Práctica 2: Visualización y Segmentación:

UNIVERSIDAD DE GRANADA

Luis González Romero, XXXXXXXXX, luisgonromero@correo.ugr.es

Escuela Técnica Superior de Ingeniería informática y Telecomunicaciones

17 de diciembre de 2020

Práctica 2: Visualización y Segmentación

Memoria sobre la segunda práctica de la asignatura Inteligencia de Negocio cursada en la ETSIIT, UGR.

Luis González Romero, XXXXXXXXX, luisgonromero@correo.ugr.es

Índice

Índice de figuras

Índice de tablas

1. Visualización

1.1. Visualización de las medidas de la primera práctica

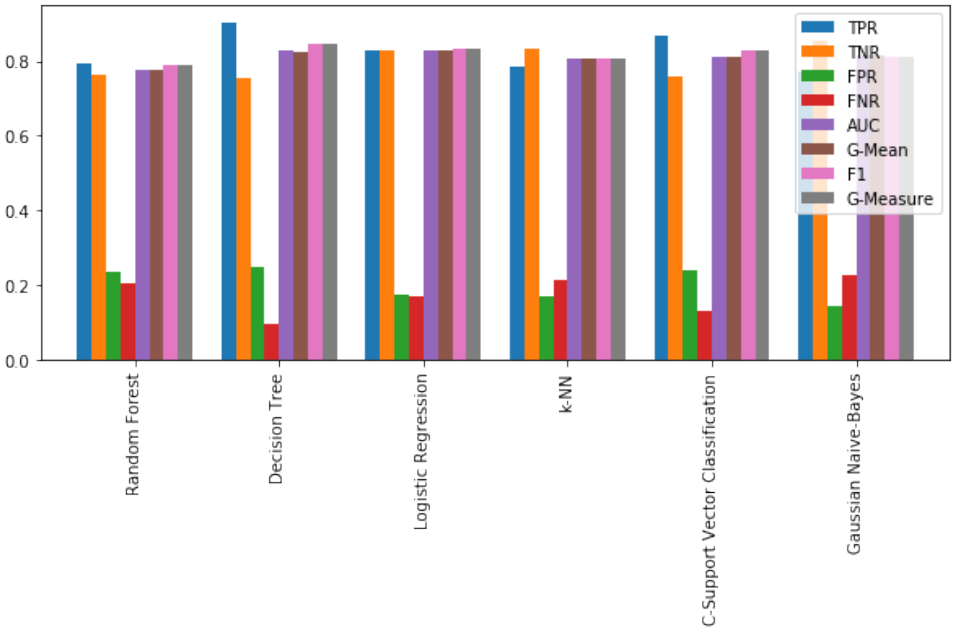


