Práctica 2: Visualización y Segmentación

Alejandro Alonso Membrilla

Grupo: viernes
Correo: aalonso99@correo.ugr.es

Contents

1 Apartado 1: Visualización

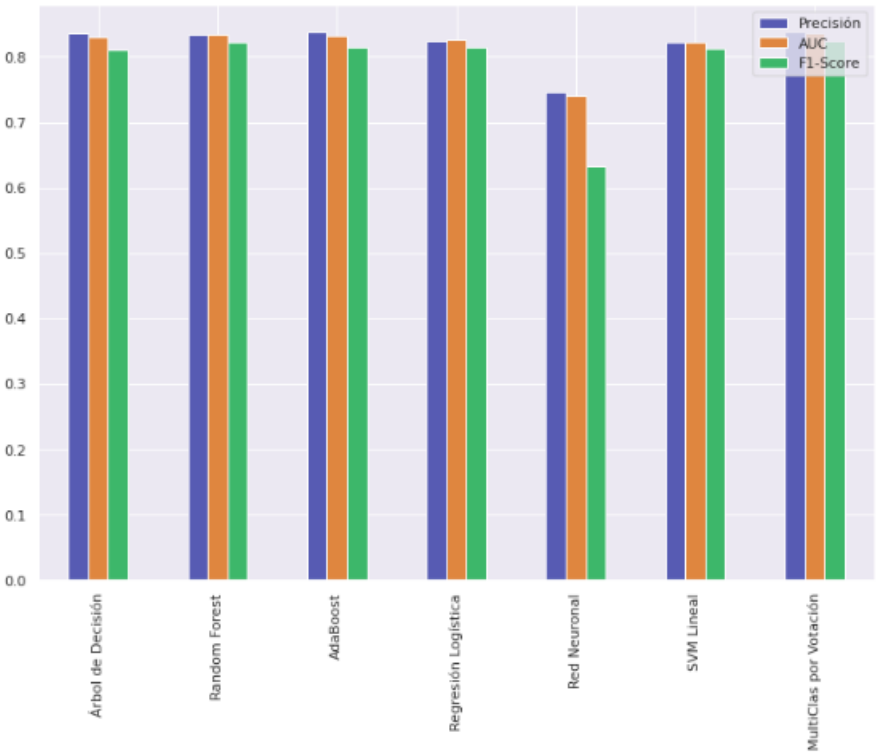
La visualización es una parte clave del análisis de datos. Visualizar el conjunto de datos y los resultados obtenidos nos ayuda a obtener información clara del problema sin tener que indagar en sus aspectos numéricos, que pueden resultar intrarables cuando nuestro conjunto de datos es muy grande o complejo. En la primera parte de esta práctica tendremos el objetivo de aplicar técnicas de visualización como complemento al estudio y clasificación de las mamografías realizado en la práctica anterior. Veremos como comparar los distintos atributos de los que tenemos información y ver su relación con la variable objetivo (malignidad del tumor), además de distintas métricas relacionadas con la calidad de los resultados de cada clasificador.

El código entregado correspondiente a este apartado consiste en el mismo cuaderno de Jupyter escrito para la práctica anterior, al que se le han añadido celdas de código que generan las gráficas pedidas para los siguientes ejercicios.

1.1 Visualización de las medidas

Procedemos a mostrar y comparar las gráficas (histogramas) con los resultados de cada algoritmo de clasificación usado en el apartado anterior. Mostraremos primero los resultados de todos los clasificadores por cada tipo de preprocesamiento, y luego visualizaremos los mismos resultados comparando distintos preprocesamientos para un mismo clasificador.

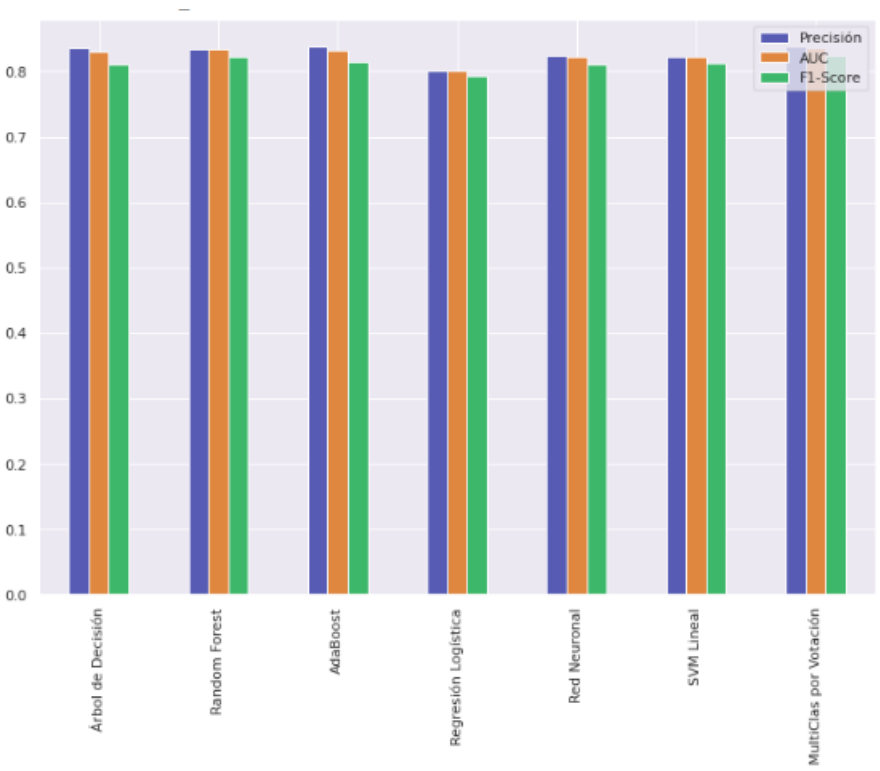
Comparación de los resultados con datos sin normalizar:



Viendo la gráfica, recordamos que la red neuronal obtenía resultados dráticamente peores cuando no normalizábamos el conjunto de datos.

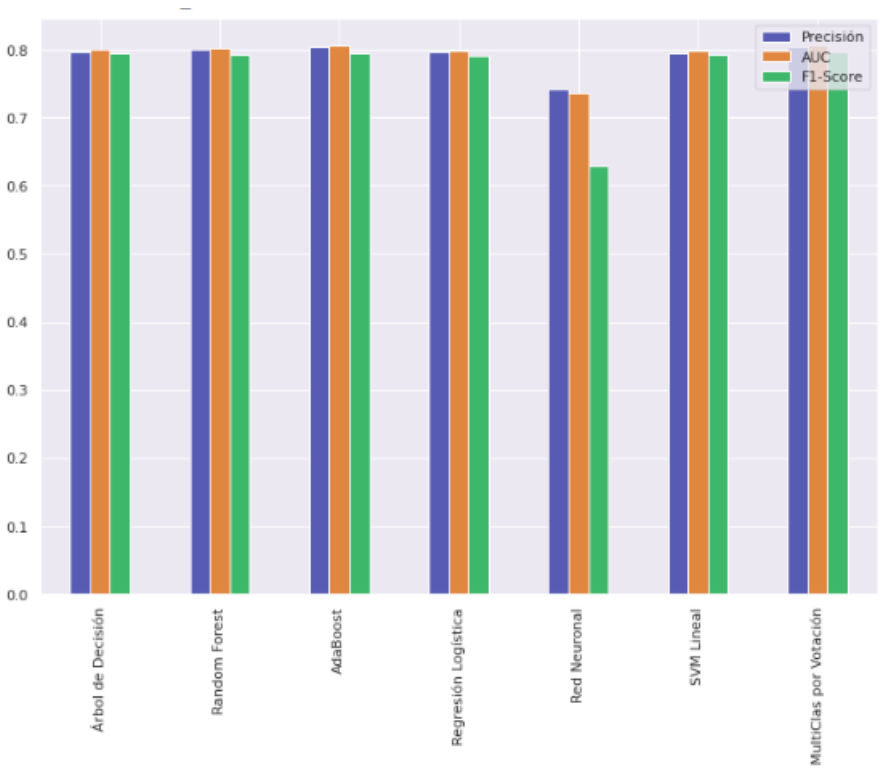
También vemos que casi todos los clasificadores obtienen resultados aproximadamente iguales, con una puntuación F1 ligeramente por debajo de la precisión.

Comparamos ahora los resultados con datos normalizados:



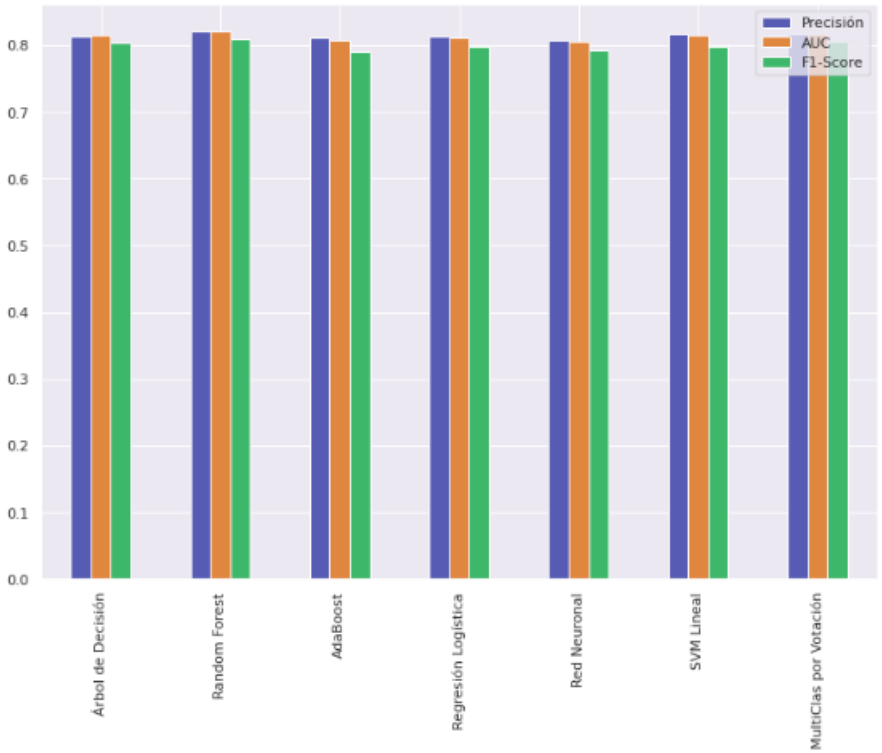
Vemos que ahora los resultados son incluso más homogéneos que en el caso anterior. Las únicas medidas que destacan son las de la regresión logística, un poco debajo de las demás, y tal vez las del multiclasificador, un poco por encima.

