Práctica 2

Clustering

08/12/2020
Inteligencia de negocio

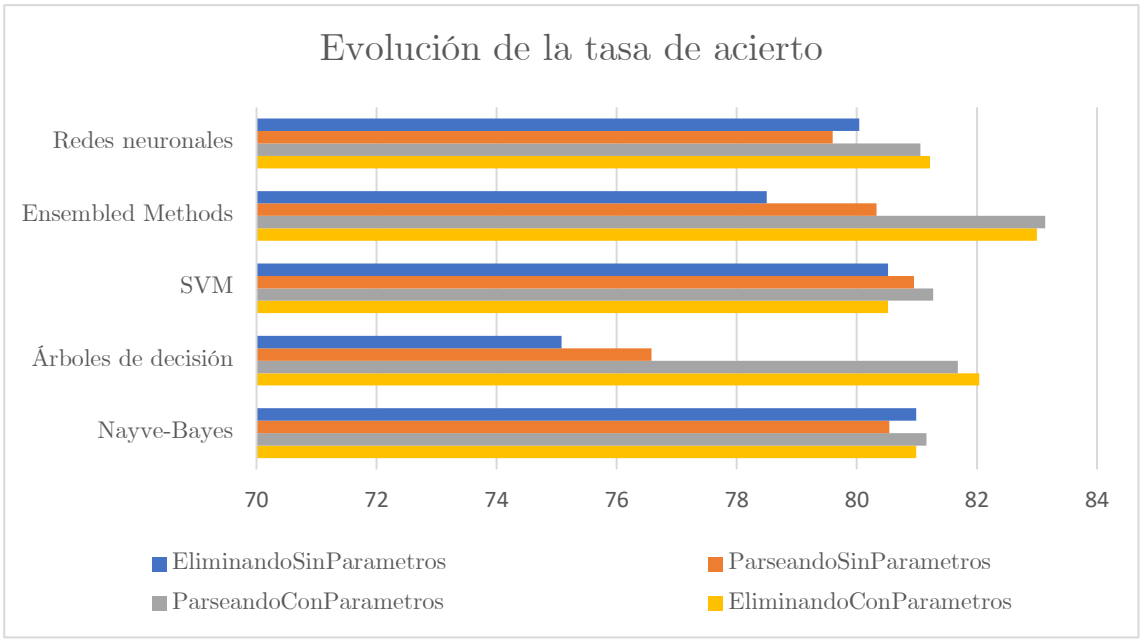
Juan Carlos González Quesada
UGR

Contenido

Parte 1

Visualización de las medidas

Representaremos un gráfico donde mostramos la evolución de cada algoritmo con sus diferentes procesamientos y dando también una visión global.

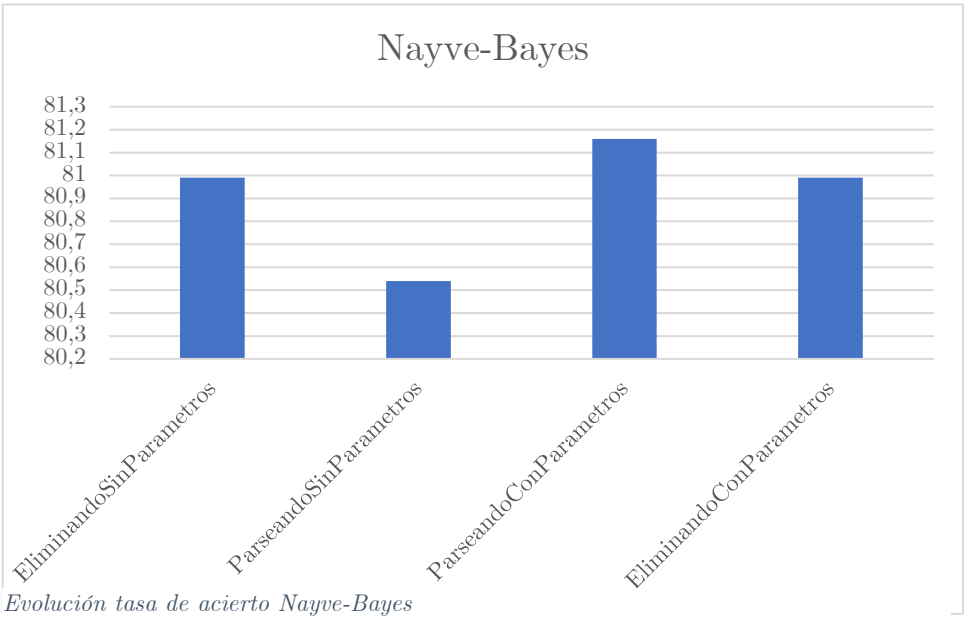


Evolución de la tasa de acierto de los algoritmos

Vamos a ver la evolución de forma individual.

Nayve-Bayes

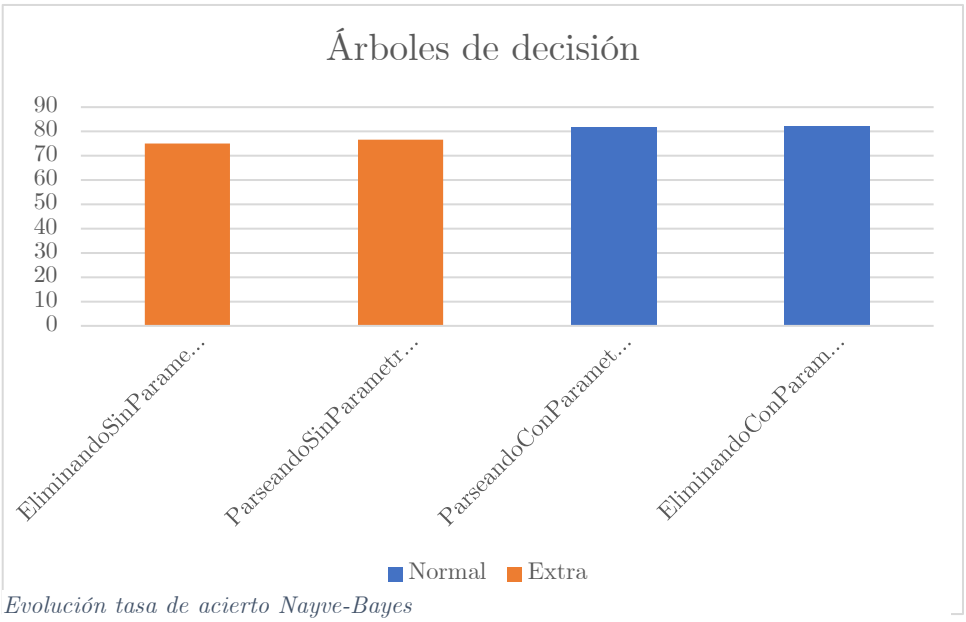
Para el algoritmo Nayve-Bayes el subtipo que mejor porcentaje nos ha dado ha sido siempre el Gaussiano.



Evolución tasa de acierto Nayve-Bayes

Árboles de decisión

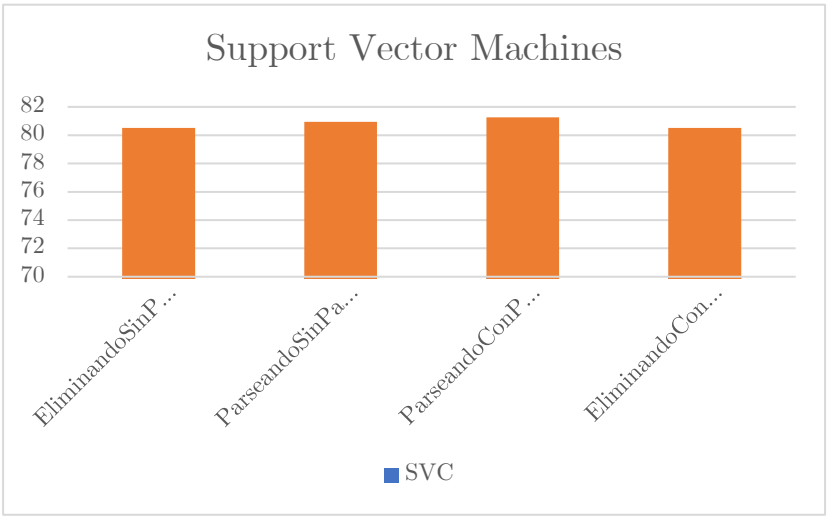
En las dos primeras ejecuciones el subtipo "extra" de los árboles de decisión, nos da mejores resultados que el "normal". Esto queda reflejado de una mejor forma en este gráfico.



Evolución tasa de acierto Nayve-Bayes

Support Vector Machines

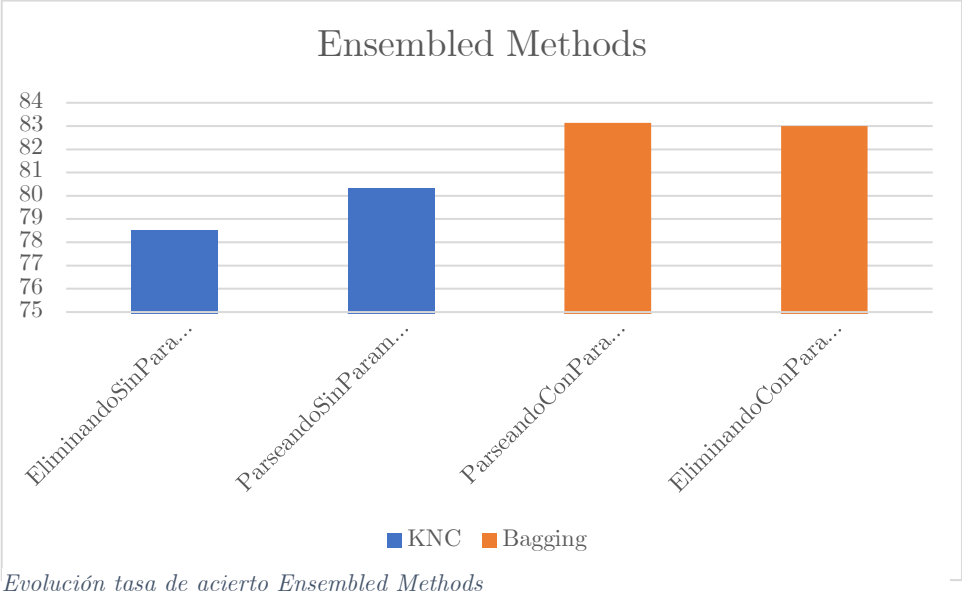
En este algoritmo sólo hay un tipo que nos da un acierto mayor.