Figura 1: Medidas obtenidas eliminando valores perdidos

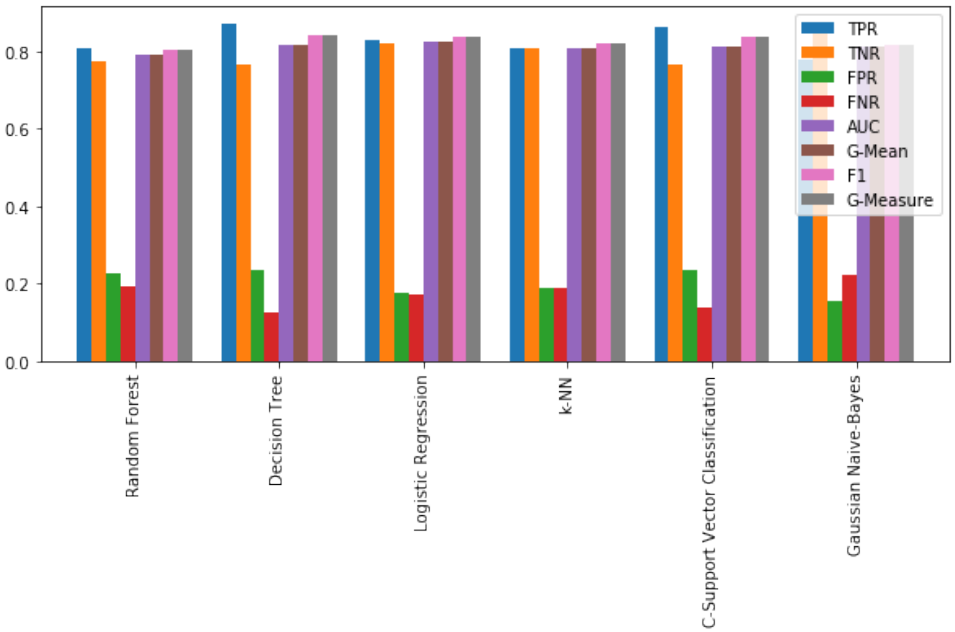


Figura 2: Medidas obtenidas tratando los valores perdidos usando la media

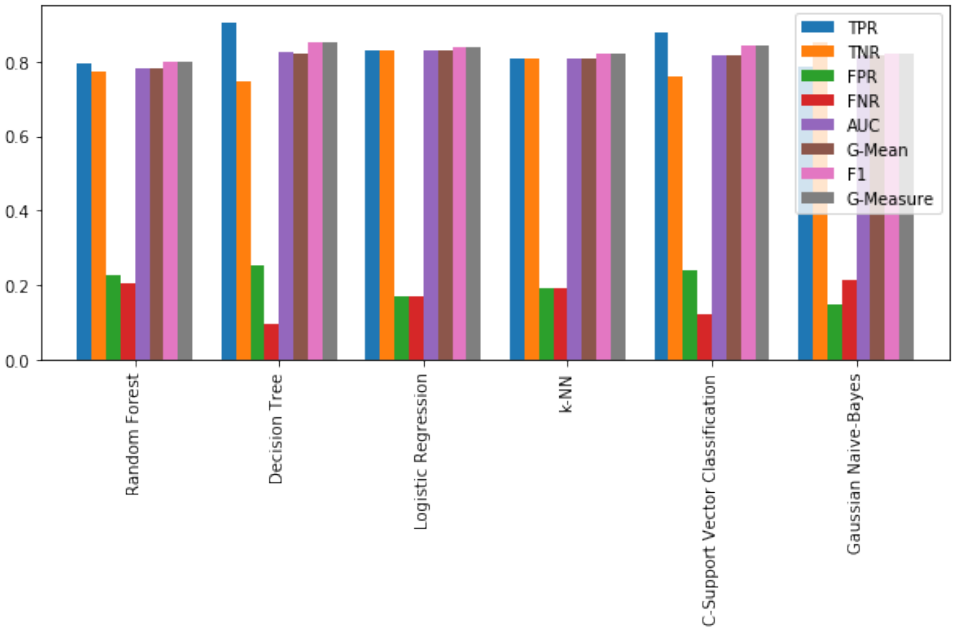


Figura 3: Medidas obtenidas tratando los valores perdidos usando la mediana

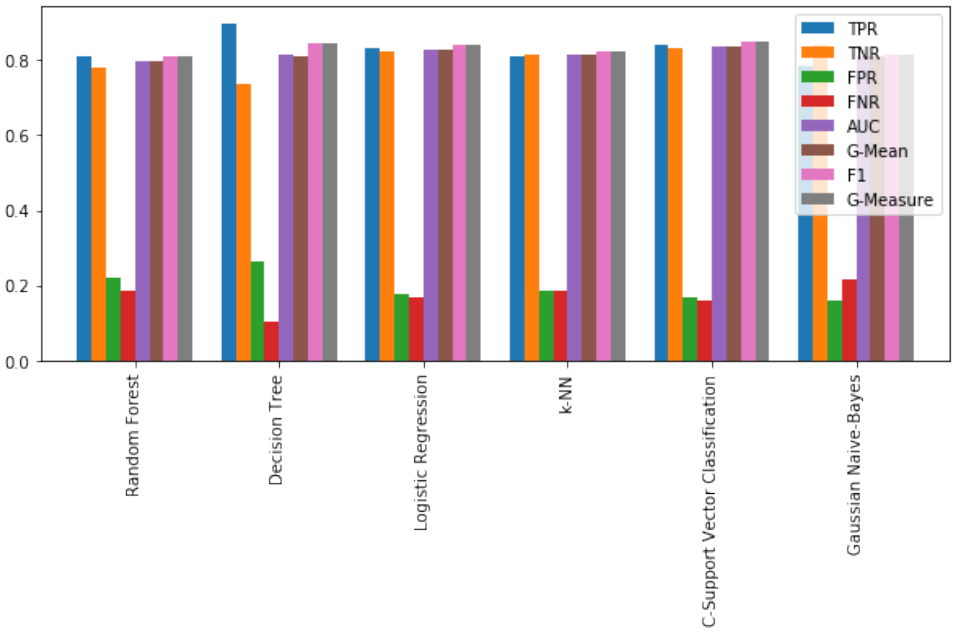


Figura 4: Medidas obtenidas tratando los valores perdidos usando el valor más frecuente

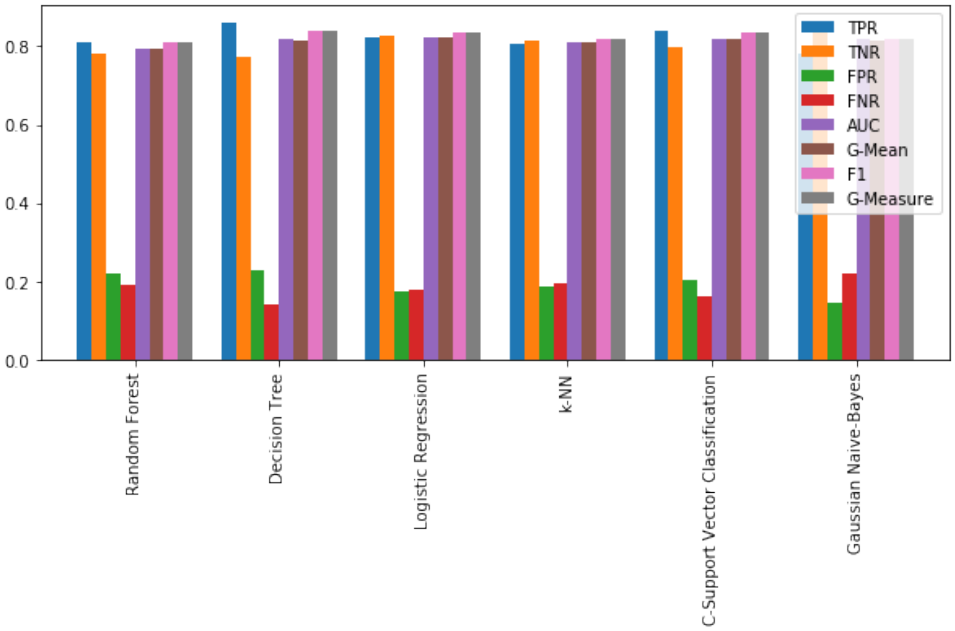


Figura 5: Medidas obtenidas tratando los valores perdidos usando k-NN con 11 vecinos