Vemos los resultados con datos reducidos mediante ACP:



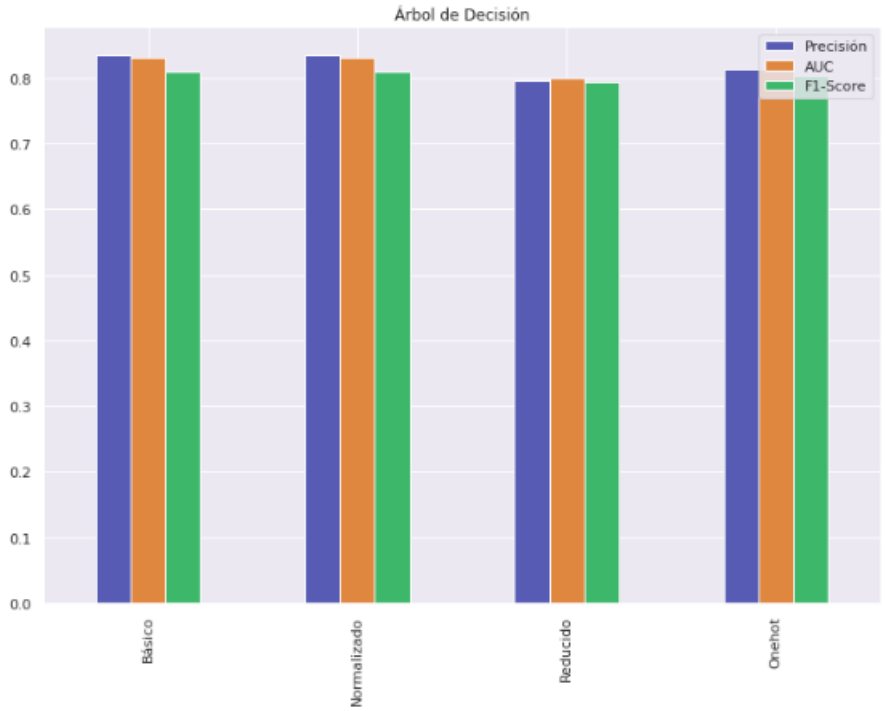
De nuevo, la red neuronal cae por debajo del resto de algoritmos, puesto que el conjunto reducido está sin normalizar. El resto de medidas están muy parejas entre sí, aunque aparentemente más bajas que las obtenidas con otros preprocesamientos. Esto quedará más claro al comparar directamente los preprocesamientos de cada clasificador entre sí.

Vemos los resultados para el preprocesamiento con vectores OneHot binarios:

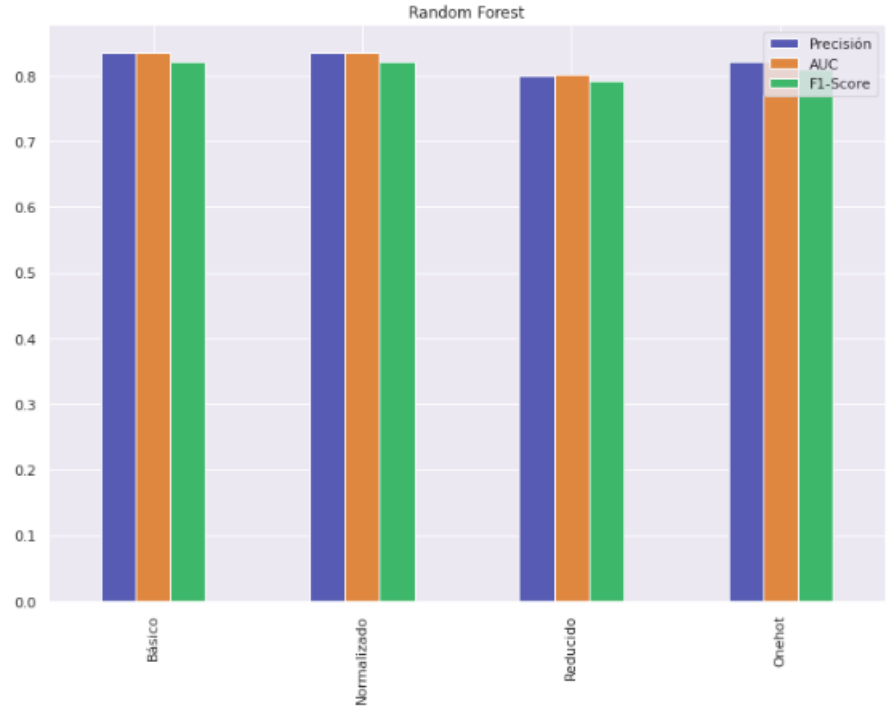


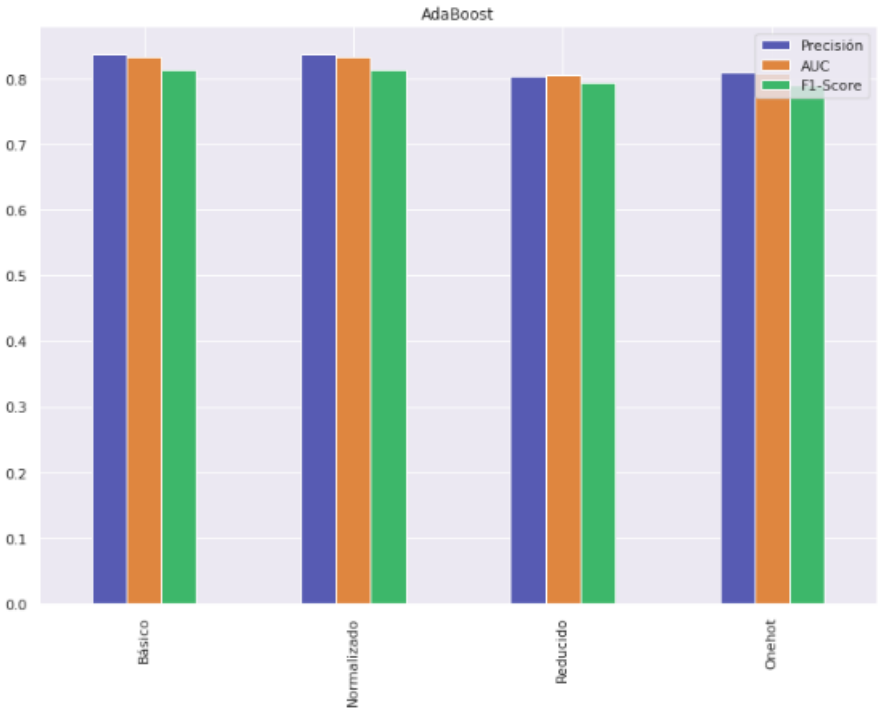
Observamos que, aunque en este caso los datos no han sido normalizados, los problemas encontrados por la red neuronal desaparecen al aplicar este preprocesado, o como mínimo son mucho menos notables.

De igual interés puede ser comparar los cambios que un mismo preprocesado puede tener sobre un algoritmo específico. Mostramos dicha comparación en las siguientes gráficas:

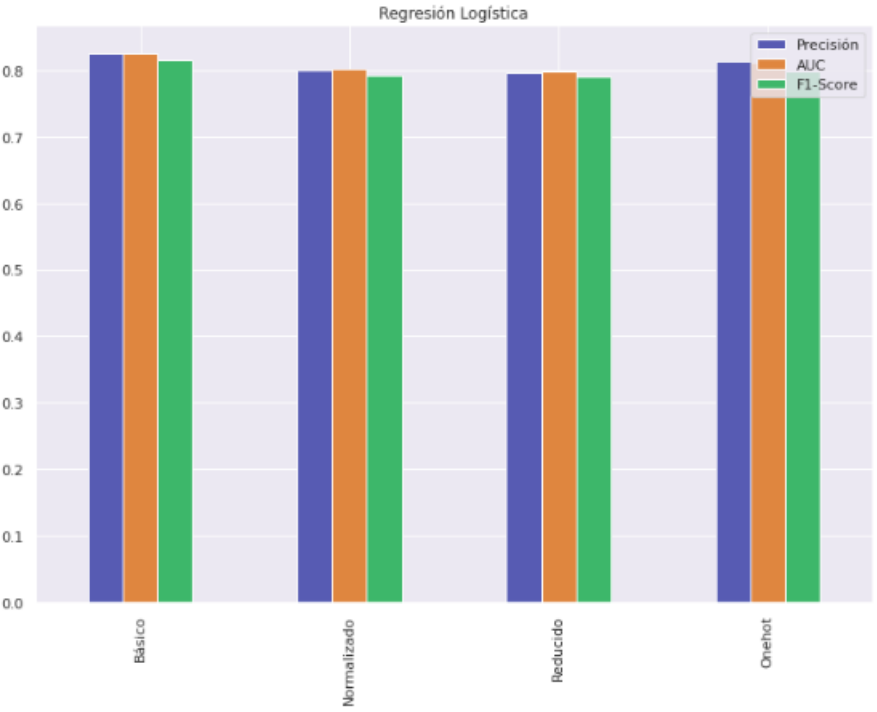


En este caso (y en muchos de los siguientes) la normalización no afecta a los resultados por las causas argumentadas en la práctica anterior. La reducción de datos, por el contrario, reduce visiblemente la calidad de los resultados obtenidos.