Evolución tasa de acierto SVM

Ensembled Methods

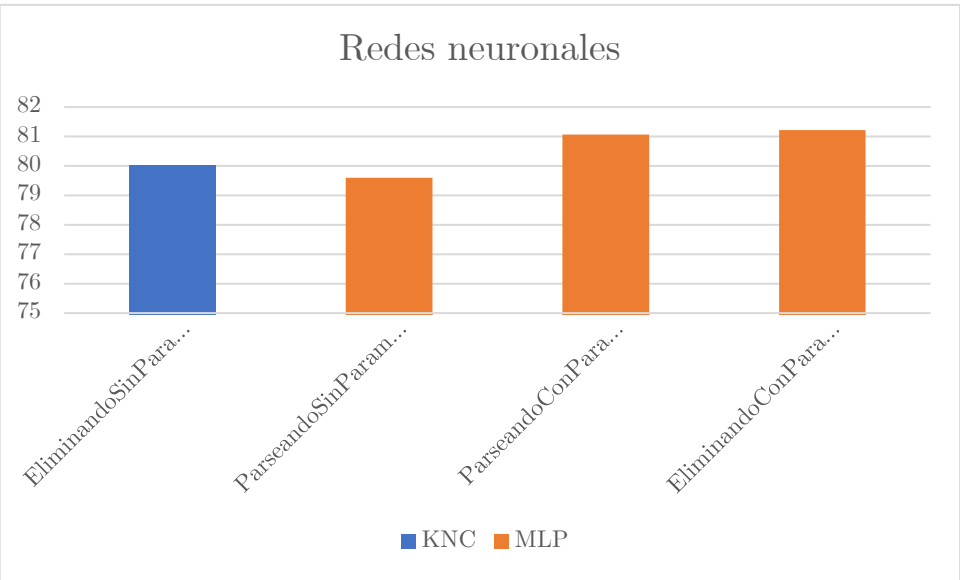
Este algoritmo presenta la tasa más alta estudiada, aunque al principio el subtipo "bagging" no es mejor que el "random forest". Finalmente, obteniendo un 83,13 % será el mejor para abordar el problema de la clasificación.



Evolución tasa de acierto Ensembled Methods

Redes neuronales

En el algoritmo de redes neuronales, el que mejores resultados nos dio fue MLP, sin embargo, vemos una curiosidad y es que, en la primera mejora, obtenemos una tasa de acierto peor que con la primera ejecución y con distinto subtipo de algoritmo.



Evolución tasa de acierto Redes neuronales

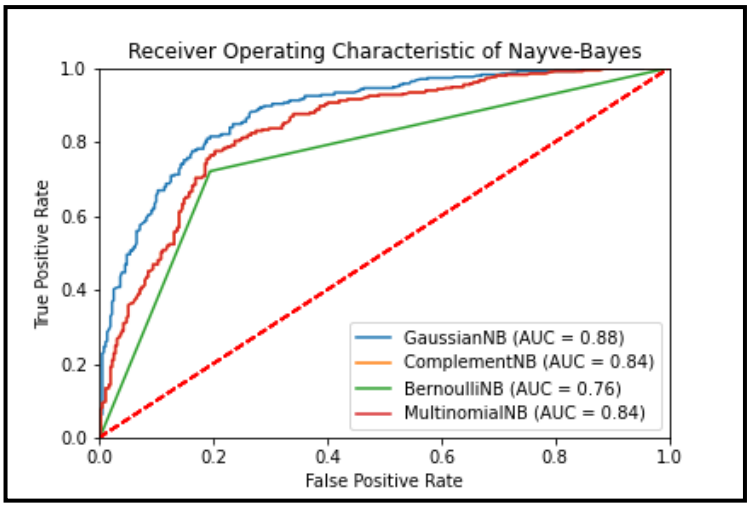
Visualización de la curva ROC

El estudio lo realizaremos basándonos en los cuatro casos que realizamos en la práctica anterior.

Algoritmo Nayve-Bayes

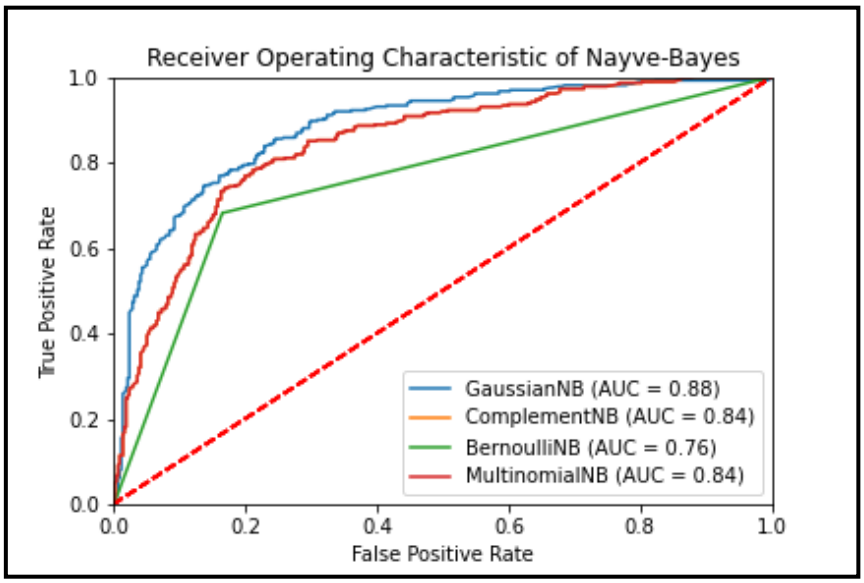
En este algoritmo vamos a ver como en todos sus subtipos, el algoritmo "Gaussiano" es el que obtiene la mejor curva y valor AUC.

Eliminando datos y sin parámetros



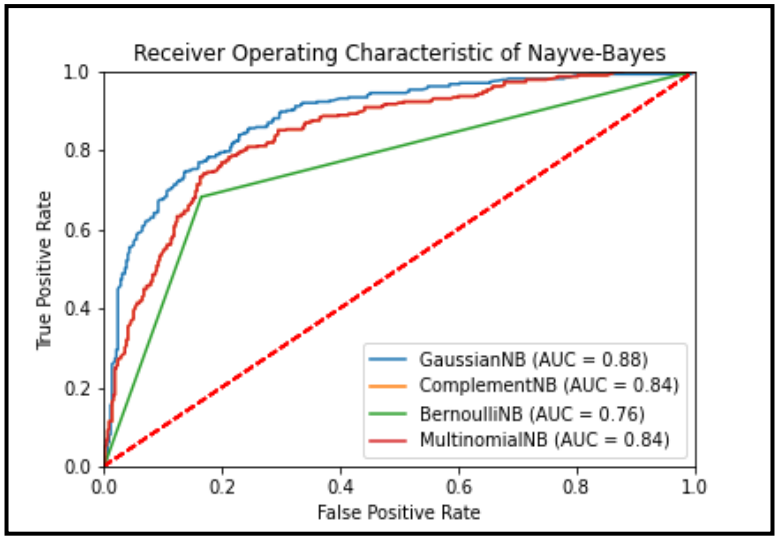
Representación de la curva ROC para el algoritmo Nayve-Bayes

Parseando datos y sin parámetros



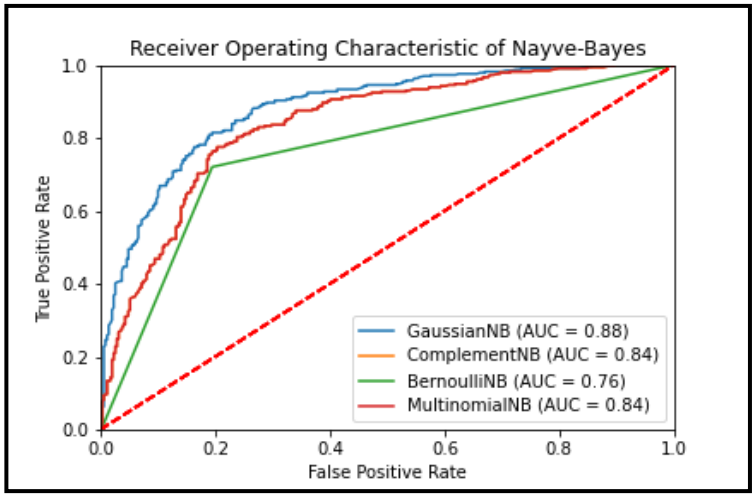
Representación de la curva ROC para el algoritmo Nayve-Bayes

Parseando datos y con parámetros



Representación de la curva ROC para el algoritmo Nayve-Bayes

Eliminando datos y con parámetros



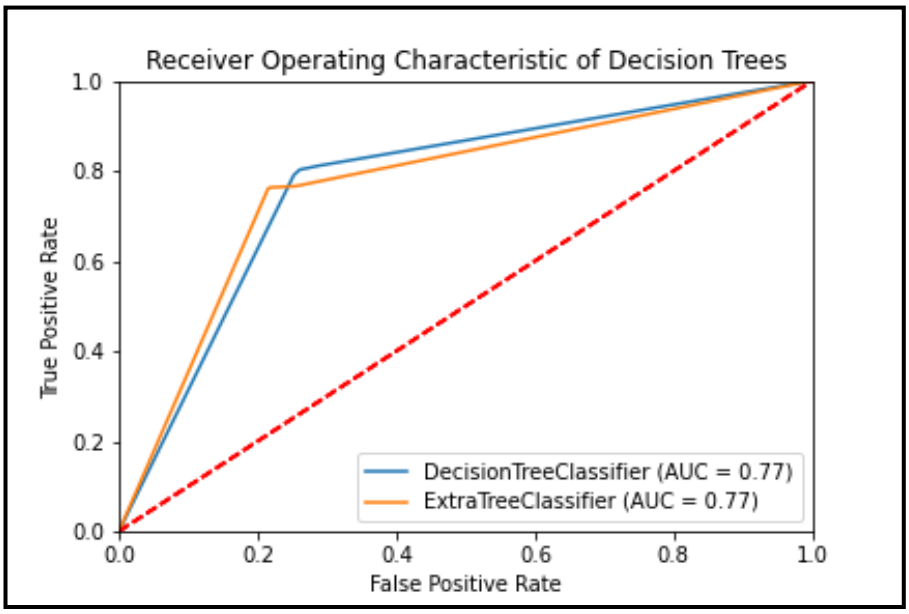
Representación de la curva ROC para el algoritmo Nayve-Bayes

Para el algoritmo de clasificación Nayve-Bayes usamos cuatro tipos. En la "*Práctica 1*" concluimos que el mejor de los tipos era el "Gaussian". Al esbozar la gráfica ROC observamos que también queda demostrado de esta forma. Para este estudio no se añadirán parámetros, puesto que no los admite.

Árboles de decisión

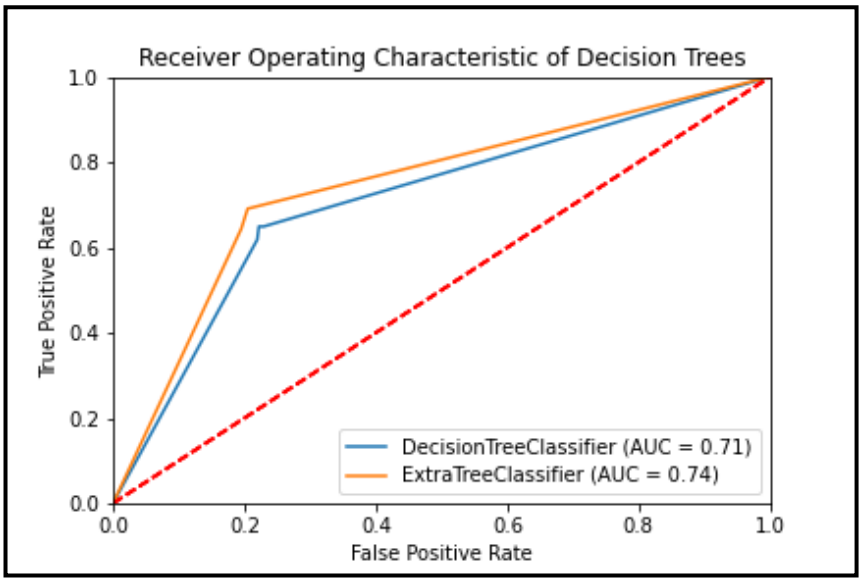
Para este algoritmo, vemos como para las dos primeras ejecuciones era mejor el algoritmo "Extra" y finalmente, acaba obteniendo mejores resultados el "normal".