1.2. Curvas ROC

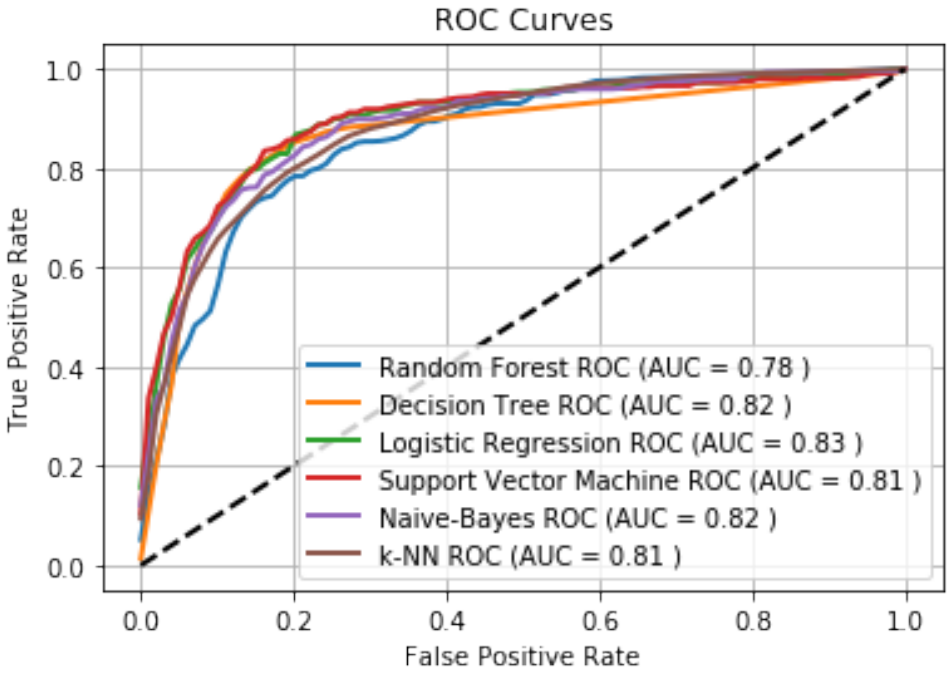


Figura 6: Curva ROC de los distintos modelos de la primera práctica

2. Segmentación

2.1. Introducción

En esta segunda práctica se ha estudiado un problema con datos sobre *accidentes de tráfico* de la Dirección General de Tráfico(DGT) en el año 2013. Los datos recogidos por la DGT contienen variables que caracterizan los accidentes, y se intentará encontrar grupos de accidentes similares dentro del dataset. Estos datos incluyen información entre los años 2008 y 2013, además de más de 30 variables. Nuestro dataset contendrá las siguientes variables, entre otras:

- MES
- HORA
- DIASEMANA
- PROVINCIA
- COMUNIDAD_AUTONOMA
- TOT_VICTIMAS
- TOT_MUERTOS
- TOT_HERIDOS_GRAVES
- TOT_HERIDOS_LEVES
- TIPO_VIA
- TIPO_INTERSEC
- PRIORIDAD
- LUMINOSIDAD
- TIPO_ACCIDENTE
- DENSIDAD_CIRCULACION

Para abordar esta práctica he usado los siguientes algoritmos:

- K-Means(Partitional)
- Agglomerative Clustering(Hierarchical)

A pesar de que el método jerárquico no necesita un valor de k , se lo paso para comparar con los obtenidos con K-Means para ese mismo de valor de k y porque en tiempo de ejecución será menos costoso.

2.2. Caso de estudio 1

El primer caso de estudio está relacionado con el tipo de intersección sobre el que se produce el accidente. Tenía interés sobre las salidas ya que pensé que quizás en salidas de autovía(por la velocidad que se circula en estas aun cruzando por poblado) y otro tipos de vías se verían diferencias en los accidentes que se producirían. Además, en mi pueblo habitual, se hicieron unos pasos de peatones en la propia salida de una y siempre me pareció algo bastante peligroso.

2.2.1. Tipo de intersección - Enlace de salida

Lo primero que pude ver es que la mayoría de accidentes se producen en autovía, asi que podría responder la pregunta que me estaba haciendo; seguidos de ramal de enlace, con autopista y via convecional con cifras similares.

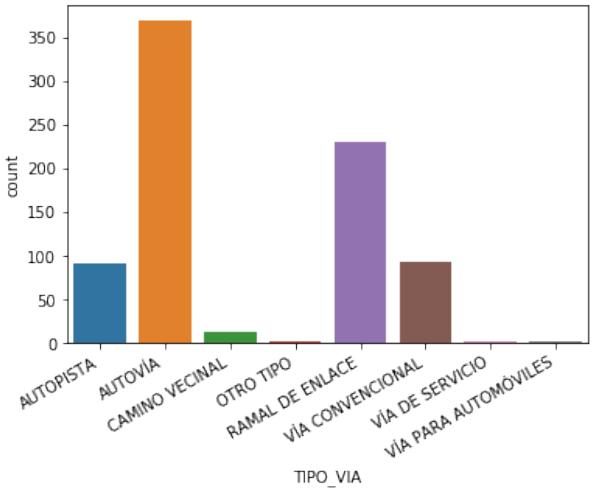
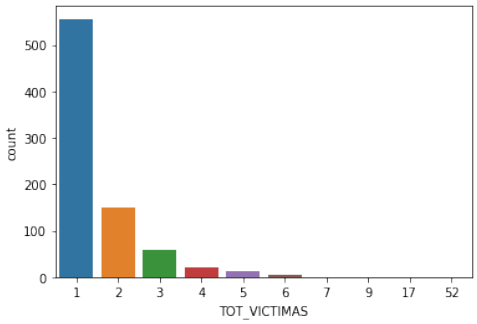
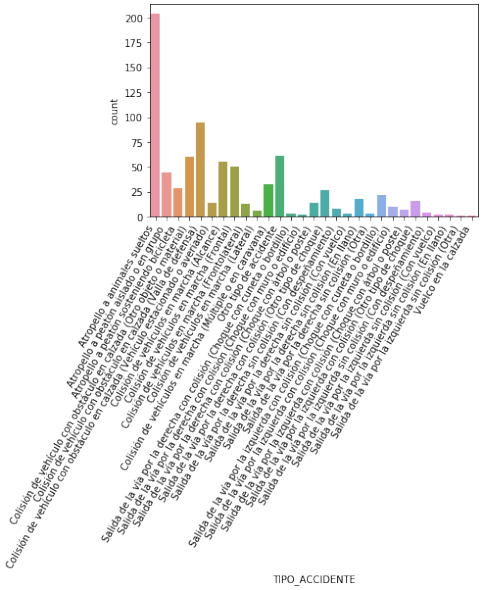