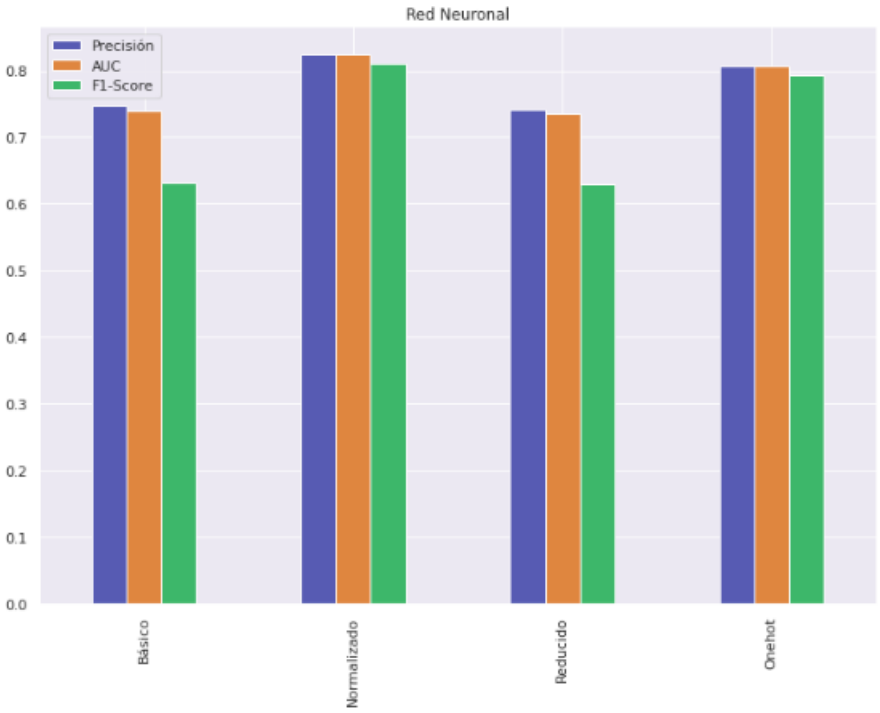




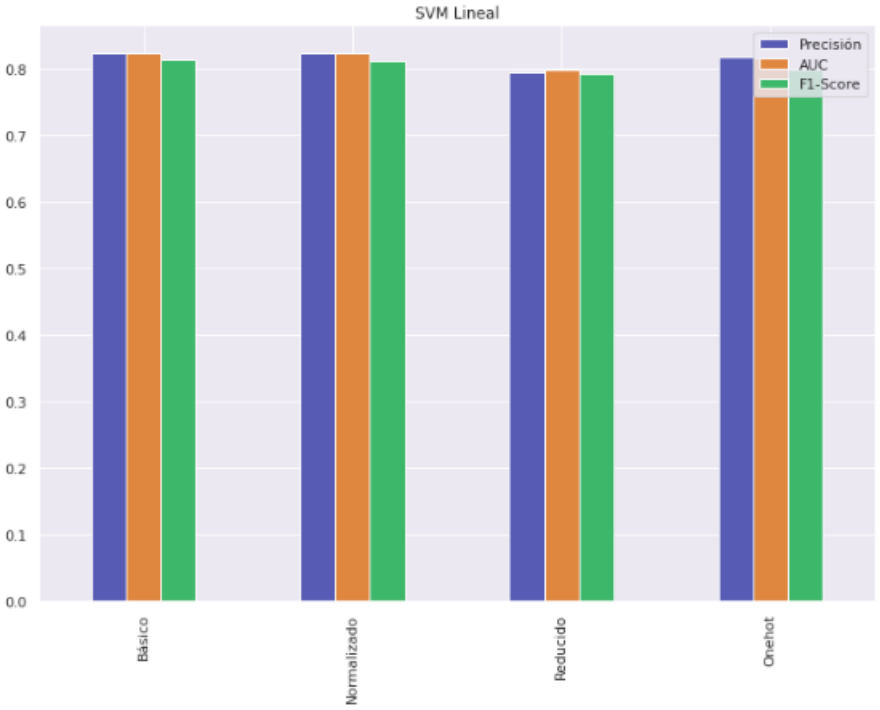
En random forest y AdaBoost los resultados son análogos a los visualizados para el árbol de decisión.

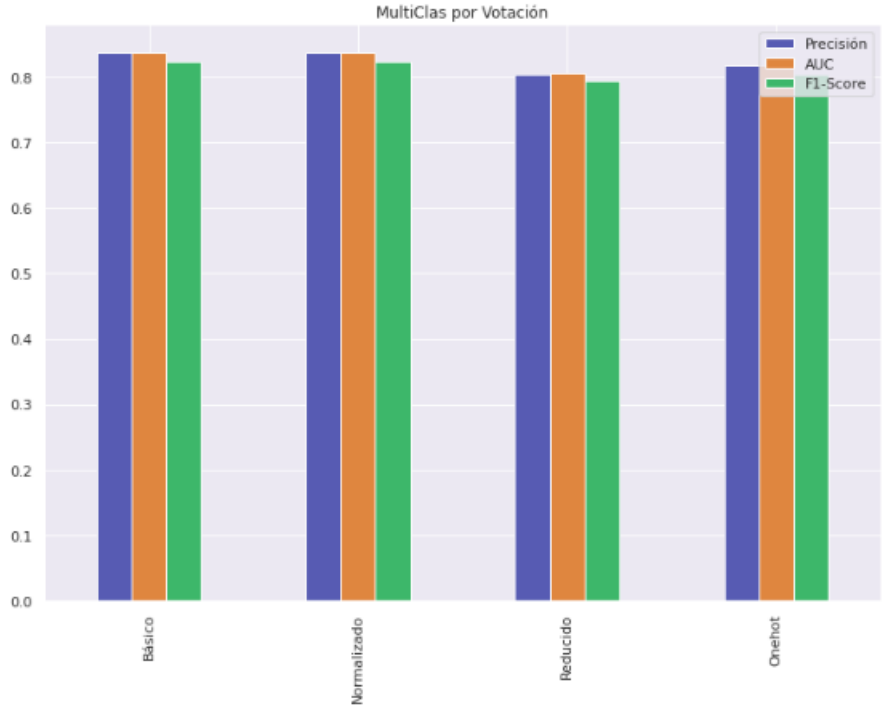


Usando la comparación anterior, vemos claramente que la normalización empeora los resultados de la regresión logística.



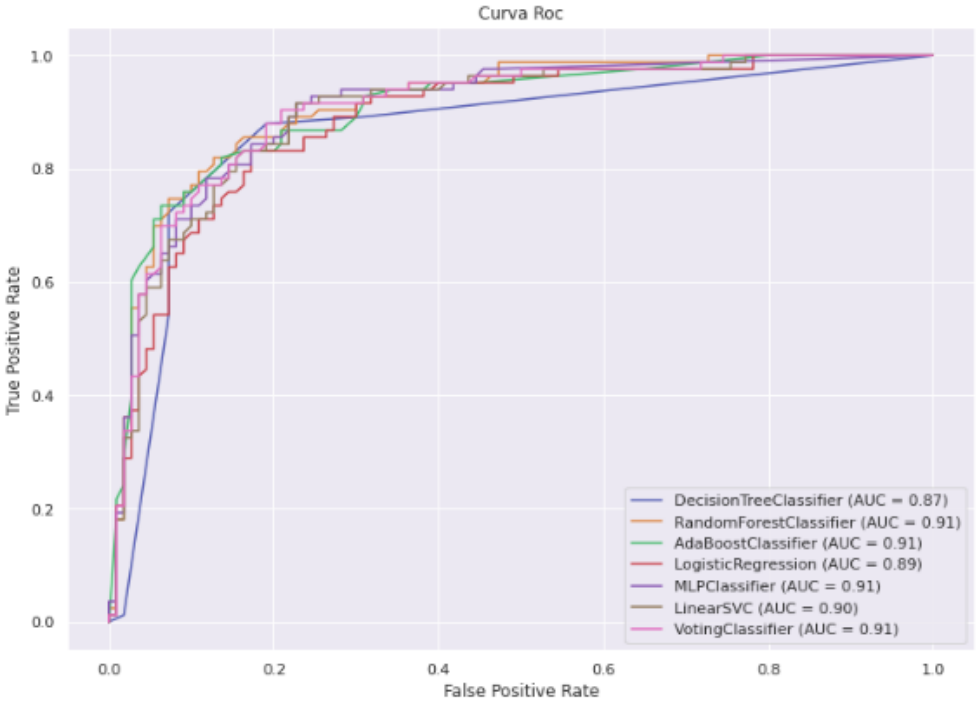
Y en el caso de la red neuronal, vemos que las medidas suben claramente al usar datos normalizados.





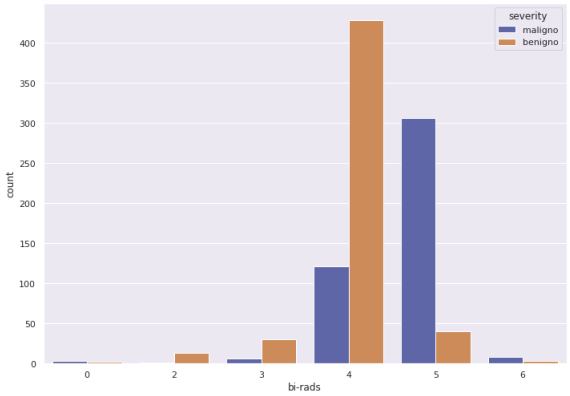
1.2 Gráficas de curva ROC

La curva ROC es un método frecuente para comparar la proporción de verdaderos positivos (tumores malignos; beneficio) frente a la de falsos positivos(tumores benignos etiquetados como malignos; coste). Para este estudio se han visualizado conjuntamente las curvas ROC de cada modelo entrenado con el dataset original (solo se han procesado los valores nulos). Dicha gráfica se muestra a continuación:

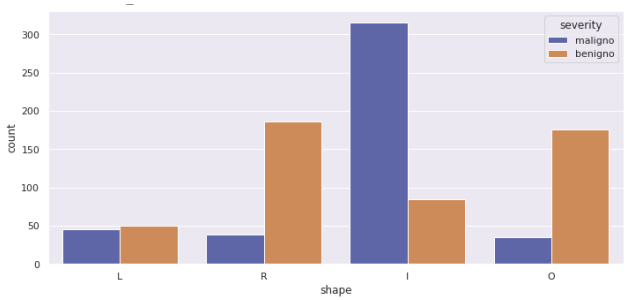


1.3 Análisis de los atributos

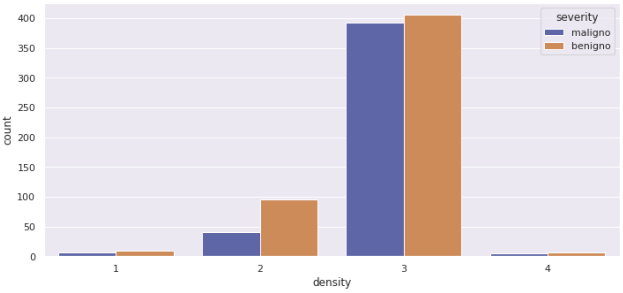
Para cada parámetro de nuestro dataset, vamos a visualizar un histograma que represente el número de tumores benignos o malignos dependiendo de cada valor de dicho parámetro. El objetivo es obtener una idea general de la dependencia del resultado de la clasificación con cada una de las variables explicativas. A continuación se muestran los histogramas con los resultados.