Eliminando datos y sin parámetros



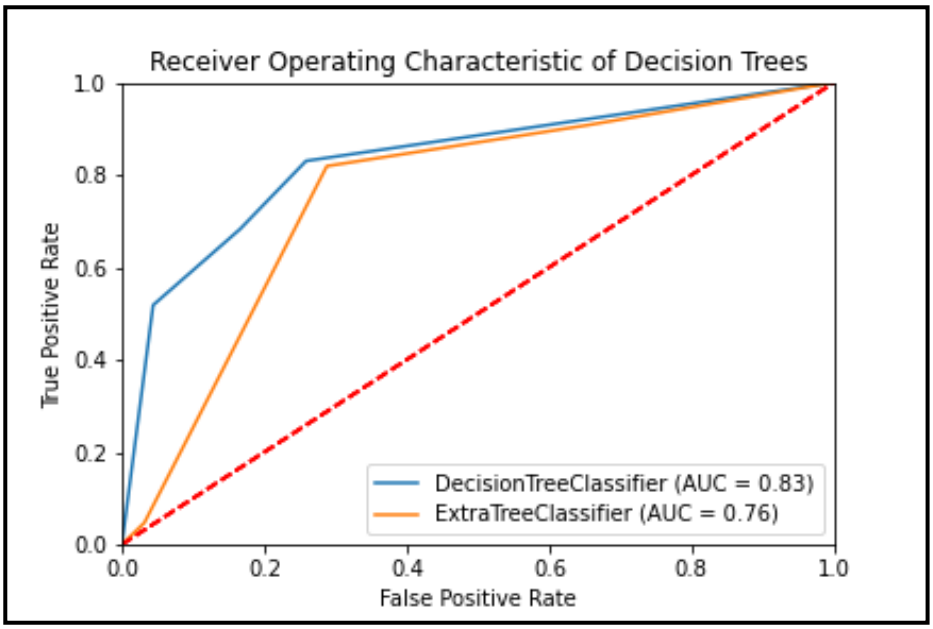
Representación de la curva ROC para el algoritmo Árboles de decisión

Parseando datos y sin parámetros



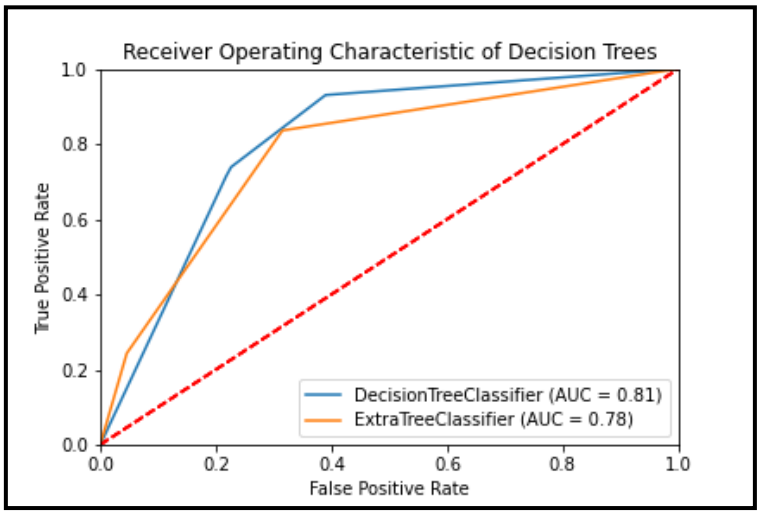
Representación de la curva ROC para el algoritmo Árboles de decisión

Parseando datos y con parámetros



Representación de la curva ROC para el algoritmo Árboles de decisión

Eliminando datos y con parámetros



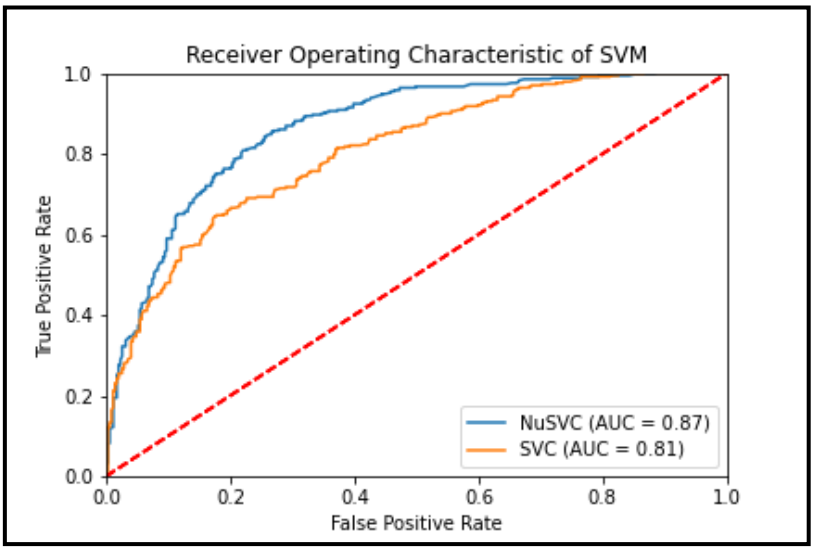
Representación de la curva ROC para el algoritmo Árboles de decisión

En este algoritmo, que ha tenido varios subtipos que nos han dado buenos resultados, vemos como el AUC no siempre beneficia al que mejor tasa de acierto tiene.

Support Vector Machines

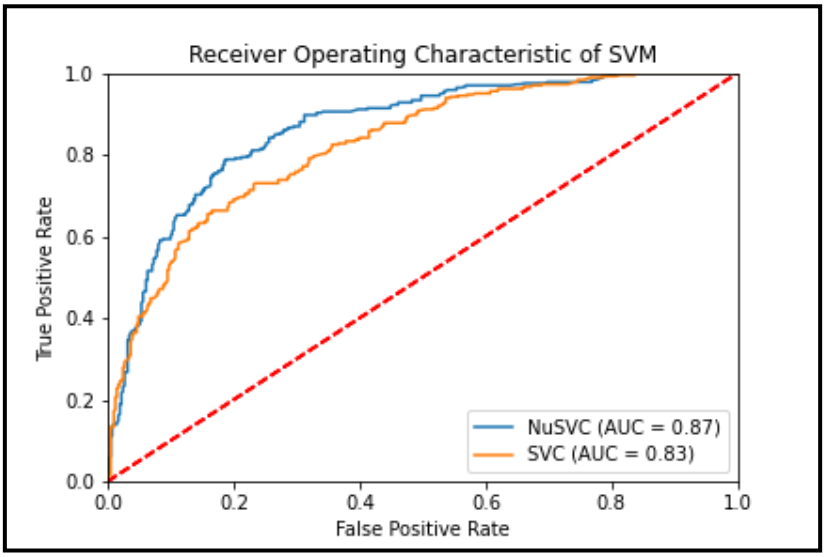
Este algoritmo tiene sólo un subtipo que nos da la mejor tasa de acierto.

Eliminando datos y sin parámetros



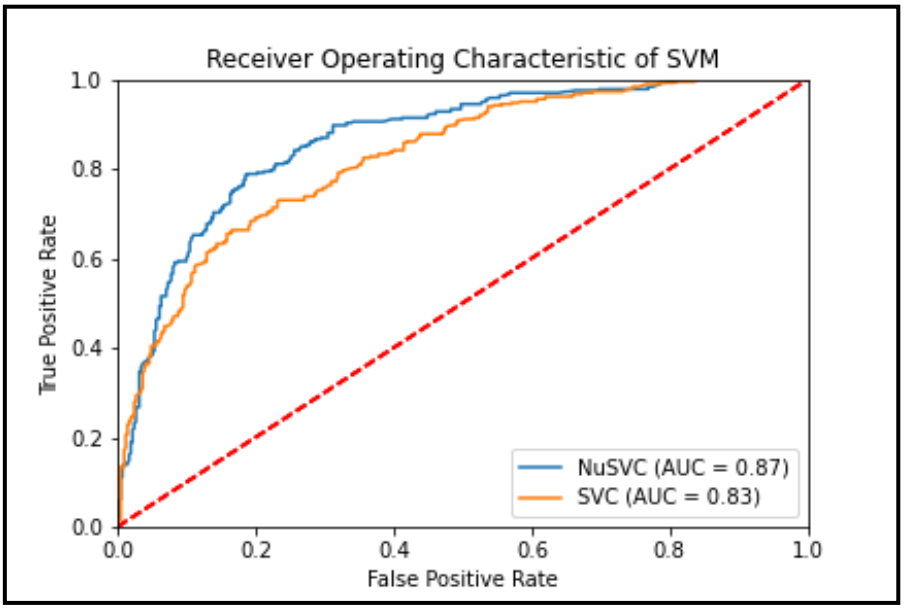
Representación de la curva ROC para el algoritmo SVM

Parseando datos y sin parámetros



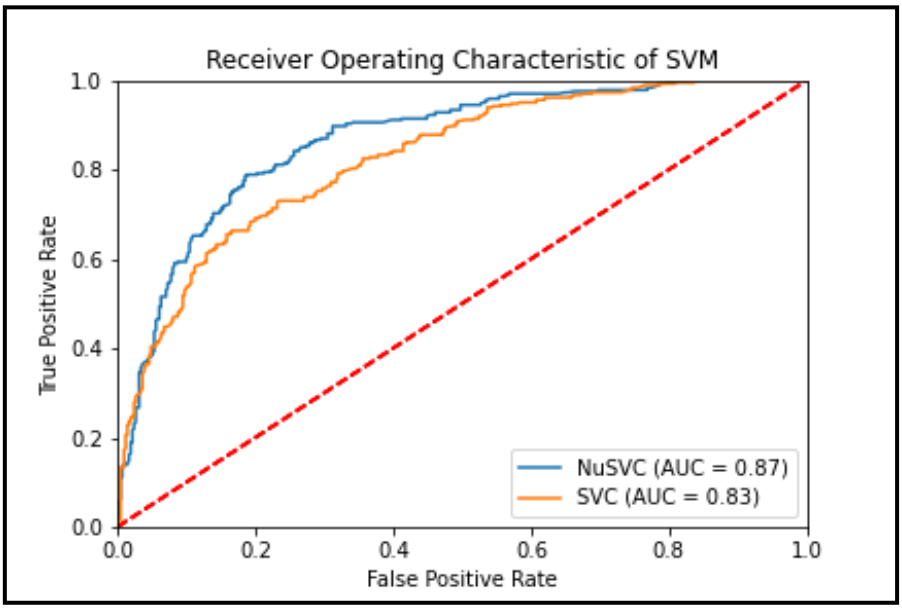
Representación de la curva ROC para el algoritmo SVM