Figura 7: Tipo de vía en intersecciones de tipo 'enlace de salida'

Viendo las victimas totales de los accidentes que se producen pense inmediatamente que se tratarían de atropellos, cosa que pude comprobar al mirar los tipos de accidentes de los que se trataban. La gran mayoría de animales sueltos y en mucha menor cantidad de peatones, también suman bastantes los casos de colisión con obstáculos seguido de colisiones de vehículos. También se puede apreciar una importante cantidad de accidentes por salida de vía(cosa que me llama la atención por un *artículo* que leí de la DGT que decía que el tipo de accidente que registra más fallecidos, es de este tipo).



(a) Número de victimas en intersecciones de tipo 'enlace de salida'



(b) Tipo de accidentes en intersecciones de tipo 'enlace de salida'

Figura 8: Víctimas y tipos de accidente en intersecciones de tipo 'enlace de salida'

2.2.1.1 Resultados de la segmentación

Para aplicar clustering he usado valores de $k \in [3, 7]$, ya que no quería probar con valores grandes de k.

Para evaluar la calidad del agrupamiento obtenido, se usarán las siguientes medidas:

- Coeficiente de Silhouette: nos dice cómo de similares son los objetos de un mismo cluster comparado con otros clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

  Sea $a(i)$ la distancia media del objeto i al resto de objetos de su cluster. Sea $b(i)$ la mínima distancia media del objeto i al resto de objetos del resto de clusters. Toma valores en [-1,1], cuanto más cercano a 1, mejor agrupados están los clusters.

  La media de todos los $s(i)$ es el coeficiente silhoutteque mide la calidad global del agrupamiento.

- Índice de Calisnki-Harabasz: razón entre la dispersión intra-cluster y la dispersión inter-cluster. Cuanto mayor es el valor, mejor es el agrupamiento.

$$\mathrm{CH}(P) = \frac{(N - |P|)\,\mathrm{inter}_{CH}(P)}{(|P| - 1)\,\mathrm{intra}_{CH}(P)}$$

N es el número de objetos
$|P|=k$ es el número de clusters

$$\mathrm{inter}_{CH}(P) = \sum_{C \in P} |C|\,\mathrm{d}(\bar{C}, \bar{X}) \text{ e } \mathrm{intra}_{CH}(P) = \sum_{C \in P} \sum_{x \in C} \mathrm{d}(x, \bar{C})$$

$k = 3$	**K-Means**	**Agg-Clustering**
Silhouette	*0,6214*	***0,6231***
Calinski	***293,1046***	283,6044

$k = 4$	**K-Means**	**Agg-Clustering**
Silhouette	*0,6584*	***0,6590***
Calinski	***335,2021***	333,9225

$k = 5$	**K-Means**	**Agg-Clustering**
Silhouette	*0,6663*	***0,6668***
Calinski	***433,5601***	431,6778

$k = 6$	**K-Means**	**Agg-Clustering**
Silhouette	***0,7593***	*0,7586*
Calinski	***656,0137***	643,6053

$k = 7$	**K-Means**	**Agg-Clustering**
Silhouette	*0,7810*	***0,7817***
Calinski	***844,7296***	825,1405

Tabla 1: Resultados obtenidos con los distintos algoritmos para las intersecciones de tipo 'enlace de salida'

Como podemos ver en cuanto al valor de Silhouette, ambos algoritmos han estado a la par, por lo que poco podemos sacar de él. En cuanto a Calinski, se puede ver como siempre con K-Means se obtiene mejor resultado.

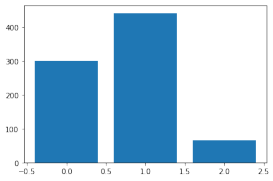
Con estos datos, el mejor agrupamiento se realiza con $k = 7$ de los que he probado. A mayor valor de k, mejor es este segun el índice Calinski.

Parece que seguirá creciendo mientras crece k, pero ya empezarían a dar demasiados numeros de clusters.