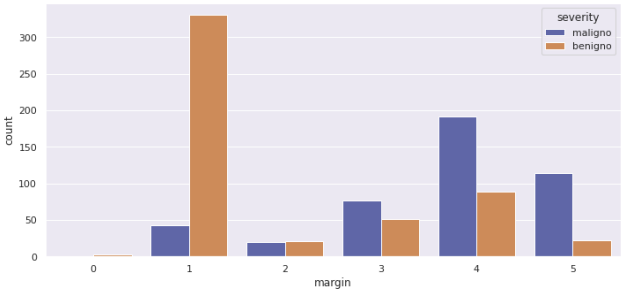
BI-RADS



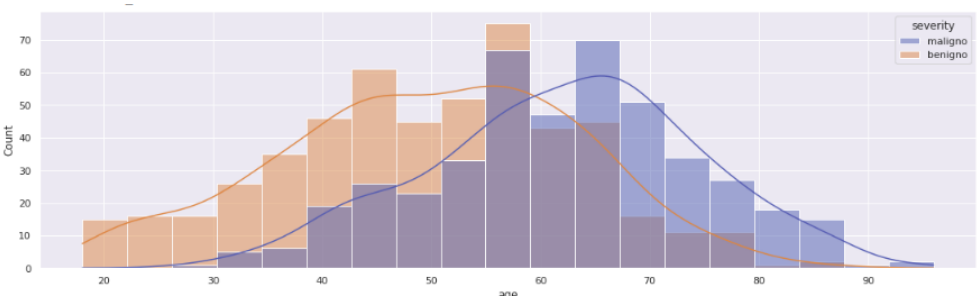
Forma



Densidad de la masa tumoral



Margen de masa



Edad (las líneas continuas marcan la tendencia de cada histograma)

Observamos los siguientes patrones:

- Una categoría BI-RADS de 4 indica que el tumor probablemente será benigno, mientras que una de 5 casi asegura su malignidad.

- Una forma irregular presenta una alta probabilidad de pertenecer a un tumor maligno, mientras que si su forma es redonda u ovalada probablemente sea benigno.
- La densidad presenta un valor de 3 en la mayor parte de las filas de la tabla de datos (luego está enormemente desbalanceada) y, para casi todos sus posibles valores, las cantidades de tumores etiquetados como benignos y malignos están relativamente equilibradas. Esto indica que existe muy poca relación entre la densidad del tumor y su malignidad o benignidad.
- Un margen de masa de tipo 1 corresponderá probablemente a un tumor benigno mientras que si es de tipo 5, aunque de estos hay una menor cantidad, probablemente el tumor se descubra como maligno.
- Las cantidades de tumores benignos y malignos se distribuyen sobre las distintas edades de forma similar a una normal, pero con medias desplazadas. La distribución de los tumores benignos se encuentra desplazada hacia edades más jóvenes, mientras que en las edades más avanzadas la probabilidad de que el tumor sea maligno aumenta.

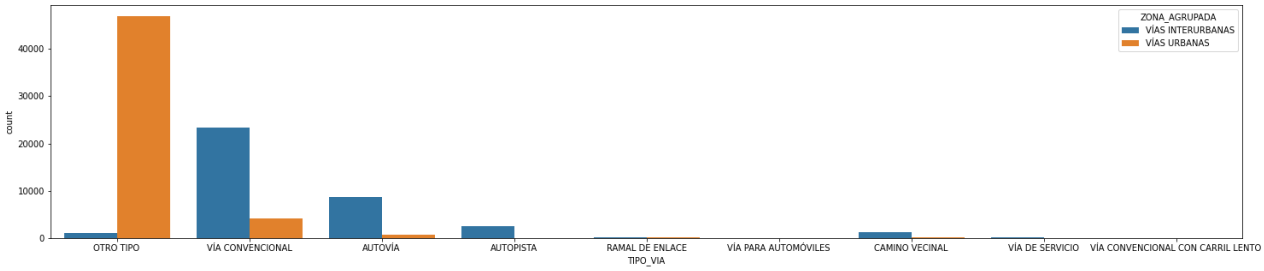
2 Apartado 2: Segmentación

2.1 Introducción

En este apartado, simularemos un trabajo realizado por una empresa aseguradora que trata de conocer mejor la naturaleza de los accidentes de tráfico ocurridos por todo el territorio español. Para ello, usaremos un conjunto de datos proporcionado por la DGT que incluye un gran número de variables sobre los 89.519 accidentes ocurridos en España a lo largo del año 2013. El tipo de estudio a analizar será el de un análisis de agrupamientos comparativo. Es decir, justificaremos el interés en un aspecto de los accidentes de tráfico, diviremos el conjunto de datos con respecto a ese parámetro, buscaremos agrupamientos usando diferentes técnicas del **análisis *cluster*** y compararemos los resultados obtenidos en cada caso.

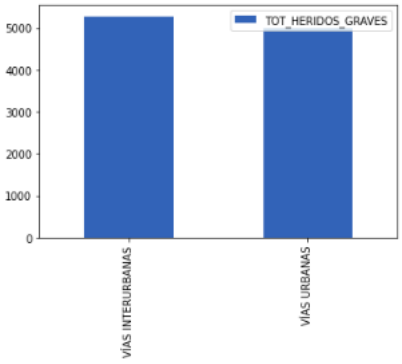
El análisis *cluster*, agrupamiento o segmentación es un problema propio de la minería de datos en el que se busca encontrar similitud entre las instancias de un conjunto de datos, basándonos en una forma prefijada de medir "distancias" entre los distintos elementos. De forma natural, la comparación de las instancias de datos basada en distancia lleva a considerar dichos elementos como "cercanos" o "alejados" unos de otros, lo que permite la segmentación del dataset que da nombre a este problema. El agrupamiento resultante es útil, puesto que da una información sobre la estructura del conjunto de datos que puede servir a expertos en una innumerabilidad de áreas a encontrar clasificaciones ocultas en un conjunto de elementos, o a confirmar una teoría o modelo anterior.

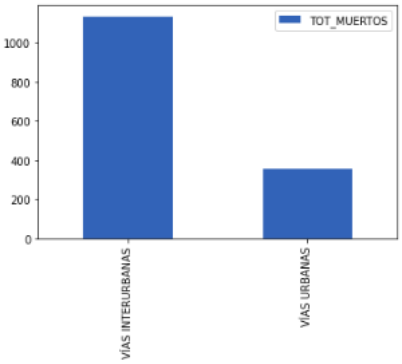
En nuestro caso, vamos a buscar grupos en los accidentes de tráfico separados por zona en la que ocurran. Esto es, dividiremos el conjunto de datos en accidentes ocurridos en **vías urbanas** y accidentes ocurridos en **vías interurbanas**. La causa es sencilla de entender, puesto la conducción y los elementos implicados en la misma difieren enormemente cuando se conduce dentro de una ciudad/municipio y cuando se sale de la misma (velocidad, tipo de carretera...). La siguiente gráfica muestra los tipos de vías en los que ha habido accidentes registrados, comparando aquellos en vías urbanas e interurbanas:



Los datos anteriores refuerzan nuestra hipótesis. La mayor parte de accidentes ocurridos en vías interurbanas ocurren en vías convencionales, autovías y autopistas. Los accidentes en vías urbanas suelen ocurrir en otro tipo de vías no convencionales, y los casos en autopistas y autovías son meramente residuales.

El número de accidentes ocurridos en vías urbanas es de 52.222 (58,34%), mientras que los accidentes en vías interurbanas son 37.297(41,66%). Existe, por tanto, un desajuste entre ambas, pero hay un número de instancias suficientes en cada caso como para estudiarlos de forma independiente. Teniendo en cuenta el dato anterior, llaman la atención las siguientes estadísticas:





Vemos que el total de muertos de los accidentes en vías urbanas supera a aquellos en vías interurbanas por diferencia. Además destaca el hecho de que ambas categorías estén parejas en cuanto a heridos graves considerando que hay un 40% más de accidentes en vías urbanas con respecto al número de casos en vías interurbanas. De nuevo, esto indica diferencias palpables entre ambos casos que justifican nuestro estudio comparado entre ambos.

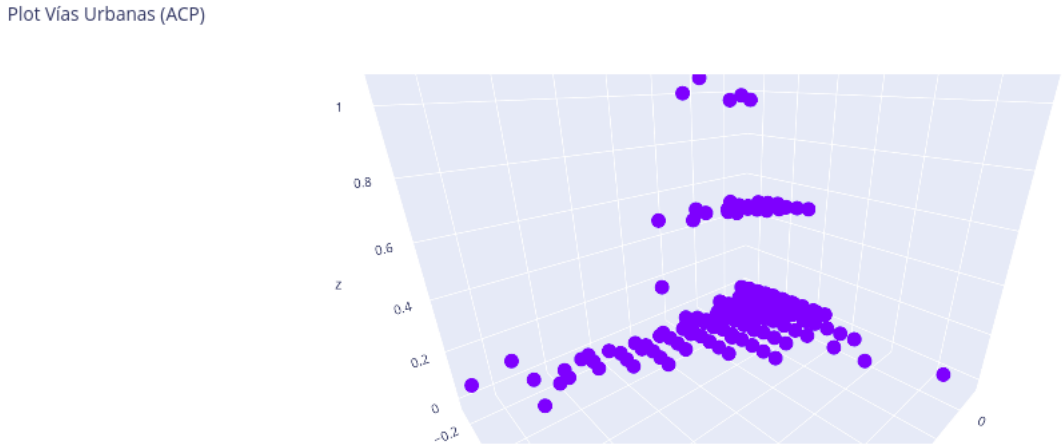
En lo que respecta a la metodología, aplicaremos dos algoritmos de segmentación diferentes: el ***K-means*** y el **agrupamiento jerárquico aglomerativo**. Se tendrán en cuenta únicamente las variables numéricas proporcionadas: el **total de víctimas**, los **números de heridos leves y graves**, el **total de vehículos implicados** y la **cantidad de muertos**. Para ambos casos, compararemos los resultados para distintos valores de k (número de *clusters*, entre 2 y 5) y mediremos la bondad de la segmentación aplicando dos medidas diferentes, la de **Calinski-Harabasz** y la de **Silhouette**, ambas ampliamente utilizadas en el campo.

2.2 Caso 1: vías urbanas

Una vez restringidos a los accidentes en vías urbanas, nos gustaría visualizar el conjunto de datos. No podemos hacerlo directamente puesto que los atributos numéricos, ya mencionados, sobre los que vamos a aplicar las técnicas de clustering son 5, y no podemos representar gráficamente un conjunto de 5 dimensiones. Sin embargo, probaremos la opción de calcular las componentes principales del conjunto de datos y representarlo en las 2 ó 3 dimensiones que mejor expliquen la varianza del conjunto.