Parseando datos y con parámetros



Representación de la curva ROC para el algoritmo SVM

Eliminando datos y con parámetros

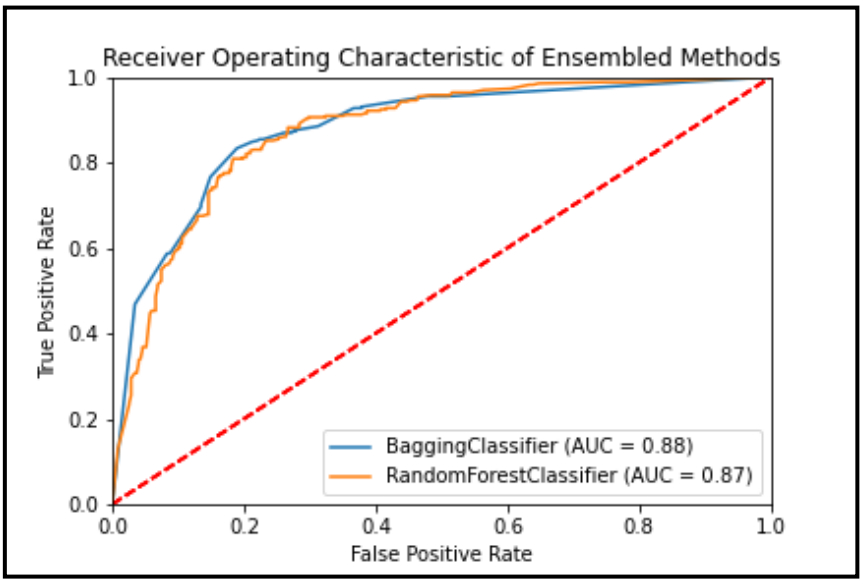


Representación de la curva ROC para el algoritmo SVM

Ensembled Methods

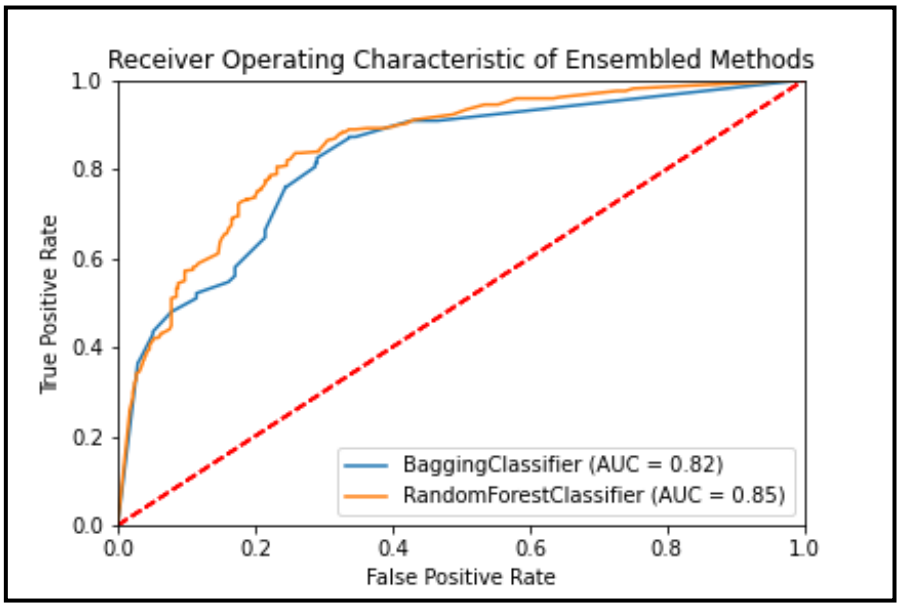
Este algoritmo tiene dos subtipos que nos dan buenos resultados. Observamos que Bagging consigue un buen AUC y obtiene la mayor tasa para el problema planteado en la "Práctica 1".

Eliminando datos y sin parámetros



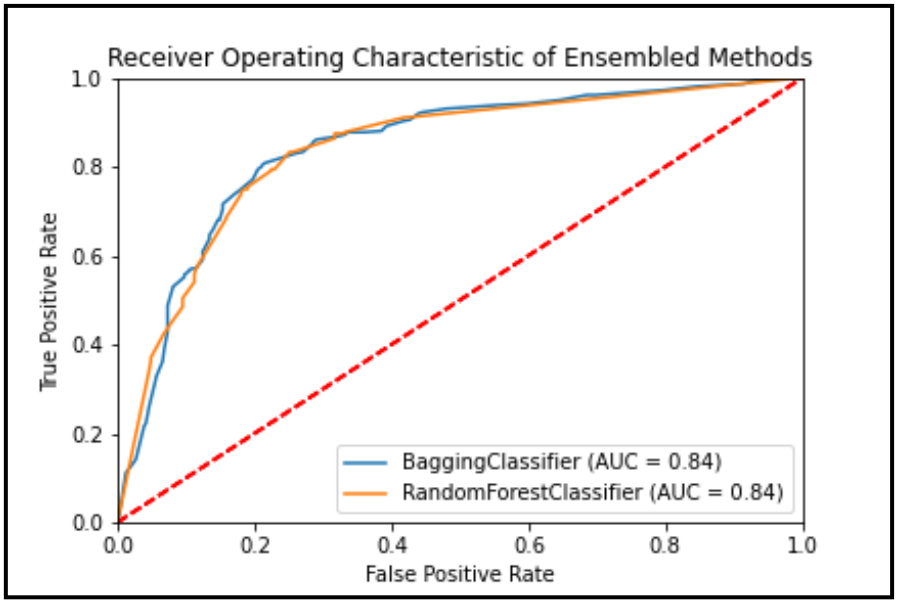
Representación de la curva ROC para el algoritmo Ensembled Methods

Parseando datos y sin parámetros



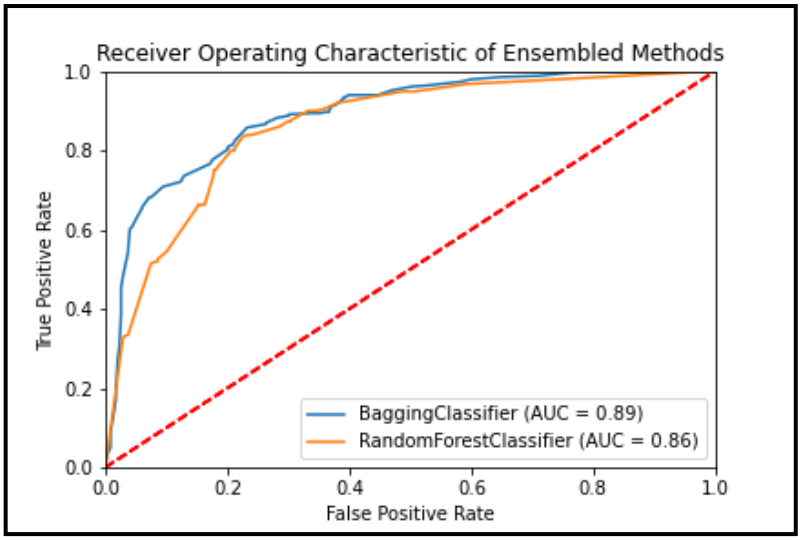
Representación de la curva ROC para el algoritmo Ensembled Methods

Parseando datos y con parámetros



Representación de la curva ROC para el algoritmo Ensembled Methods

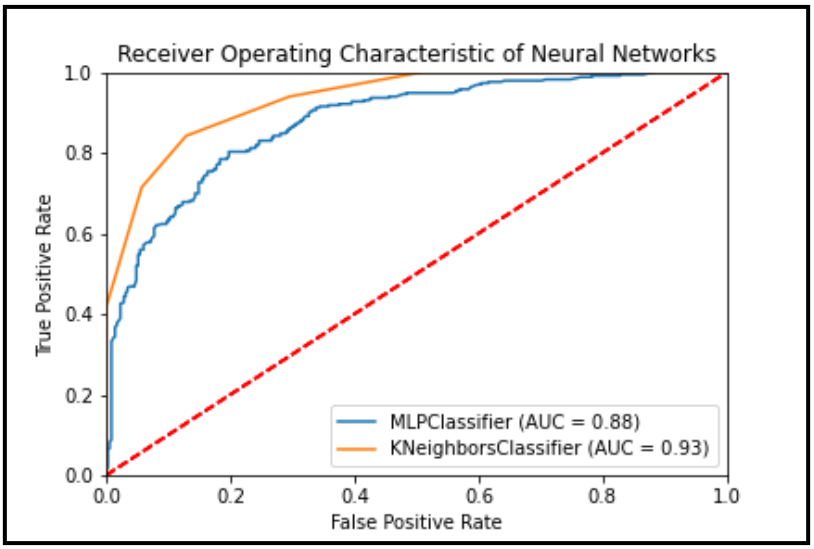
Eliminando datos y con parámetros



Representación de la curva ROC para el algoritmo Ensembled Methods

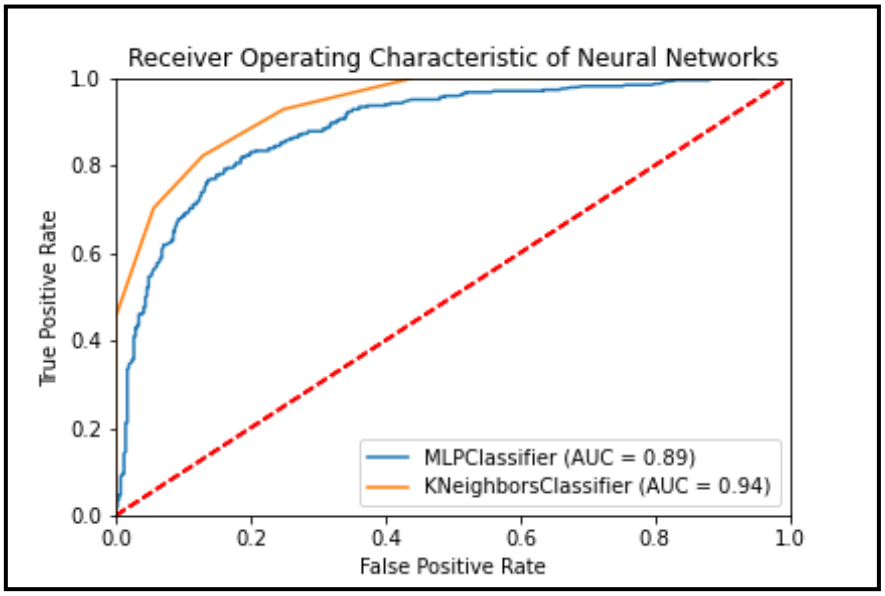
Redes Neuronales

Eliminando datos y sin parámetros



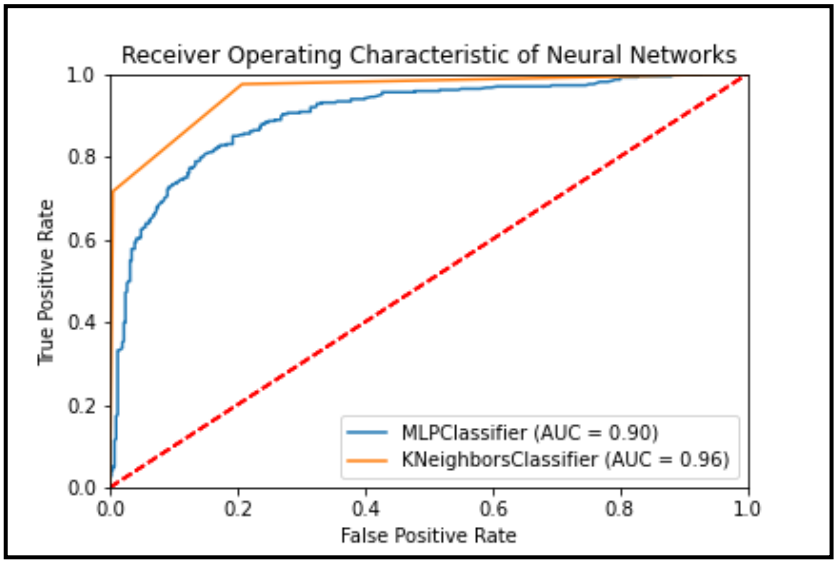
Representación de la curva ROC para el algoritmo Redes Neuronales