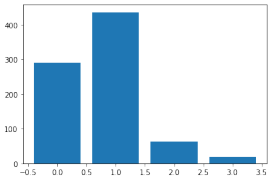
2.2.1.2 Interpretación de la segmentación

Figura 9: Mapas de calor para el algoritmo K-Means, tipo 'enlace de salida'

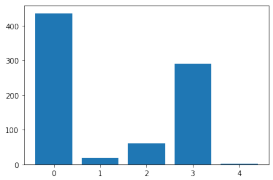
Figura 10: Mapas de calor para el algoritmo Agglomerative Clustering, tipo 'enlace de salida'



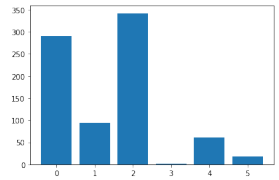
(a) $k = 3$



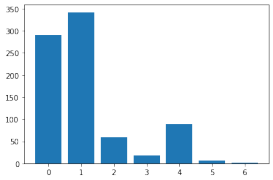
(b) $k = 4$



(c) $k = 5$

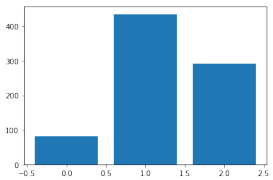


(d) $k = 6$

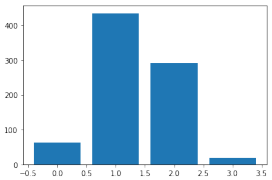


(e) $k = 7$

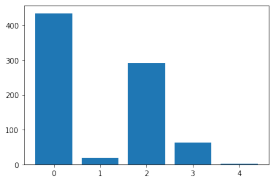
Figura 11: Número de instancias en cada cluster(K-Means), tipo 'enlace de salida'



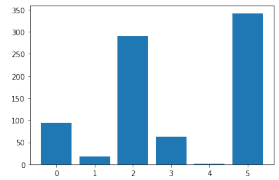
(a) $k = 3$



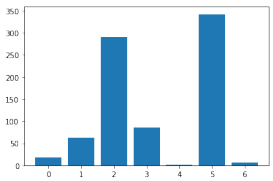
(b) $k = 4$



(c) $k = 5$



(d) $k = 6$



(e) $k = 7$

Figura 12: Número de instancias en cada cluster(Agglomerative-Clustering), tipo 'enlace de salida'

Por lo general en mi caso particular, hay muy pocas victimas por accidente y pocos heridos. Exceptuando un caso que parece ser de un accidente de un autobús por la cantidad de víctimas.

Se puede ver como uno de los clusters siempre tiene 1 vehículo implicado(posiblemente los atropellos y las colisiones con obstáculos) y otro con el valor máximo de victimas(52), con este último siempre estando solo en su cluster. También se ven los que tienen algunos vehículos implicados, que serán choques entre vehículos.

Las pocas cifras de muertos que hay, seguramente quedan en los clusters con muy pocas instancias en ellos, al igual que pasará con los heridos graves.

Como vista global, hay muchos menos heridos de los que me esperaba con mi primera hipótesis, así como obviamente mortalidad quedando agrupados los casos con un solo vehículo(atropellos) y el resto se forman dependiendo de los pocos heridos que haya o si algún vehículo ha estado implicado.

2.2.2. Tipo de intersección - Enlace de entrada

Ahora veremos el caso opuesto, las situaciones cuando se producen estos accidentes en las intersecciones de entrada.

En cuanto al tipo de vía sigue siendo la mayoría en autovía, pero en este caso también en autopista incluso superando a esta otra. También en las ramas de enlace se reduce mucho la cantidad de accidentes en comparación al caso opuesto, al igual que en vías de servicio.

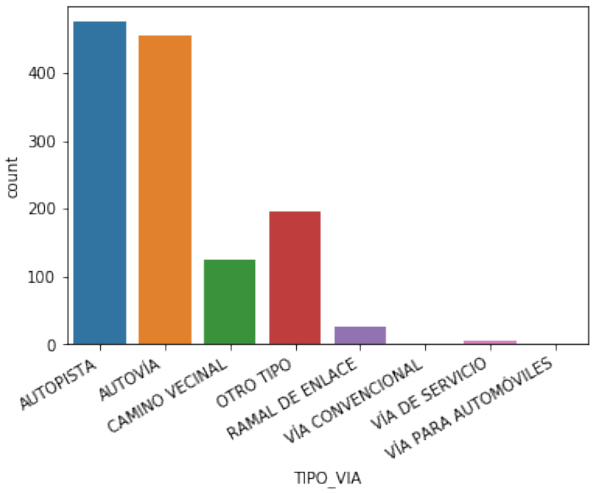
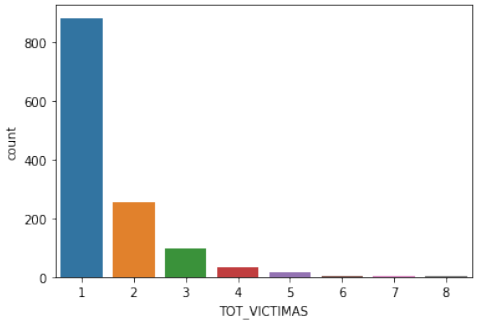


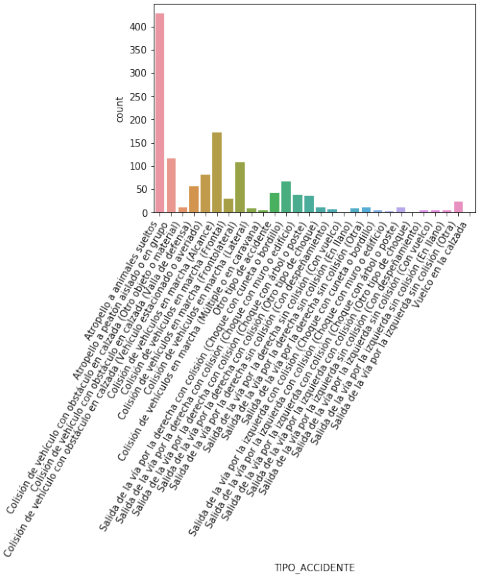
Figura 13: Tipo de vía en intersecciones de tipo 'enlace de entrada'

El número de víctimas es casi igual al que hay en los enlaces de salida.