Una vez calculadas las componentes principales, vemos que su porcentaje de varianza explicada es 0,5990 en el caso de la primera, 0,1721 la segunda y 0,1338 la tercera. Las 3 componentes principales explican, por tanto, más del 90% de la varianza del conjunto, lo que supondremos suficiente para hacernos una idea de cómo se agrupan sus elementos.

Los accidentes ocurridos en vías urbanas se distribuyen en sus 3 componentes principales de la siguiente forma:

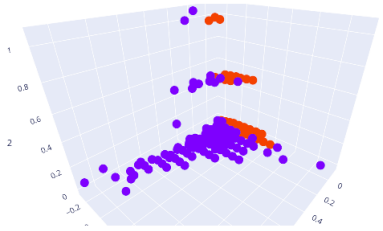


2.2.1 K-means

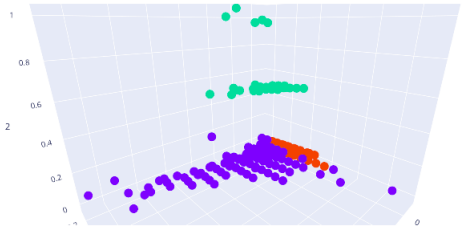
Procedemos a aplicar el K-medias utilizando el paquete correspondiente de la biblioteca Scikit-Learn. A modo de anotación, para cada valor de K se ha aplicado el algoritmo K-medias 100 veces, y nos hemos quedado con aquella solución que da una mejor inercia (suma de la distancia al cuadrado de cada instancia a su correspondiente *cluster*). Esto se ha hecho a través del parámetro `n_init` al instanciar la clase `KMeans` de Scikit-Learn que ejecuta el algoritmo.

Para cada valor de K, mostraremos los resultados en una gráfica como la de la sección anterior:

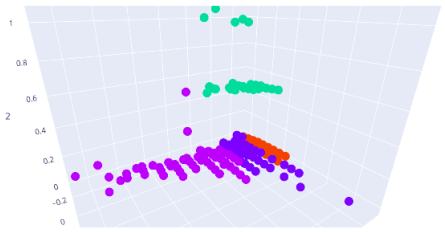
K = 2



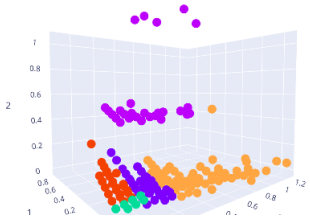
K = 3



K = 4



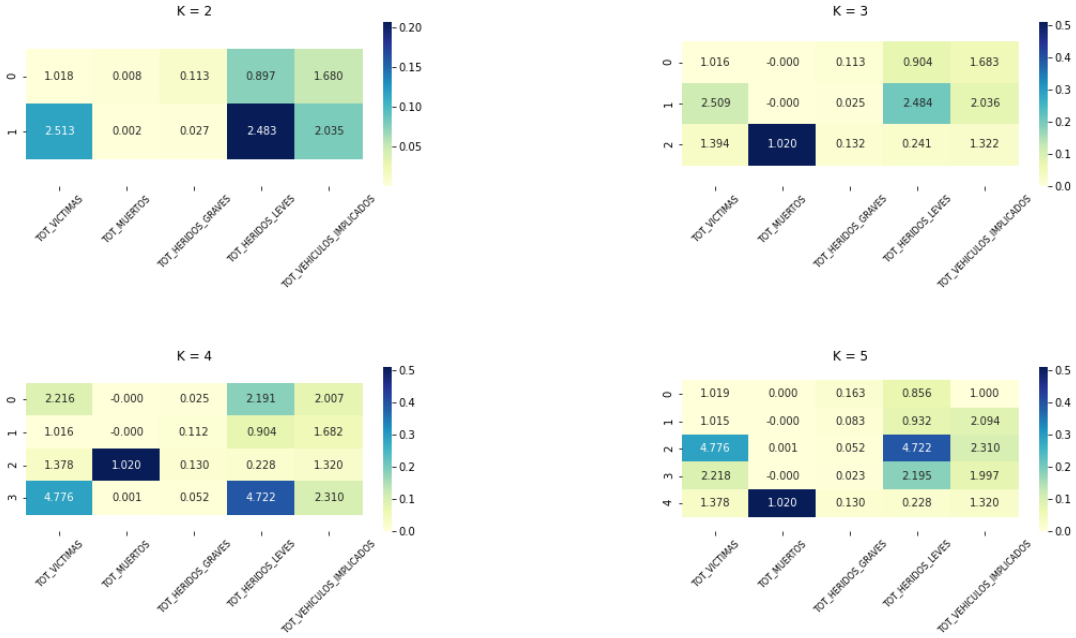
K = 5 (rotado)



En la siguiente tabla mostramos las medidas tomadas sobre el agrupamiento resultante para cada valor de K:

K	Silhouette	Calinsky-Harabasz
2	0.6038274224598406	31943.592964312647
3	0.6186178663600829	27219.744201358662
4	0.6003596510819257	30629.99613688806
5	0.6731184154821989	36913.874156197926

Un valor de K de 5 es el que obtiene las medidas más altas. A continuación visualizaremos los mapas de calor en representación de los centroides escogidos para cada valor de K:



Vemos que cualquiera de las opciones obtenidas presenta considerables diferencias entre los centroides escogidos, justificando una segmentación del conjunto de datos. El análisis de los centroides se realizará al final de este caso, pudiendo comparar previamente estos resultados con los obtenidos por el clustering aglomerativo.

2.2.2 Clustering Aglomerativo

La técnica de agrupamiento que probaremos ahora es un tipo de clustering jerárquico denominado aglomerativo. Este algoritmo sigue un proceso *bottom-up*: cada instancia empieza como un *cluster* en sí misma, y a lo largo de sucesivas etapas se van fusionando aquellos grupos satisfaciendo un criterio dado. El criterio utilizado en esta práctica será el de *"ward"*, que busca minimizar la suma de las diferencias entre los elementos de un mismo *cluster* al cuadrado.

La ventaja de este proceso es que, una vez se ha construido el árbol completo y si se decide guardar la estructura obtenida, se puede dividir el conjunto para cualquier valor de K sin tener que repetir el proceso. Este ha sido el procedimiento seguido y el código usado para dicha tarea se muestra a continuación:

```
from sklearn.cluster import AgglomerativeClustering
results = AgglomerativeClustering(2, memory='clustcache').fit(data_vias_urbanas_norm)

labels = {}
centroids = {}
```

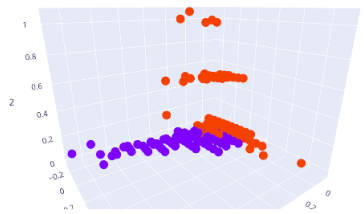
```
for k in [2,3,4,5]:
        results = AgglomerativeClustering(k, memory='clustcache').fit(data_vias_urbanas_norm)
        labels[k] = results.labels_
        centroids[k] = calcular_centroides(
                norm(raw_data.loc[raw_data["ZONA_AGRUPADA"]=="VÍAS URBANAS"][atributos]),
                results.labels_)
        df_k_clusters = pd.DataFrame(data_vias_urbanas_red)
        df_k_clusters["Label"] = [ "{}".format(l) for l in results.labels_ ]
        fig = px.scatter_3d(df_k_clusters, x=0, y=1, z=2,
                        title="K = {}".format(k), color="Label")
        fig.show()
```

Donde las últimas 4 líneas dentro del bucle `for` están destinadas a la visualización del conjunto segmentado.

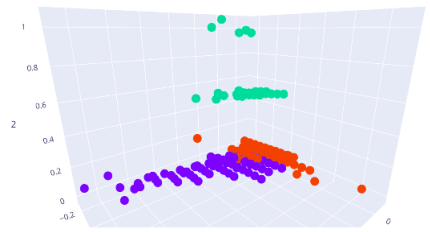
Es importante destacar que el mayor coste computacional del algoritmo de agrupamiento se da al instanciar la clase `AgglomerativeClustering` (línea 2). Cada vez que se llama a esta función dentro del bucle solamente se lee la caché generada en el fichero *clustcache*.

Procedemos a mostrar la segmentación hecha para cada valor de K:

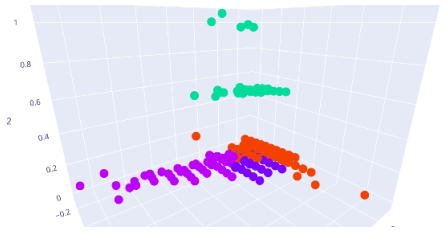
K = 2



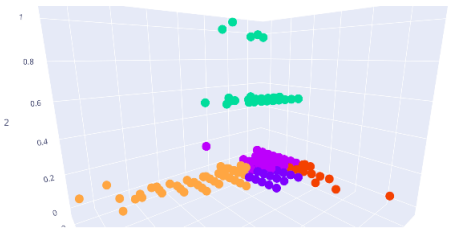
K = 3



K = 4



K = 5

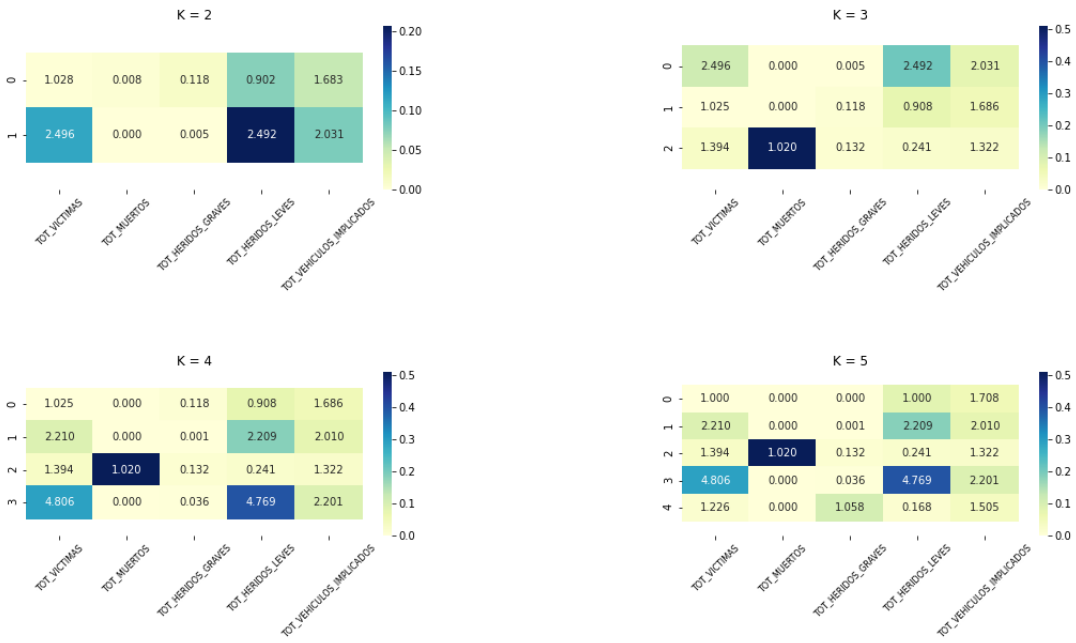


Y, a continuación, las puntuaciones obtenidas con cada valor de K:

K	Silhouette	Calinsky-Harabasz
2	0.5969261993026124	30731.91258220163
3	0.6162067241199223	26413.63565692129
4	0.5988853343499478	29276.18147847532
5	0.6525631958067118	34523.77756682497

De nuevo, un valor de K=5 es el que obtiene mejores resultados, lo cual refuerza nuestro análisis exploratorio y parece indicar que, en efecto, hay 5 grupos diferenciables dentro de los casos en vías urbanas. También observamos que las puntuaciones generales del algoritmo jerárquico aglomerativo son menores que las del *K-means*.