Parseando datos y sin parámetros



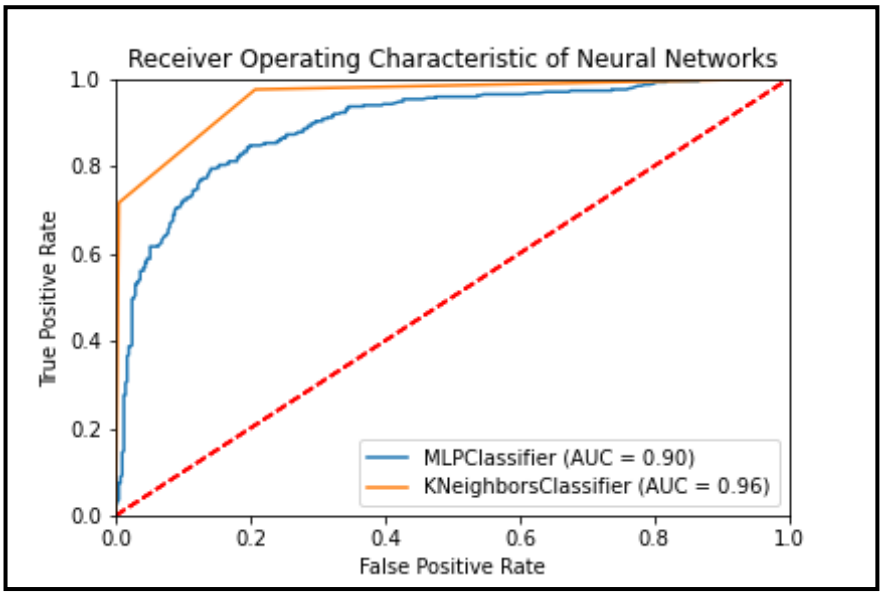
Representación de la curva ROC para el algoritmo Redes Neuronales

Parseando datos y con parámetros



Representación de la curva ROC para el algoritmo Redes Neuronales

Eliminando datos y con parámetros



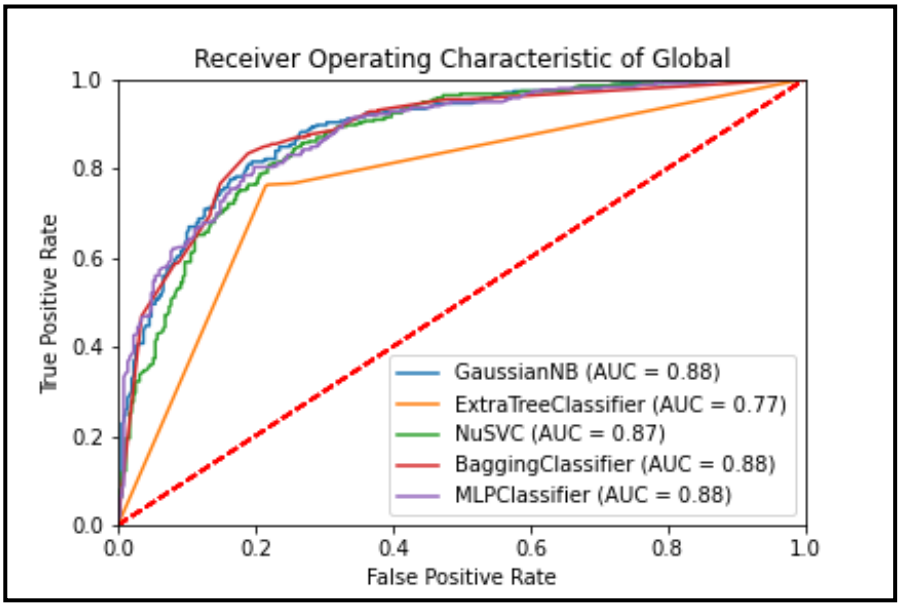
Representación de la curva ROC para el algoritmo Redes Neuronales

Este algoritmo nos da como resultado un AUC medio prácticamente superior a 90 (0.9 sobre 1). Sin embargo, no es el que mejor tasa de acierto obtiene en ningún caso.

Resumen de mejores subtipos

Eliminando datos y sin parámetros

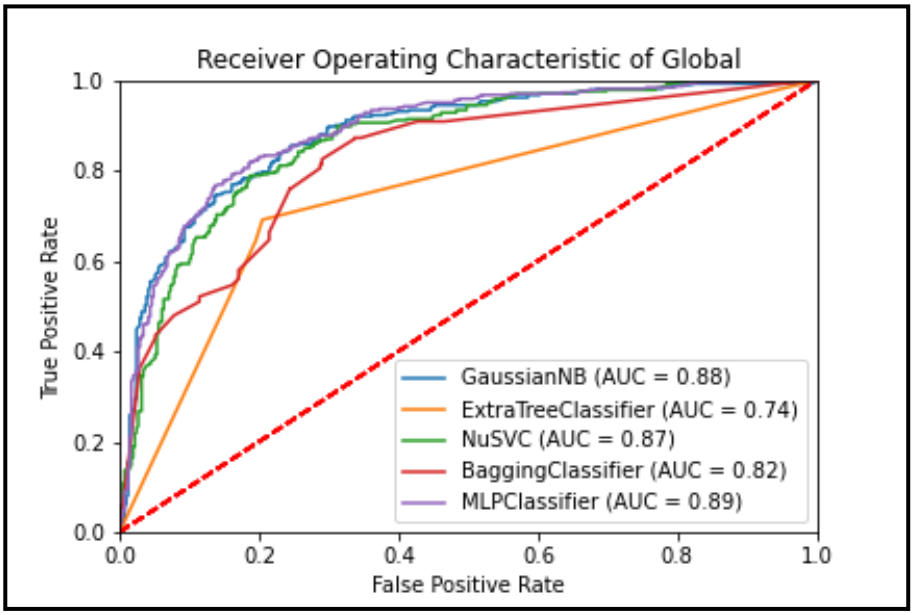
Para este caso, la mejor tasa nos la daba el algoritmo Nayve-Bayes que, como observamos en la gráfica, también obtiene un buen AUC.



Representación de la curva ROC para la representación global de las mejores tasas de acierto

Parseando datos y sin parámetros

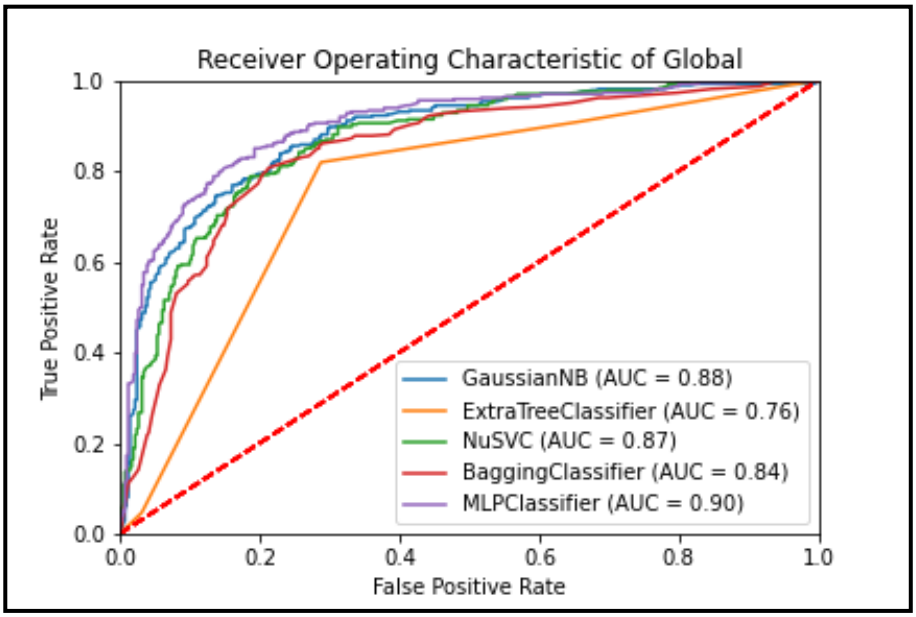
A continuación, se muestra una gráfica en la que vemos cómo el algoritmo SVM no nos da la mejor tasa ya que no posee el mejor de los valores AUC. En este caso sería el algoritmo de redes neuronales el que destaca por tener mejor valor AUC.



Representación de la curva ROC para la representación global de las mejores tasas de acierto

Parseando datos y con parámetros

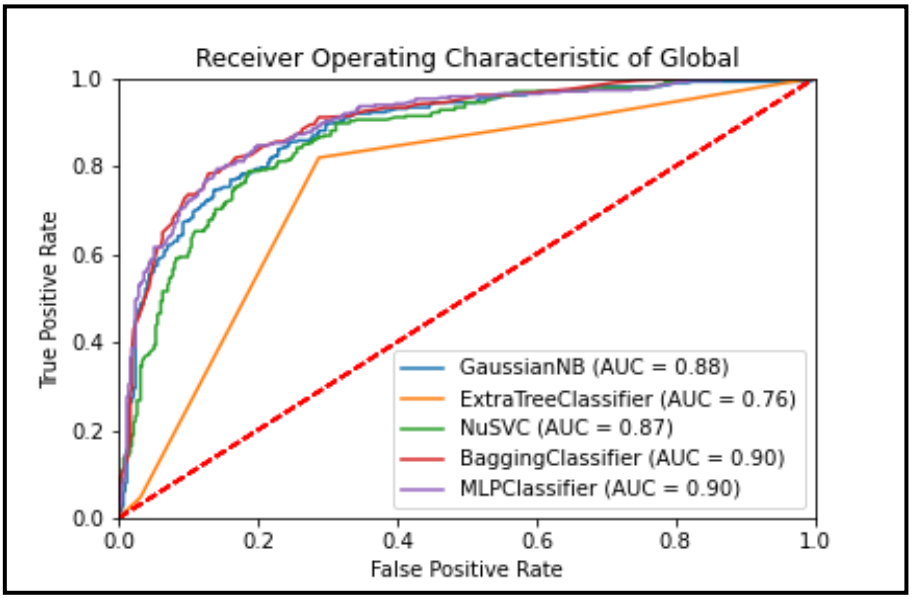
Para este caso, la mejor tasa nos la daba el algoritmo Ensembled Methods. Sin embargo, no obtenemos un buen AUC respecto al resto de algoritmos. Esto contrasta con su tasa de acierto, un 83,13 %, siendo la más alta obtenida.