En cuanto al tipo de accidentes que se producen, se mantiene el dominio de los atropellos a animales seguido por peatones(aunque reducido con respecto al caso opuesto). Las colisiones con obstáculos se ve reducida(en bastante medida) y la colisión con vehículos se ve aumentada.



(a) Número de victimas en intersecciones de tipo 'enlace de entrada'



(b) Tipo de accidentes en intersecciones de tipo 'enlace de entrada'

Figura 14: Víctimas y tipos de accidente en intersecciones de tipo 'enlace de entrada'

2.2.2.1 Resultados de la segmentación

Para evaluar la calidad del agrupamiento obtenido, se usarán las mismas medidas de evaluación que en el caso anterior.

$k = 3$	K-Means	Agg-Clustering
Silhouette	***0,6001***	*0,5951*
Calinski	***789,7076***	746,7932

$k = 4$	K-Means	Agg-Clustering
Silhouette	***0,5706***	*0,5217*
Calinski	***818,6911***	689,6661

$k = 5$	K-Means	Agg-Clustering
Silhouette	***0,6568***	*0,6238*
Calinski	***846,7492***	693,9017

$k = 6$	K-Means	Agg-Clustering
Silhouette	***0,6795***	*0,6470*
Calinski	***824,9188***	739,5186

$k = 7$	K-Means	Agg-Clustering
Silhouette	***0,7088***	*0,6663*
Calinski	***932,9841***	813,6500

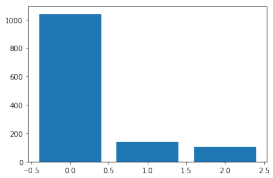
Tabla 2: Resultados obtenidos con los distintos algoritmos para las intersecciones de tipo 'enlace de entrada'

Para este caso también dan valores parecidos para Silhouette y se mantiene el mismo patrón que en el caso opuesto, K-Means queda por encima(mejor agrupamiento global) que con Agglomerative-Clustering.

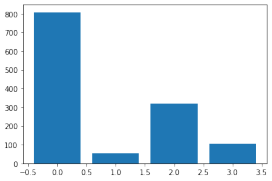
2.2.2.2 Interpretación de la segmentación

Figura 15: Mapas de calor para el algoritmo K-Means, tipo 'enlace de entrada'

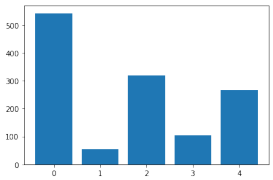
Figura 16: Mapas de calor para el algoritmo Agglomerative Clustering, tipo 'enlace de entrada'



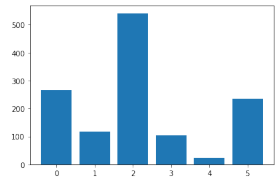
(a) $k = 3$



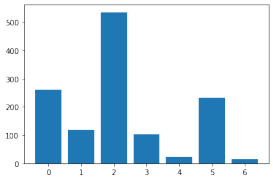
(b) $k = 4$



(c) $k = 5$

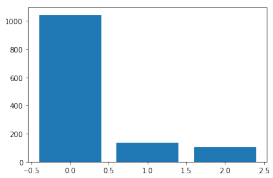


(d) $k = 6$

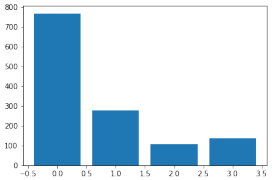


(e) $k = 7$

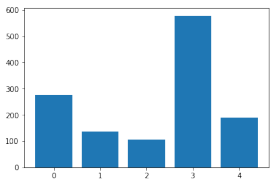
Figura 17: Número de instancias en cada cluster(K-Means), tipo 'enlace de entrada'



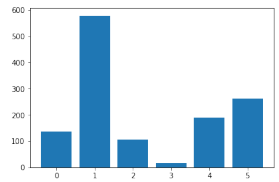
(a) $k = 3$



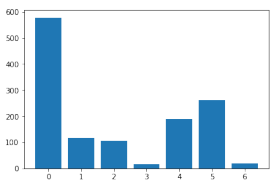
(b) $k = 4$



(c) $k = 5$



(d) $k = 6$



(e) $k = 7$

Figura 18: Número de instancias en cada cluster(Agglomerative-Clustering), tipo 'enlace de entrada'

En este caso parece apreciarse como ya aparecen más víctimas y heridos que en el caso anterior(seguramente sea por el aumento de colisiones con vehículos).

Vuelve a aparecer un cluster cuando aumenta el valor de k con un solo vehículo implicado de media(atropellos), y termina agrupandose a los pocos accidentes con casos mortales en un cluster en su mayoría.

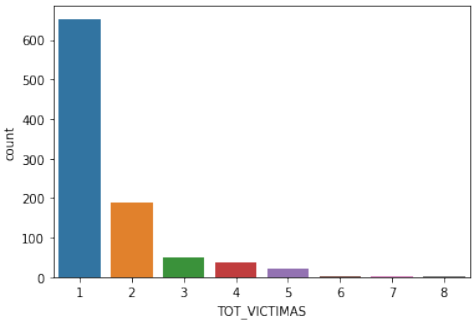
Los clusters parecen mantenerse conforme aumenta la k formándose otros nuevos para especificar más.

Como conclusión, aparece un aumento de heridos(tanto leves como graves) y de víctimas totales.