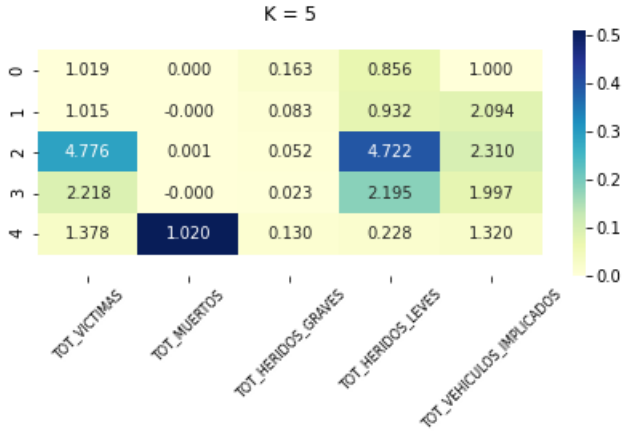
A continuación visualizamos los *heatmaps* con los centroides de cada uno de los grupos para cada valor de K:



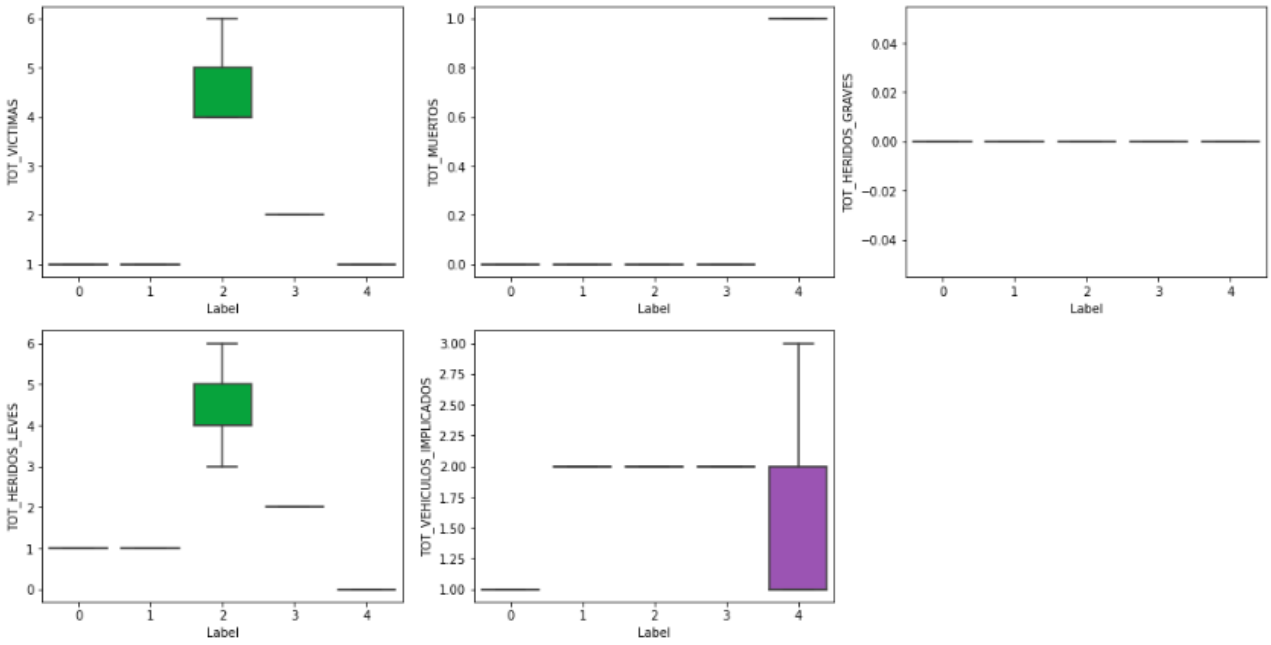
Es importante destacar el hecho de que los resultados obtenidos son muy similares a los del *K-means*, lo cual vuelve a indicar que los resultados obtenidos no son fruto de usar un algoritmo inadecuado, sino que realmente corresponden con distintos grupos que podemos encontrar en el dataset.

2.2.3 Interpretación de la segmentación

Hemos estudiado los accidentes de tráfico sucedidos en vías urbanas y hemos realizado un análisis *cluster* sobre los mismos empleando dos algoritmos: el *K-means* y el agrupamiento aglomerativo. Ambos apuntan a la siguiente conclusión: entre los valores de K estudiados (entre 2 y 5), un número de 5 grupos es el que parece obtener los mejores resultados de acuerdo a las medidas de Calinski-Harabasz y Silhouette. Basándonos en estas mismas medidas, elegiremos el resultado obtenido por el K-means para un valor de K=5 y analizaremos los centroides resultantes.



Para completar la información presentada incluimos los siguientes diagramas de cajas para cada variable comparando los distintos grupos:



Procedemos a enumerar los siguientes aspectos destacables de cada *cluster*:

1. Los accidentes del primer grupo implican a un único coche y presentan una cantidad baja en víctimas y heridos, y no provoca muertes de una forma estadísticamente significativa. Este grupo presenta 15.558 casos.

2. Los accidentes del segundo grupo son similares en consecuencias a los primeros, pero tienden a tener 2 coches implicados en el accidente. Como el número de víctimas ronda en torno a 1, asumimos que es común en los accidentes de este tipo el hecho de que sólamente uno de los implicados se vea perjudicado. El número de muertes, de nuevo, es prácticamente nulo. Este grupo presenta 25.963 casos.

3. El siguiente grupo está caracterizado por un número mucho mayor de víctimas, y 2 ó más coches impicados. Estas víctimas suelen acabar levemente heridas. De nuevo, en este tipo de accidentes no hay una cantidad significativa de muertes. Este grupo presenta 1.185 casos.

4. Este grupo es similar al anterior. Sin embargo, aunque el número de vehículos implicados ronda en torno a la misma media, el número de víctimas es de menos de la mitad. En este tipo de accidentes tampoco hay muertos. Este grupo presenta 9.169 casos.

5. El último grupo es, de forma general, el de los accidentes que acaban en muerte (por lo general, la de una sola persona si vemos la media y la desviación). Vemos que tiende a haber entre 1 y 2 víctimas y 1 ó 2 vehículos implicados de media, aunque este grupo presenta una gran varianza en ese aspecto. El número de heridos, tanto leves como graves, es muy bajo, especialmente en comparación con el número de muertos. Concluimos que este tipo de accidentes tienen una mortalidad muy alta. Este grupo presenta 347 casos.

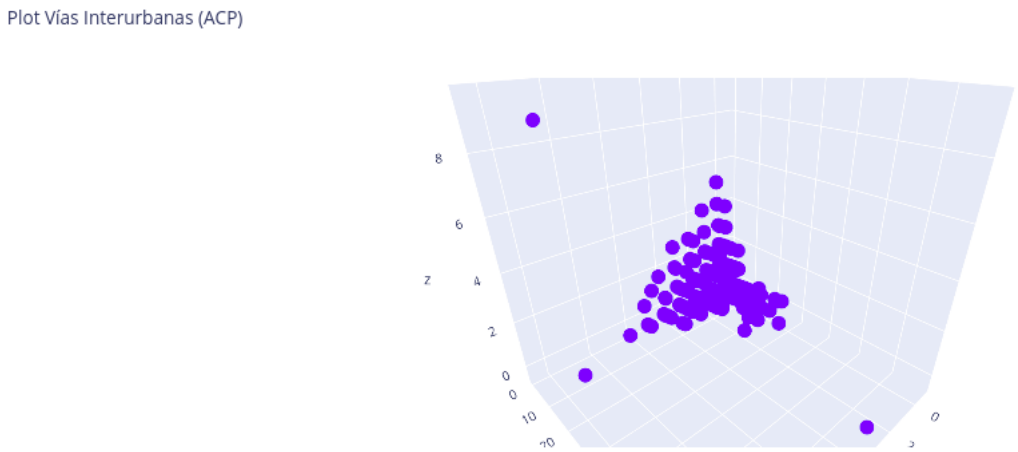
De forma general, también podemos destacar que el número de heridos graves, salvo los muertos en el último grupo, es muy bajo.

2.3 Caso 2: vías interurbanas

En este caso también queremos visualizar los elementos del conjunto de datos y la segmentación realizada, para lo que repetiremos el análisis de componentes principales de la sección anterior, esta vez para el caso de accidentes en vías interurbanas.

Una vez calculadas las componentes principales, vemos que su porcentaje de varianza explicada es 0,5465 en el caso de la primera, 0,2441 la segunda y 0,1217 la tercera. Las 3 componentes principales explican, por tanto, más del 90% de la varianza del conjunto, lo que supondremos suficiente para hacernos una idea de cómo se agrupan sus elementos.