Representación de la curva ROC para la representación global de las mejores tasas de acierto

Eliminando datos y con parámetros

Para este caso, la mejor tasa nos la daba el algoritmo SVM que, como observamos en la gráfica, también obtiene, junto a redes neuronales, el mejor AUC observado hasta el momento.



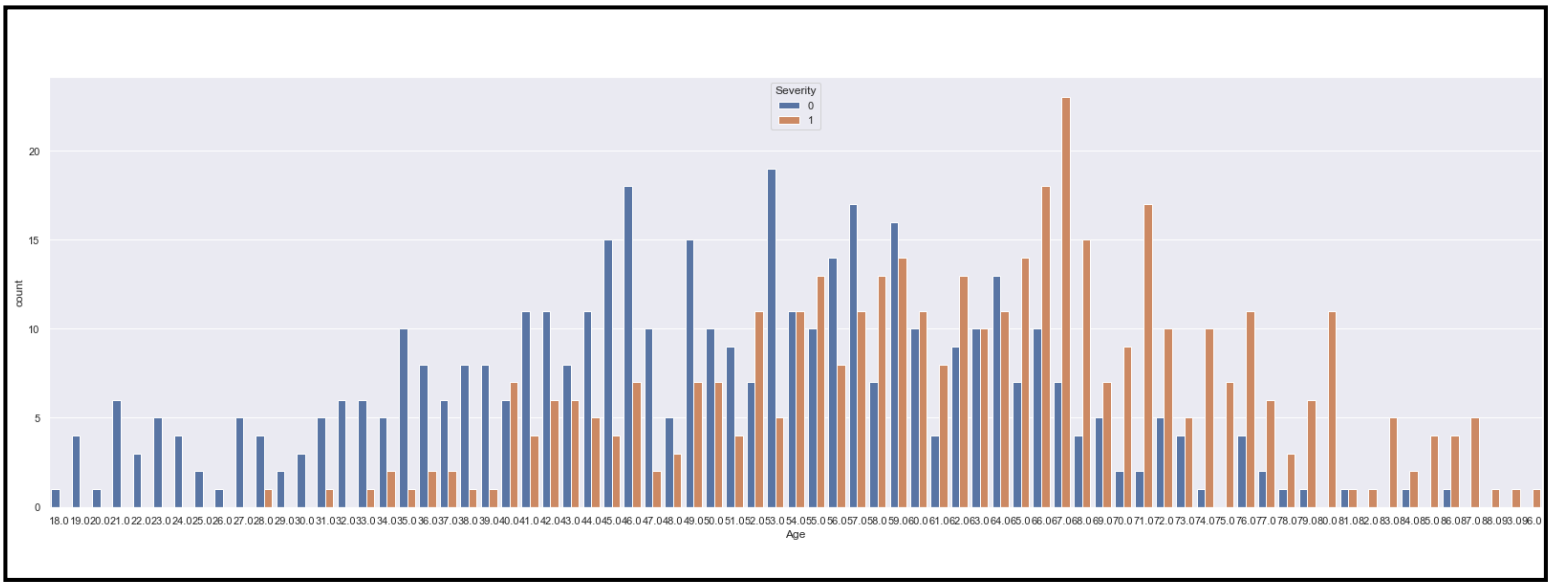
Representación de la curva ROC para la representación global de las mejores tasas de acierto

Análisis de los atributos

Primer estudio

En numerosos estudios[1] se ha hablado que a ciertas edades se aumenta la probabilidad de tener cáncer de mama. La edad de la que se habla es entorno a los 50 años.

En nuestro primer estudio veremos si guardan relación la edad con los casos positivos de cáncer de mama. Se va a utilizar el conjunto de valores con los datos perdidos eliminados, para evitar una modificación de los datos y un estudio más certero. ¿Será la edad un factor importante a la hora de padecer cáncer de mama?



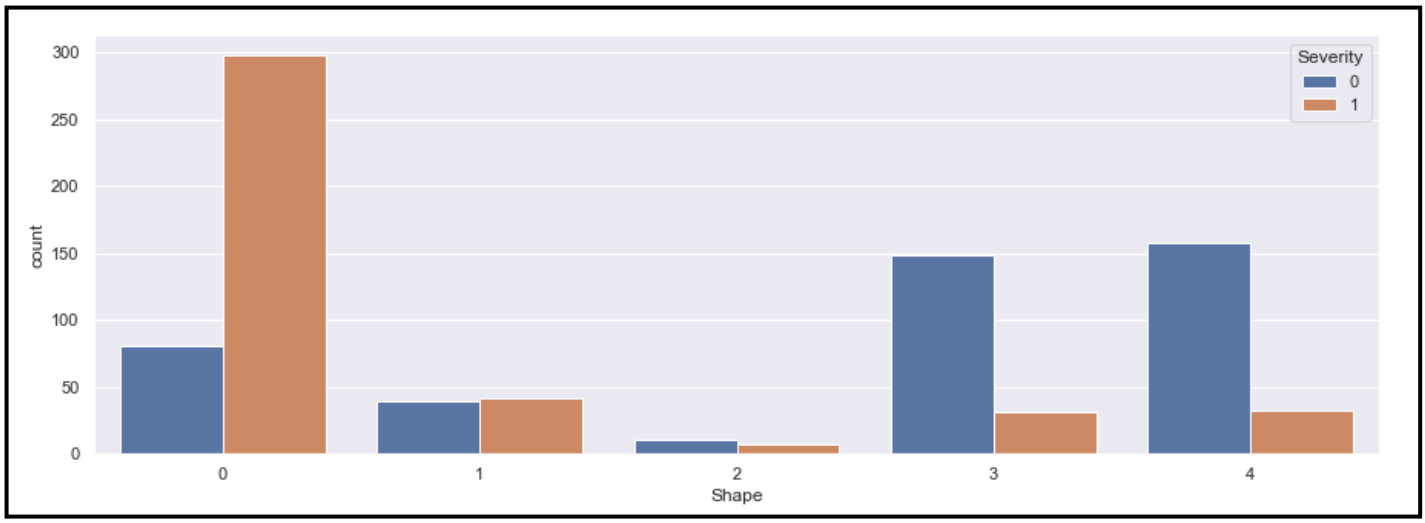
Gráfica que relaciona la edad con la severidad

Observamos como a medida que se va aumentando la edad, los positivos que se han detectado de cáncer aumentan considerablemente. Nos damos cuenta que, aunque hay positivos desde los 28 años, cuanto más volumen de positivos encontramos es en el rango de edades de 48 a 81 años. También nos encontramos con unos pocos casos hasta los 96 años. Por tanto, podemos concluir que la edad es un factor importante pero no determinante para la aparición del cáncer.

[1] https://www.cancer.net/es/tipos-de-c%C3%A1ncer/cancer-de-mama/estad%C3%ADsticas

Segundo estudio

Otros estudios[2] nos indican, que la forma en la que se encuentre el conjunto de células en la mama influye de manera muy notable en la predicción de su severidad. ¿Será un factor determinante?



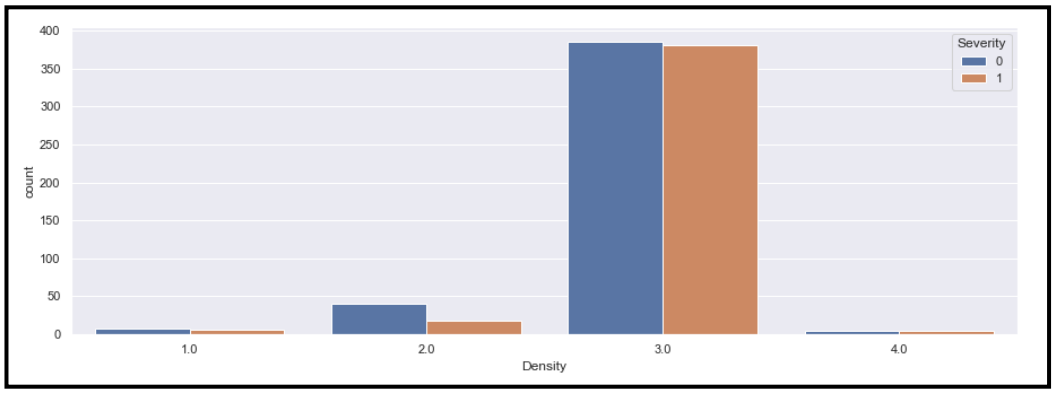
Gráfica que relaciona la forma tumoral con la severidad

Significado de los números: "*0 -> Redondeada, 1 -> Ovalada, 2 -> Lobulada, 3 -> Irregular, 4 -> No definida*".

Observando estos gráficos, vemos que hay un claro factor que determina la alta posibilidad de padecer un cáncer de mama y es que tenga una forma redondeada. Esto nos lleva a nuestro último caso de estudio. Suponemos que la forma redondeada es porque hay una alta concentración de materia tumoral. ¿Es también un factor de alta probabilidad su alta densidad?

Tercer estudio

Ahora comparamos la severidad de los resultados con la densidad de la masa tumoral.



Gráfica que relaciona la densidad tumoral con la severidad

Significado de los números: "*1 -> Alta, 2 -> Media, 3 -> Baja, 4 -> No definida*".

[2] https://www.cancer.net/es/tipos-de-cancer/cancer-de-mama/diagnostico

Tras visualizar el gráfico, nos damos cuenta de que el informe realizado es erróneo, puesto que es la densidad "baja" es la que proporciona el máximo número tanto de positivos como de negativos. Esto nos hace plantearnos que quizás no exista una relación entre la forma y la densidad del tumor con la severidad. Así que lo comprobaremos de la siguiente forma:

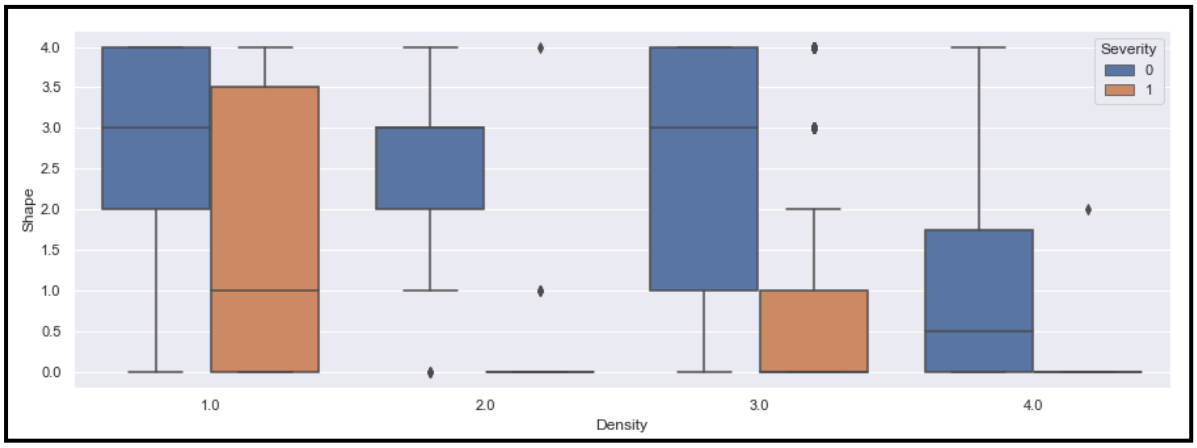


Diagrama de caja que representa la relación entre densidad y forma con su severidad

Tras este diagrama de caja, podemos decir que hay una pequeña relación entre los positivos con alta densidad y una forma redondeada. Lo observamos en la mediana de la severidad positiva y que su lado más grande es el izquierdo y posee una asimetría negativa, en el otro caso que estudiamos, que es el de densidad baja, tiene una alta relación con la forma lobulada, que lo observamos en su mediana y su asimetría positiva.