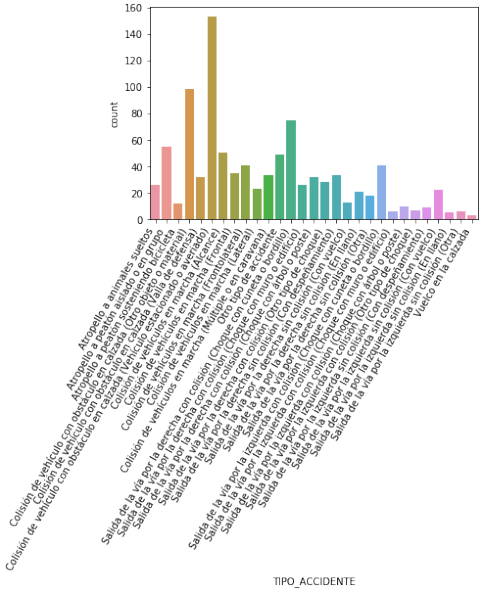
2.3. Caso de estudio 2

Para este segundo caso de estudio, se tendrán en cuenta los accidentes comprendidos desde las 12 horas a las 6 horas(a.m para un caso y p.m para el opuesto) en Andalucía, para ver como difieren las cifras de víctimas comparando por el día y por la madrugada.

2.3.1. Primera franja horaria: 12-6 a.m



(a) Número de victimas de 12 a 6 a.m



(b) Tipo de accidentes en intersecciones de 12 a 6 a.m

Figura 19: Víctimas y tipos de accidente en intersecciones de 12 a 6 a.m

2.3.1.1 Resultados de la segmentación

k = 3	**K-Means**	**Agg-Clustering**
Silhouette	***0,5659***	*0,5519*
Calinsky	***573,8087***	479,1132

k = 4	**K-Means**	**Agg-Clustering**
Silhouette	***0,5813***	*0,5736*
Calinsky	***582,3789***	495,6096

k = 5	**K-Means**	**Agg-Clustering**
Silhouette	***0,6309***	*0,5629*
Calinsky	***659,8113***	555,6438

k = 6	**K-Means**	**Agg-Clustering**
Silhouette	*0,7001*	***0,7026***
Calinsky	***685,6643***	657,3857

k = 7	**K-Means**	**Agg-Clustering**
Silhouette	*0,6734*	***0,7088***
Calinsky	*629,0390*	***643,1446***

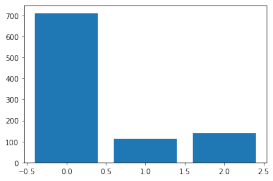
Tabla 3: Resultados obtenidos con los distintos algoritmos para el tramo de la madrugada(12-6 a.m)

Ocurre algo parecido en parte al primer caso de estudio, el valor de Silhouette es parecido para ambos algoritmos excepto en algunos casos($k = 5$ y $k = 7$). Con el índice Calinski podemos ver que para $k = 6$ usando K-Means obtiene el mejor resultado entre los valores de k escogidos, por lo que podríamos decir que es el mejor agrupado. Aun empatando con Agg para $k = 7$ tiene mejor Calinski.

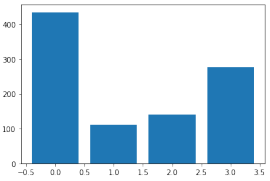
2.3.1.2 Interpretación de la segmentación

Figura 20: Mapas de calor para el algoritmo K-Means, A.M

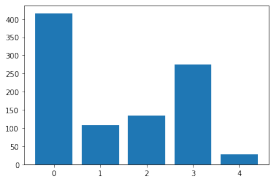
Figura 21: Mapas de calor para el algoritmo Agglomerative Clustering, A.M



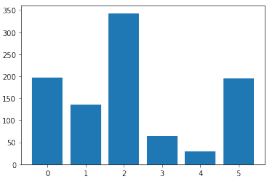
(a) $k = 3$



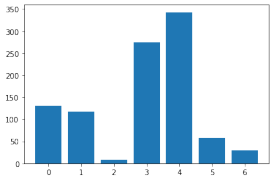
(b) $k = 4$



(c) $k = 5$

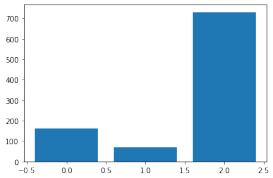


(d) $k = 6$

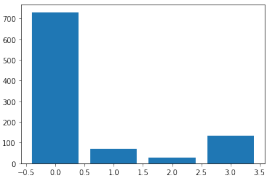


(e) $k = 7$

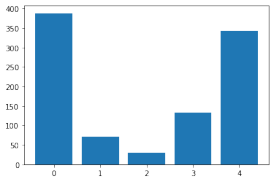
Figura 22: Número de instancias en cada cluster(K-Means), A.M



