations. First, patients with a BI-RADS value of 5 are extremely more likely to have a malignant tumor. Second, a lot of patients in the data who have BI-RADS of 4, have a benign tumor. It is also more likely to have a malignant tumor with BI-RADS 4 then with any other (apart from 5), so the value 4 is not as predictive as 5.

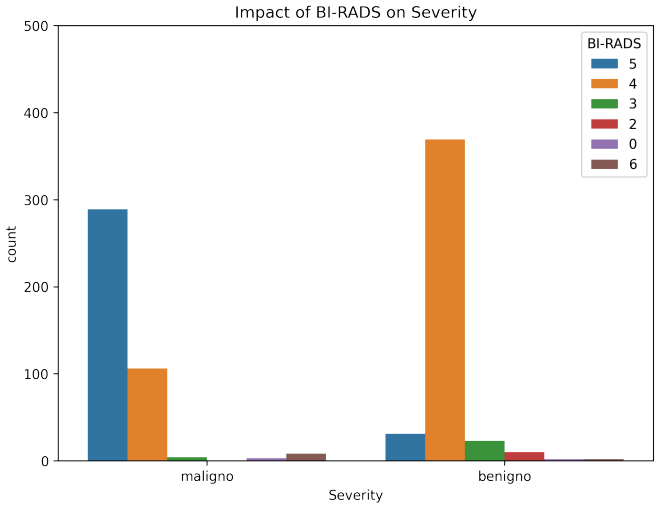

Figure 37: Impact of BI-RADS on Severity

Next, fig. 38 shows the impact of Margin on Severity. Here we can observe that if the mass margin is circumscribed (1), then the tumor is more likely to be benign. On the other hand, a malignant tumor occurs more often if the mass margin is speculated or ill-defined.

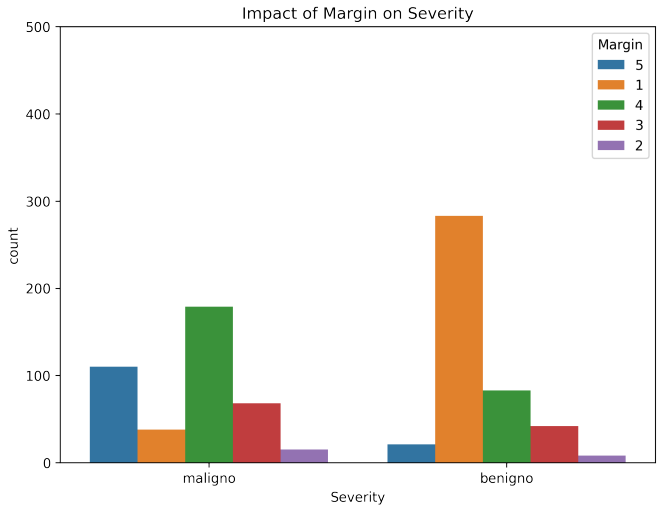

Figure 38: Impact of mass margin on severity

In addition, the impact of the shape of the abnormal mass detected on the severity is plotted in fig. 39. It shows that most patients with malignant tumor have a irregular (I) shape. On the contrary, patients with a round and oval shape have a higher probability of a benign tumor. For patients with not defined or lobular shapes the chances for both tumors are equal.

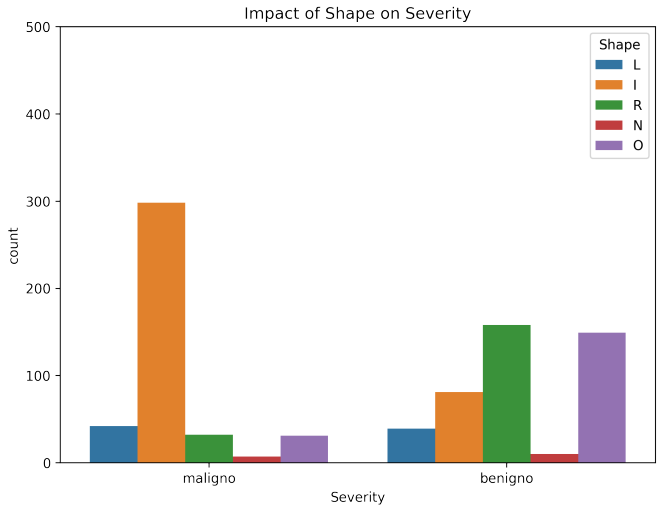

Figure 39: Impact of shape on severity

Fig. 40 shows the impact of density on severity. The chances of both tumor types are more or less equal no matter what values density takes on.

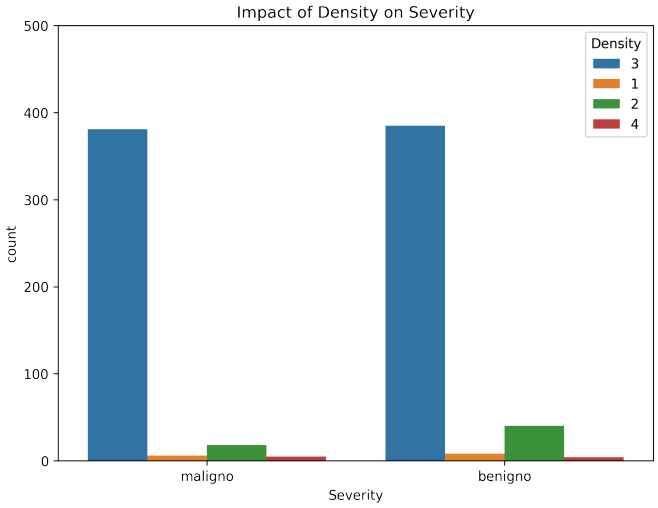

Figure 40: Impact of density on severity

Given the above, BI-RADS is the most important feature to predict the tumor type. On the contrary, Density gives us the least information on the severity. Yet, since there are 5 different features in the dataset, a plot showing the frequency of one feature and it's impact on the severity is not sufficient. We need to experiment with different combinations of features to see how they affect together the severity type.

References

[1] `https://realpython.com/k-means-clustering-python/` *K-Means Clustering in Python: A Practical Guide*, last access 09-12-2020.

[2] `https://towardsdatascience.com/explaining-dbscan-clustering-18eaf5c83b31` *Explaining DBSCAN Clustering*, last access 09-12-2020.

[3] `https://en.wikipedia.org/wiki/Elbow_method_(clustering)` *Elbow method (clustering)*, last access 09-12-2020.