Parte 2

Para el desarrollo de esta práctica usaré los algoritmos Kmeans y Birch.

Enfoque general

¿Sobre qué se puede realizar un estudio de accidentes de coche? Quizás se vengan muchas ideas a la cabeza: Lluvia, exceso de velocidad, alcohol, distracciones al volante, malos descansos, dolores de cabeza... Estos son mis dos casos de uso:

Caso 1

Cuando realizas el examen teórico del coche o ves algún informe [3]de la DGT[4], puedes leer que las carreteras convencionales son más peligrosas que las autovías. Esto deriva a que en las carreteras convencionales se produzcan un gran número de accidentes respecto a las autovías. Para completar el informe, visualizaré también si hay relación con los factores climatológicos. Mi comparación será entre las clases BUEN TIEMPO y LLUVIA, donde consideraré LLUVIA FUERTE y LLOVIZNANDO. Esto se debe a que la lluvia fuerte implica que los coches puedan resbalar debido a la abundante agua y la llovizna implica que los accidentes sean peligrosos con las primeras gotas ya que se mezclan con sustancias que se encuentran en la carretera. Lo mismo pasa con el TIPO_VÍA, que usaré la convencional y la convencional con carril lento.

Justificación del tipo de vía

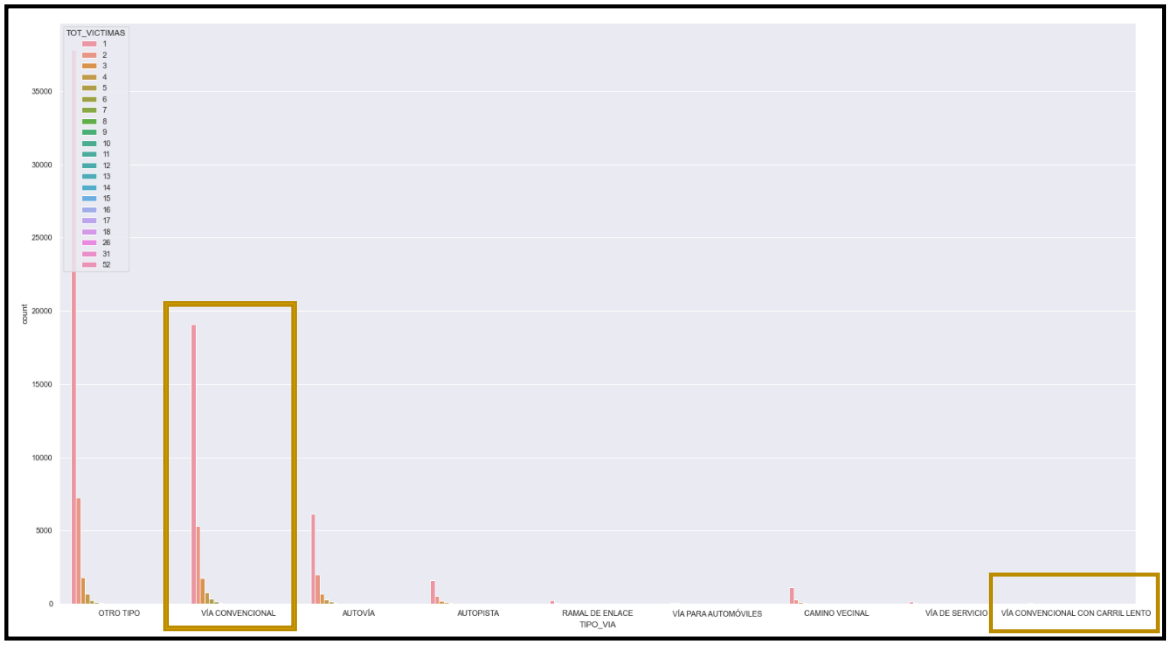


Gráfico ilustrativo que relaciona el tipo de vía con el número de víctimas

Sin contar con las vías que no se han clasificado, vemos como la incidencia de accidentes con víctimas se agrupan dentro de los dos tipos y, a simple vista, se observa como las vías convencionales tienen mayor tasa de accidentes.

---

[3] https://www.coches.net/noticias/carreteras-convencionales-poco-seguras

[4] http://www.dgt.es/es/prensa/notas-de-prensa/2016/20160509-dos-cada-tres-fallecidos-accidente-trafico-producen-carreteras-convencionales.shtml

Justificación del factor atmosférico

Aunque parezca que se pueden sufrir más accidentes cuando el clima es desfavorable para los conductores, podemos observar en la gráfica como llegamos a una contradicción. La mayoría de los accidentes ocurren con un buen clima y día.

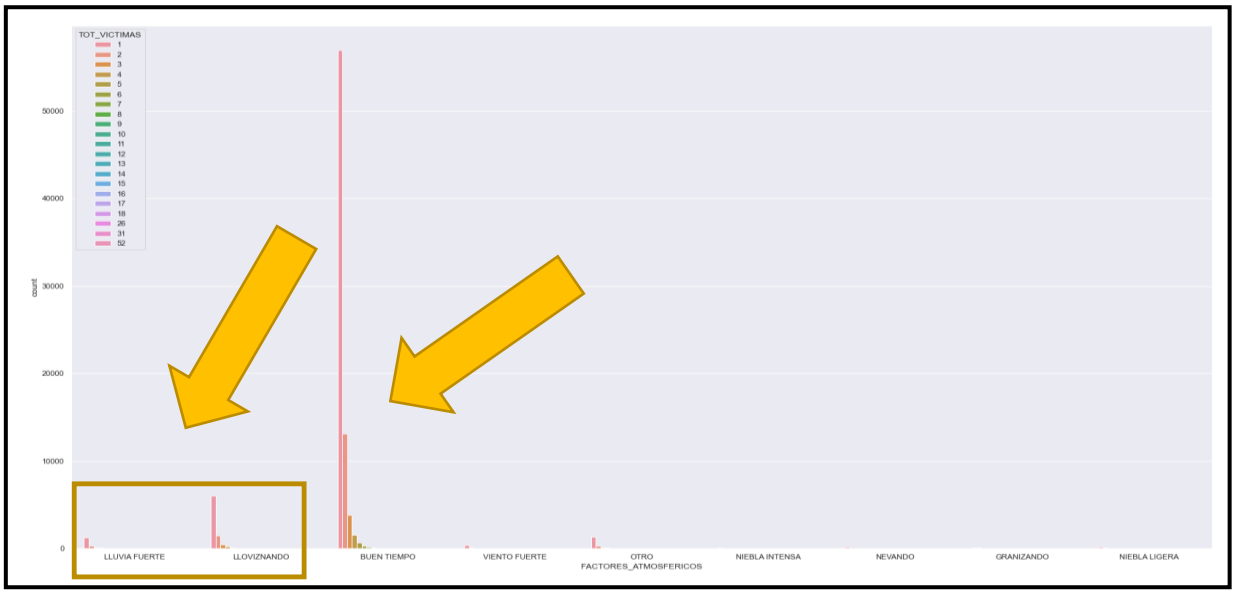


Gráfico que nos relaciona los factores atmosféricos con el número de víctimas

Trabajando con algoritmos

Para realizar las comparativas y obtener una buena conclusión, vamos a comparar los clústeres (de las mejores mediciones) BUEN TIEMPO con los de LLUVIA. De esta manera, podremos verificar si para distintas condiciones, pero igual vía, obtenemos el mismo tipo de "gráfico de centroides"

Interpretación de la segmentación

¿Qué valores de K escoger?

Para ello vamos a utilizar el método "*Elbow*", que es un método con el que podemos observar de una forma visual cuales son los mejores valores para los clústeres. Se obtiene midiendo el valor WCSS ("*Within-Cluster-Sum-of-Squares*"), el cual disminuye al aumentar el número de clústeres, de forma que los mejores valores son cuando se forma un "codo".

Para nuestro estudio escogeremos 4, 6, 10 y 16. Ya que es donde se observa un mayor número de codos.

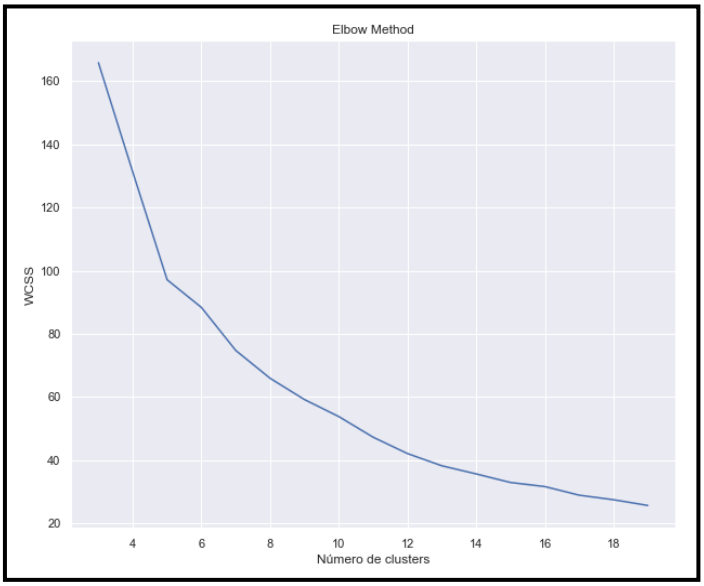


Gráfico que ilustra el valor WCSS

Relación y tamaño de los clústeres

El tamaño para el DF Soleado es de 23.087 y para Lluvia 3.454.