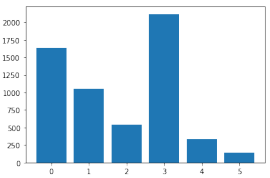
(a) $k = 3$



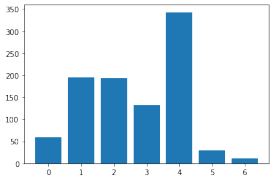
(b) $k = 4$



(c) $k = 5$



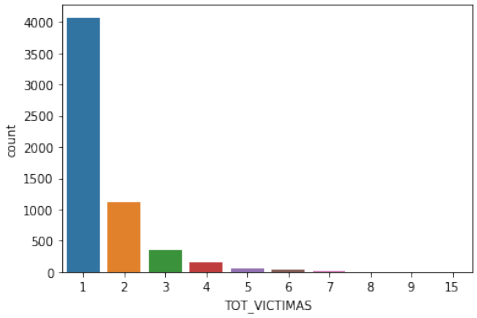
(d) $k = 6$



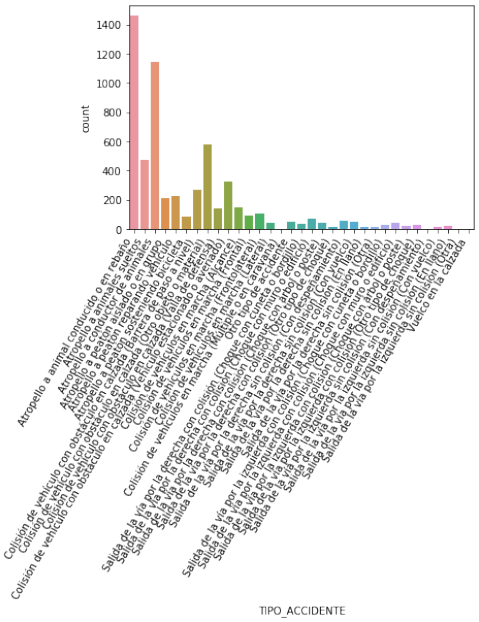
(e) $k = 7$

Figura 23: Número de instancias en cada cluster(Agglomerative-Clustering), A.M

2.3.2. Primera franja horaria: 12-6 p.m



(a) Número de victimas de 12 a 6 p.m



(b) Tipo de accidentes en intersecciones de 12 a 6 p.m

Figura 24: Víctimas y tipos de accidente en intersecciones de 12 a 6 p.m

Durante el día lo que predomina es el atropello ed animales conducidos o en rebaño y conductores de animales, seguido por animales sueltos y colisiones.

2.3.2.1 Resultados de la segmentación

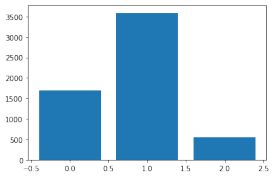
$k = 3$	K-Means	Agg-Clustering	$k = 4$	K-Means	Agg-Clustering
Silhouette	*0,5815*	***0,5818***	Silhouette	***0,6701***	*0,6312*
Calinsky	***3.586,8308***	*3.533,2931*	Calinsky	***4.519,7336***	*3.758,7649*
$k = 5$	K-Means	Agg-Clustering	$k = 6$	K-Means	Agg-Clustering
Silhouette	***0,6824***	*0,6677*	Silhouette	***0,7527***	*0,7485*
Calinsky	***4.237,4522***	*3.785,2146*	Calinsky	***4.354,3507***	*4.127,1006*
$k = 7$	K-Means	Agg-Clustering			
Silhouette	*0,7533*	***0,7538***			
Calinsky	***4.211,6426***	*4.077,9610*			

Tabla 4: Resultados obtenidos con los distintos algoritmos para el tramo del día(12-6 p.m)

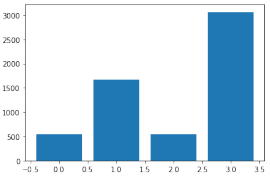
2.3.2.2 Interpretación de la segmentación

Figura 25: Mapas de calor para el algoritmo K-Means, P.M

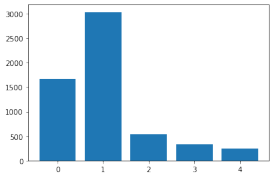
Figura 26: Mapas de calor para el algoritmo Agglomerative Clustering, P.M



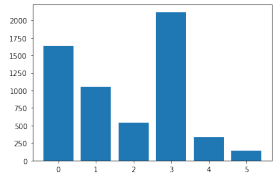
(a) $k = 3$



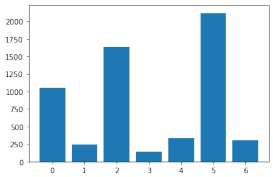
(b) $k = 4$



(c) $k = 5$

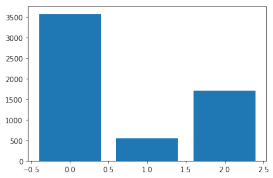


(d) $k = 6$

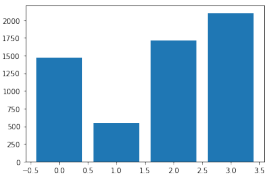


(e) $k = 7$

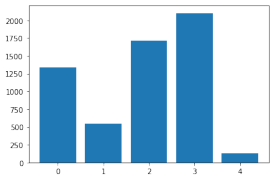
Figura 27: Número de instancias en cada cluster(K-Means), P.M



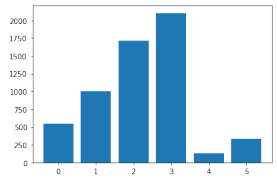
(a) $k = 3$



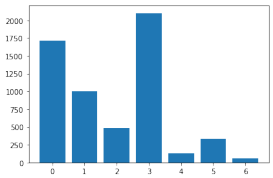
(b) $k = 4$



(c) $k = 5$



(d) $k = 6$



(e) $k = 7$

Figura 28: Número de instancias en cada cluster(Agglomerative-Clustering), P.M

En los agrupamientos se puede ver como separa bien, en unos teniendo apenas heridos y un vehículo implicado, serán atropellos. Y como dependiendo del número de heridos leves y vehiculos implicados va segmentando las instancias.

En comparación con el tramo de la madrugada, aquí aparecen cifras de muertos

mayor que cero en más de un cluster, cuando en el caso opuesto se agrupan todas en uno(parece ser colisión con obstáculo o salida de vía). Además de mayores cifras en heridos leves y victimas en distintos clusters con respecto a la madrugada.