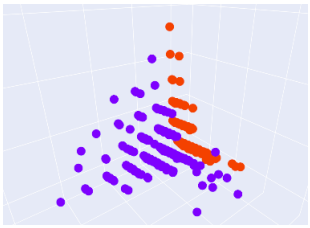
Los accidentes ocurridos en vías urbanas se distribuyen en sus 3 componentes principales de la siguiente forma:

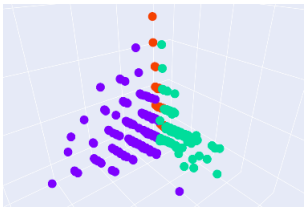


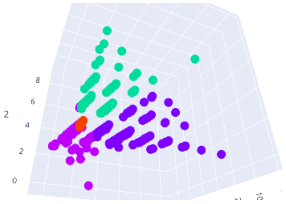
2.3.1 K-means

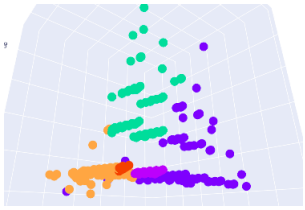
Procedemos a aplicar el K-medias en las mismas condiciones que para el caso 1: un valor de K entre 2 y 5 y el mismo número de ejecuciones, solo cambiando el conjunto de datos sobre el que se aplica.

Para cada valor de K, mostraremos los resultados en una gráfica como la de la sección anterior:

K = 2 

K = 3 

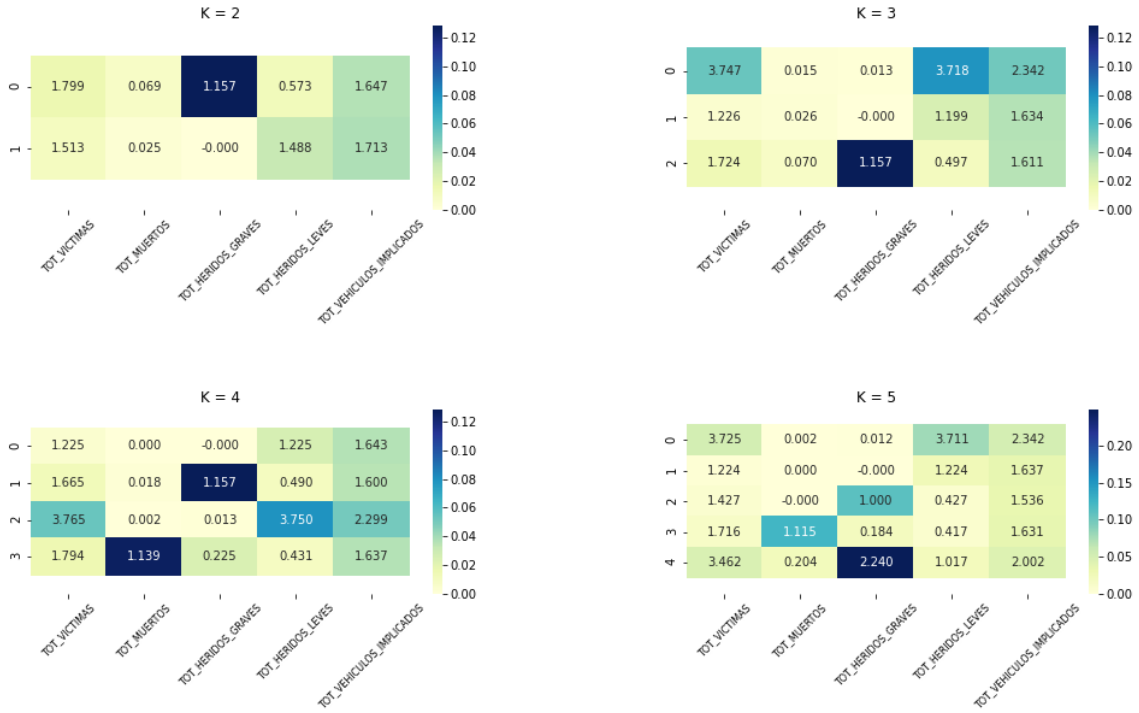
K = 4 (rotado) 

K = 5 (rotado) 

En la siguiente tabla mostramos las medidas tomadas sobre el agrupamiento resultante para cada valor de K:

K	Silhouette	Calinsky-Harabasz
2	0.7176559933335068	31386.16858002395
3	0.6164151855278602	25945.89079892083
4	0.6511249052114755	24356.94591686606
5	0.6623403806170062	25462.961947760883

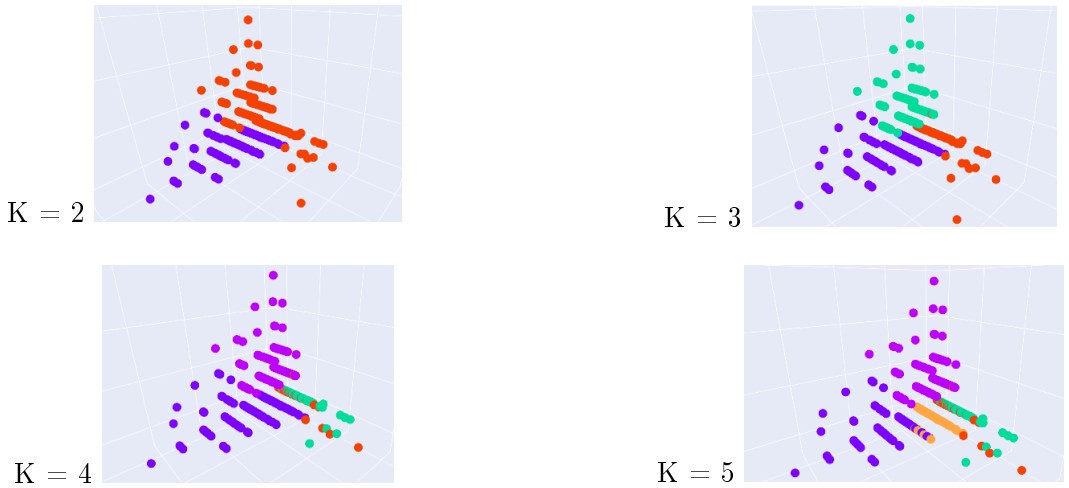
Un valor de K de 2 es el que obtiene las medidas más altas. A continuación visualizaremos los mapas de calor en representación de los centroides escogidos para cada valor de K:



Observamos que para cualquier valor de K se muestran diferencias entre los distintos centroides, más o menos destacables, y que dichos centroides son considerablemente distintos a los observados en el caso 1.

2.3.2 Clustering Aglomerativo

Repetimos el clustering aglomerativo realizado para el caso anterior, esta vez sobre el conjunto de accidentes en vías interurbanas. La segmentación hecha para cada valor de K queda así:



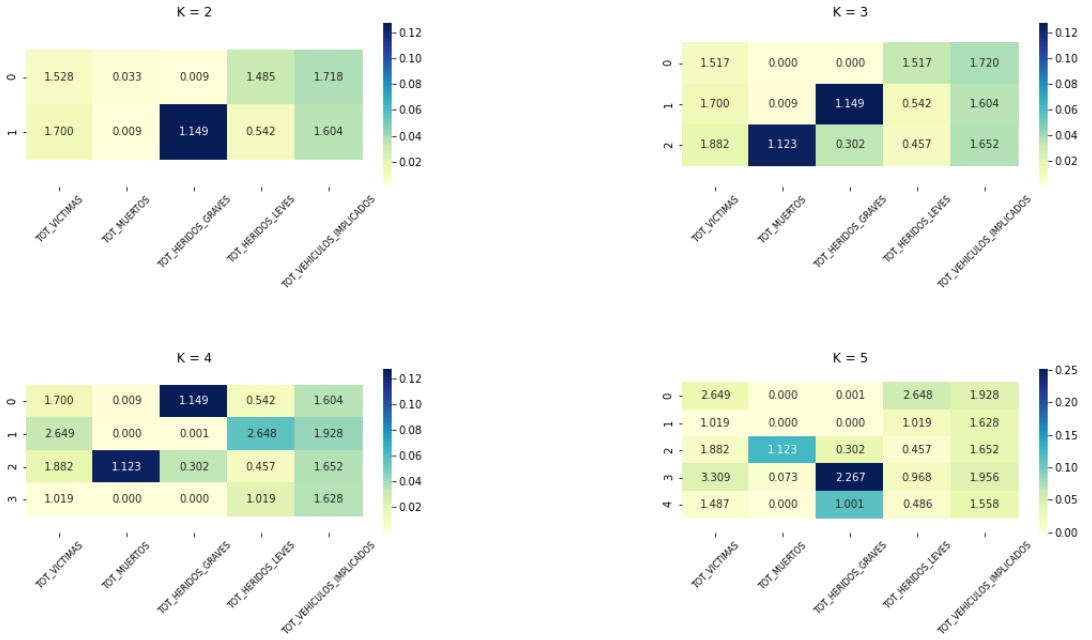
Observamos que, al usar 4 ó 5 *clusters*, aumenta la superposición y entremezclado de los mismos.

A continuación vemos las puntuaciones obtenidas con cada valor de K:

K	Silhouette	Calinsky-Harabasz
2	0.7028659361263889	27420.55762894603
3	0.7274855636316319	21324.00500712306
4	0.553976738595942	21822.229857465456
5	0.5657015423161276	21867.78893146342

En este caso sí que existe un conflicto, tanto al comparar las dos medidas utilizadas, que dan como ganadores a valores de K distintos, como al cotejar estos resultados con los obtenidos por el K-medias, puesto ambos dan como claros ganadores a valores de K diferentes. En cualquier lugar, los dos números de *clusters* a tener en cuenta son K=2 y K=3.

Saldremos de dudas comparando directamente los centroides mediante los *heatmaps* correspondientes a cada K, que se visualizan a continuación:

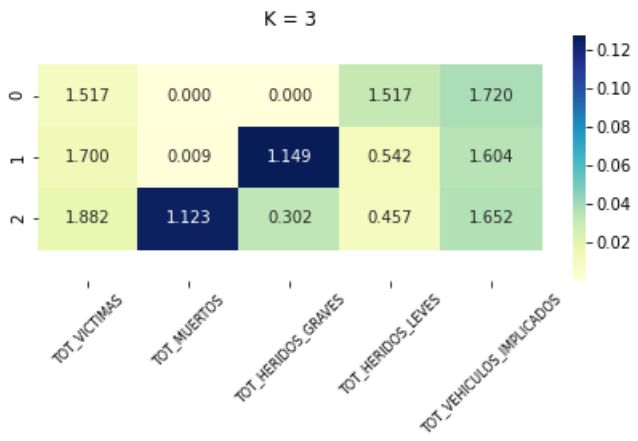


En este caso los centroides para K=2, K=4 y K=5 son relativamente similares a los obtenidos mediante el K-medias, pero para K=3, que además obtiene medidas muy diferentes al usar el clustering aglomerativo, los centroides son bastante diferentes.

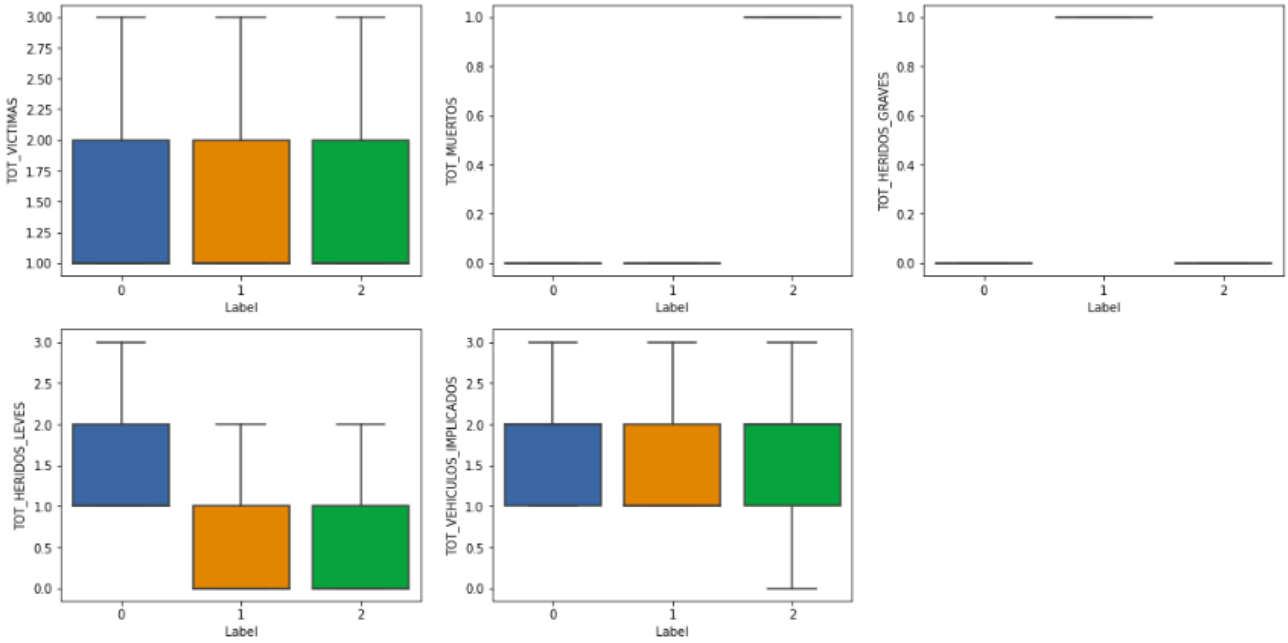
Mientras que para K=3 el K-medias subdividía uno de sus grupos (el grupo con la etiqueta 1) en dos subgrupos, uno con números de víctimas, heridos leves y vehículos implicados más bajos y otro con números más altos, pero sin mayores diferencias entre sí, el clustering jerárquico ha logrado separar un conjunto con una gran cantidad de muertos de los otros subgrupos, prácticamente sin víctimas mortales. Por este motivo, que parece aportar información importante sobre la casuística de los accidentes en vías interurbanas, y por la buena puntuación Silhouette que ha obtenido, esta vez elegiremos la segmentación en tres grupos hecha por el clustering aglomerativo como la mejor para los accidentes en este tipo de casos.

2.3.3 Interpretación de la segmentación

Vistos los resultados de cada algoritmo y dada la argumentación anterior, revisaremos los grupos calculados por el clustering aglomerativo y trataremos de describir las diferencias entre cada uno de ellos. Recordamos que los centroides finales son los siguientes:



Para completar la información presentada incluimos los siguientes diagramas de cajas para cada variable comparando los distintos grupos:



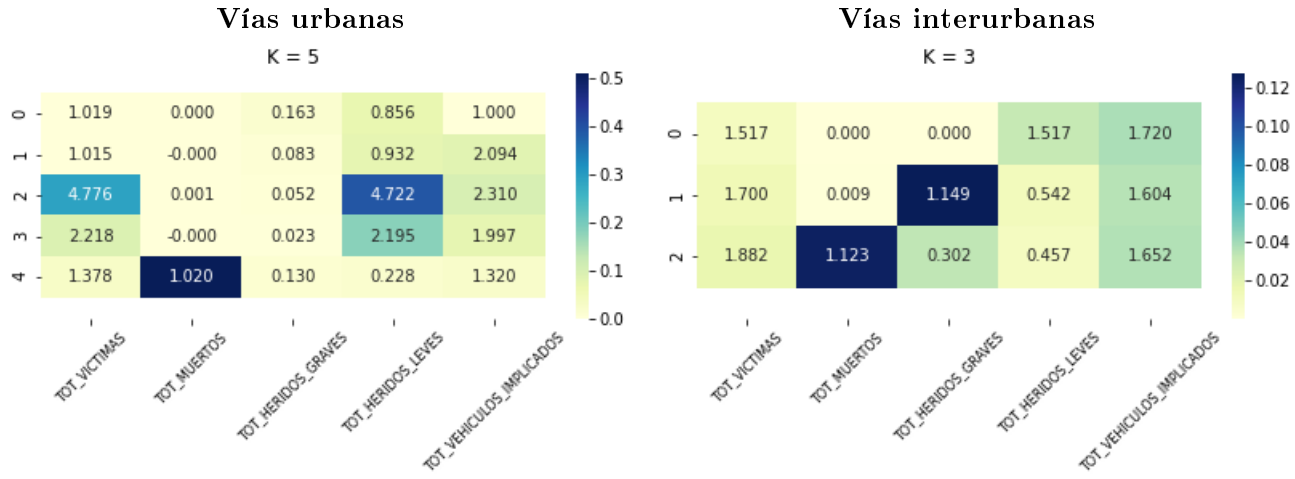
Procedemos a enumerar los siguientes aspectos destacables de cada *cluster*:

1. En el primer grupo el total de heridos leves es más elevado que en el resto. Estos accidentes no presentan heridos graves ni muertos. Este grupo presenta 31.996 casos.
2. Los accidentes del segundo grupo se caracterizan por presentar heridos graves (entre 1 y 2 heridos graves, de media). No tiene muertos y el número de heridos leves es muy bajo. Este grupo presenta 4.327 casos.
3. El tercer grupo es el único que presenta accidentes con una cantidad significativa de muertos. Mueren entre 1 y 2 personas en este tipo de accidentes. Presenta heridos graves, pero muy por debajo del grupo anterior. El número de heridos leves es también bajo. Este grupo presenta 974 casos.

Aunque, ciertamente, el número de instancias en el primer grupo es mucho mayor que en los otros dos, el segundo y el tercer grupo presentan un tamaño y unas diferencias suficientemente notables como para que no se los considere *outliers*, sino grupos independientes.

2.4 Análisis Comparativo

En esta sección compararemos ambos casos, viendo los parecidos y resaltando detenidamente las diferencias que las segmentaciones realizadas nos han confirmado o dado a conocer. Revisamos los centroides obtenidos en cada caso:



Hay un hecho que destaca a simple vista al ver los resultados de ambos agrupamientos. Mientras que las instancias estudiadas en el caso 1 han terminado segmentadas en 5 *clusters*, los accidentes analizados en el caso 2 han terminado en 3 *clusters* diferentes. Además, viendo los resultados de los algoritmos empleados, utilizar en el caso 1 un valor de K de 2 ó 3, como se ha barajado para el caso 2, hubiese supuesto unos resultados mucho peores. Al mismo tiempo, para el caso 2 hubiese sido impensable, de acuerdo con las medidas realizadas, concluir nuestro estudio tomando una segmentación con 5 *clusters*. Esto confirma, o al menos refuerza notablemente, la hipótesis que se afirmó al principio del apartado 2 cuando se justificó el estudio de los accidentes dividiendo en zonas urbanas e interurbanas: ambos casos presentan diferencias importantes y sus respectivos conjuntos poseen una estructura distinta. De no ser así, se esperaría que esta variable no tuviese influencia en las demás y que los accidentes en zonas urbanas e interurbanas siguieran la misma distribución, lo que los llevaría a agruparse en subconjuntos de naturaleza similar.

La segunda desigualdad es la relativa a los accidentes con heridas graves. Mientras que en el caso 2, con 3 grupos ha sido suficiente para encontrar un subconjunto del dataset caracterizado por presentar heridas graves, en el caso 1 no ha sucedido de la misma forma incluso obteniendo 5 grupos distintos. Por un lado, esto puede tener relación con el hecho que señalamos en la Introducción sobre la menor proporción de heridos graves en los accidentes en vías urbanas con respecto a los accidentes en interurbanas. Sin embargo, esta diferencia no es suficientemente elevada como para justificar por sí sola la disimilitud entre ambas segmentaciones. Posiblemente, en el caso de las vías urbanas, los accidentes con heridos graves se distribuyan de forma más heterogénea (difíciles de agrupar), entremezclados entre el resto de accidentes. Esto explica el hecho de que, si miramos el *heatmap* final con los centroides escogidos para el caso 1, veamos que la columna de heridos graves es distinta de 0 en cualquiera de los 5 *clusters* formados.

Por otro lado, aunque ya sabíamos que la mayoría de los percances no conllevan consecuencias directas graves, vemos que los agrupamientos obtenidos en ambos casos confirman este hecho. En el caso de las vías urbanas, el grupo correspondiente a los incidentes que acaban en la muerte de alguien consta de 347 instancias (un 0,66%). Sobre el caso en vías interurbanas, entendemos que el primer *cluster* engloba los accidentes análogos a los 4 primeros grupos del caso anterior, puesto que no presenta muertos ni heridos de gravedad y su varianza en número de víctimas, de heridos leves y de vehículos implicados es elevada. Este grupo tiene 31.996 casos (un 85,79%).

3 Bibliografía

- *The Elements of Statistical Learning*, J.H. Friedman, R. Tibshirani, T. Hastie.
- Sección de Scikit-Learn sobre *clustering*.
- Documentación de la API de Scikit-Learn.
- Artículo en *Medium* sobre el análisis del número de *clusters*.
- Representación de un *dataframe* mediante histogramas con Pandas.
- Artículo en *towardsdatascience.com* sobre el *clustering* jerárquico.
- Guía de uso de la función scatter_3d de plotly usada para visualizar los gráficos en 3D.