	Comunidad	Buen tiempo	Lloviendo
	Nº Clusters	Tamaño de cada clúster	
Kmeans	4	2: 9762, 1: 9655, 0: 1862, 3: 1808	0: 1744, 2: 1175, 3: 309, 1: 226
	6	0: 8569, 1: 8051, 5: 2829, 3: 1842, 4: 1383, 2: 413	5: 1404, 2: 996, 1: 335, 3: 300, 4: 215, 0: 204
	10	1: 8051, 2: 6408, 8: 2166, 7: 1577, 0: 1413, 5: 1313, 3: 1162, 4: 432, 6: 395, 9: 170	2: 1404, 1: 675, 3: 428, 5: 336, 0: 165, 8: 160, 9: 109, 7: 99, 6: 48, 4: 30
	16	1: 6721, 2: 6254, 8: 2166, 7: 1537, 12: 1105, 3: 1101, 0: 1036, 5: 746, 14: 435, 15: 411, 10: 377, 11: 322, 4: 293, 6: 256, 9: 169, 13: 158	1: 1377, 0: 660, 12: 320, 10: 251, 5: 164, 4: 159, 9: 108, 6: 90, 7: 89, 2: 82, 14: 44, 13: 42, 8: 31, 3: 19, 11: 17, 15: 1

	Comunidad	Buen tiempo	Lloviendo
	Nº Clusters	Tamaño de cada clúster (Balanceo)[5]	
Birch	4	0: 19524, 2: 1794, 1: 1768, 3: 1	1: 2165, 2: 1199, 3: 78, 0: 12
	6	2: 19475, 1: 1788, 0: 1768, 4: 49, 5: 6, 3: 1	5: 2159, 2: 1199, 0: 78, 1: 10, 3: 6, 4: 2
	10	0: 19426, 4: 1775, 2: 1764, 1: 49, 9: 25, 3: 24, 8: 13, 5: 6, 6: 4, 7: 1	5: 2159, 1: 1180, 0: 39, 2: 39, 8: 19, 6: 9, 3: 5, 4: 2, 9: 1, 7: 1
	16	10: 19133, 1: 1716, 14: 1681, 0: 293, 4: 94, 11: 48, 2: 48, 9: 25, 7: 18, 8: 13, 5: 6, 3: 5, 6: 4, 15: 1, 13: 1, 12: 1	3: 1915, 13: 1119, 12: 244, 5: 61, 6: 32, 8: 20, 1: 19, 4: 17, 2: 9, 14: 6, 0: 5, 7: 2, 10: 2, 11: 1, 9: 1, 15: 1

Comparativa de mediciones

Silhouette

Comunidad	Buen tiempo				Lluvia			
Nº Clusters	4	6	10	16	4	6	10	16
Kmeans	0.584	0.674	0.768	0.875	0.643	0.738	0.811	0.895
Birch	0.479	0.479	0.468	0.444	0.521	0.522	0.519	0.601

Calinsky

Comunidad	Buen Tiempo				Lluvia			
Nº Clusters	4	6	10	16	4	6	10	16
Kmeans	12941.98	13770.96	14208.48	15572.65	2296.6	2323.96	2565.96	2725.16
Birch	5255.97	3505.58	2119.94	1631.46	910.45	600.33	444.76	569.97

[5] A esto nos referiremos cuando hablemos del balanceo de clústeres.

El primer análisis lo haré respecto a las medidas y tamaños obtenidos. Queda patente que el algoritmo con mejores resultados es el Kmeans, y quizás sea por su ventaja de no converger a un mínimo global, sino a uno local.

Sin embargo, al realizar esta comparativa entre las dos clases, vemos como el tamaño para SOLEADO es mucho mayor que para LLUVIA, sin embargo, los mejores resultados y agrupamientos se obtienen con la segunda clase. Esto lo veremos de manera gráfica en el siguiente punto, pero de forma numérica visualizamos un mejor reparto en los clústeres de la segunda clase que para la primera, y esto sucede en los dos algoritmos.

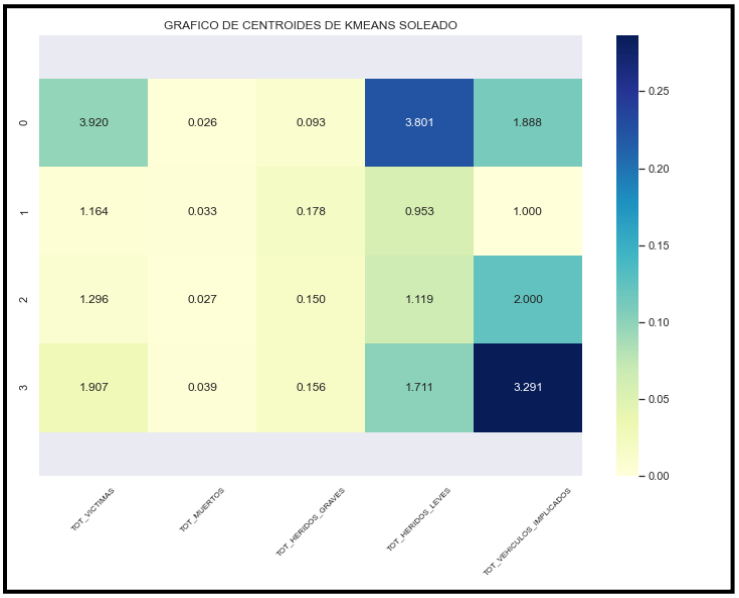
La mejor medida la obtenemos para el caso de los 16 clústeres: esto es algo normal, debido a que para el algoritmo Kmeans, cuanto más clúster tenga mayor será la medida Silhuoete. Nos tenemos que quedar con la medida que en relación con su Calinsky nos dé el mayor porcentaje: Será para el número de clúster 6.

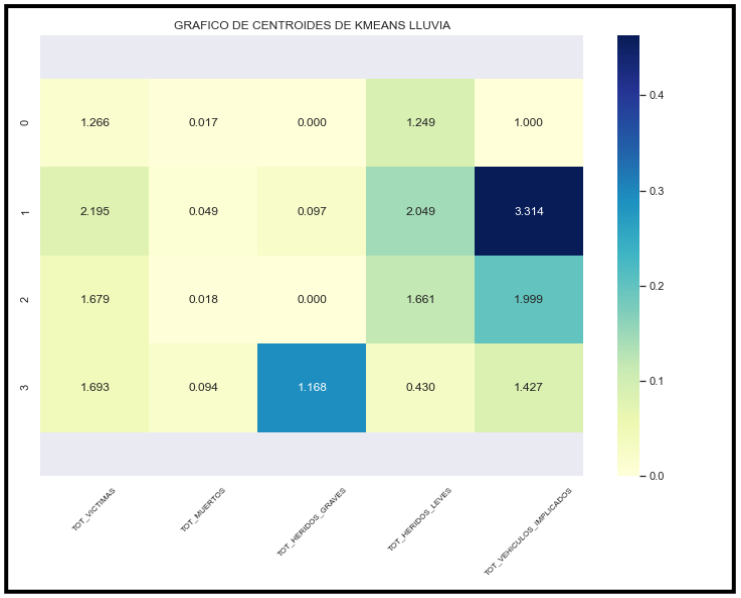
Gráficas de centroides

K = 4

Kmeans

Para un número de clústeres pequeños, tenemos el problema de que queremos incluir en solo 4 clústeres muchos datos con una cantidad elevada de variables. Al representar el gráfico de los centroides, vemos como no encontramos similitudes para las dos clases.

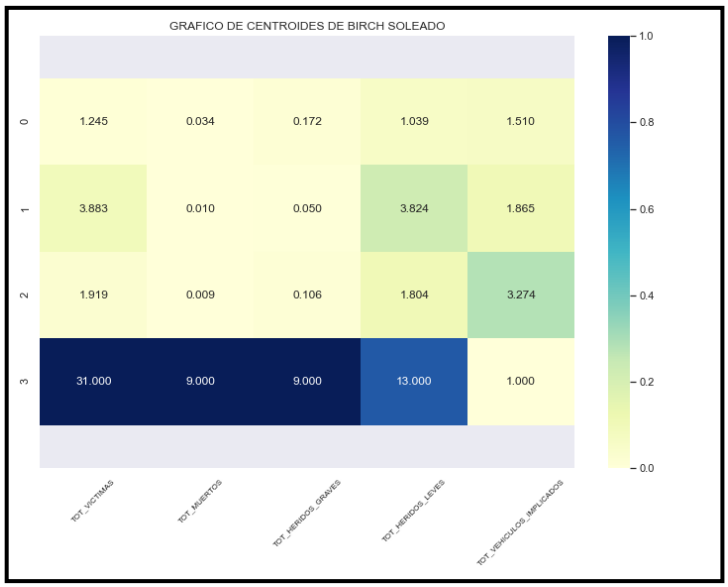


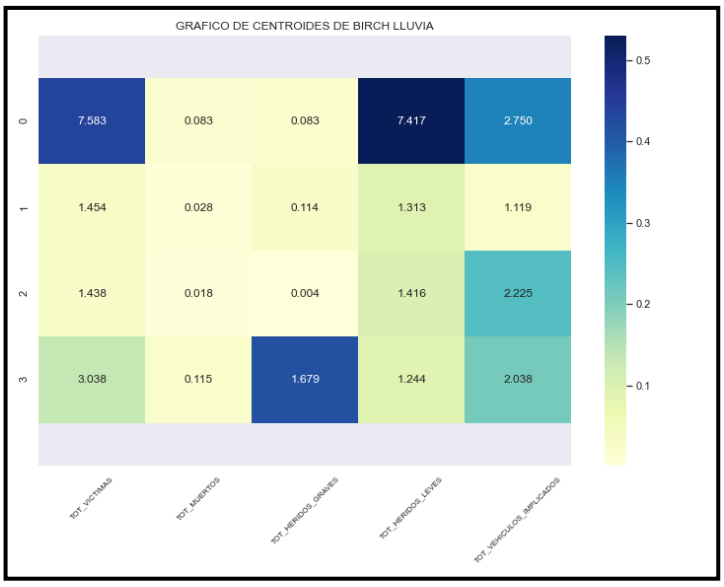


Observamos como para ambas clases los atributos “heridos leves” y “vehículos implicados”, son atributos clave en los clústeres. Vemos una similitud: un número alto de vehículos implica un alto número de heridos leves. Pero para la clase SOLEADO un alto porcentaje de heridos leves implica vehículos y un alto número de víctimas, sin embargo, para la clase LLUVIA un alto número de heridos graves, implica una pequeña cantidad de vehículos implicados.

Birch

Para el algoritmo Birch, no obtenemos algo similar al algoritmo Kmeans.

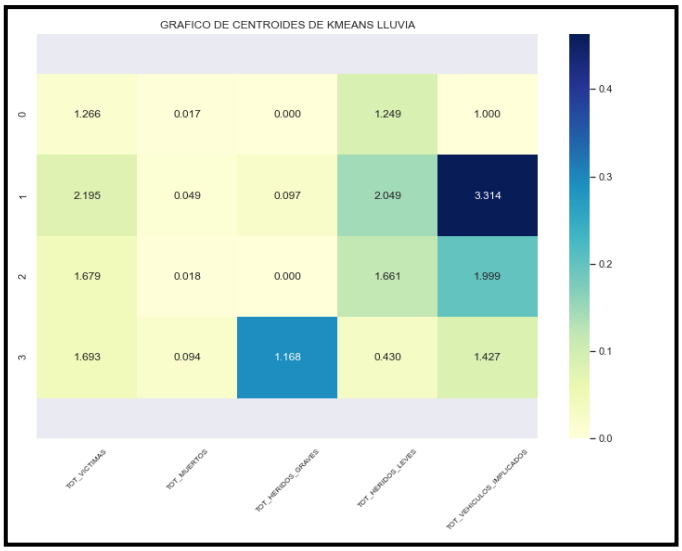


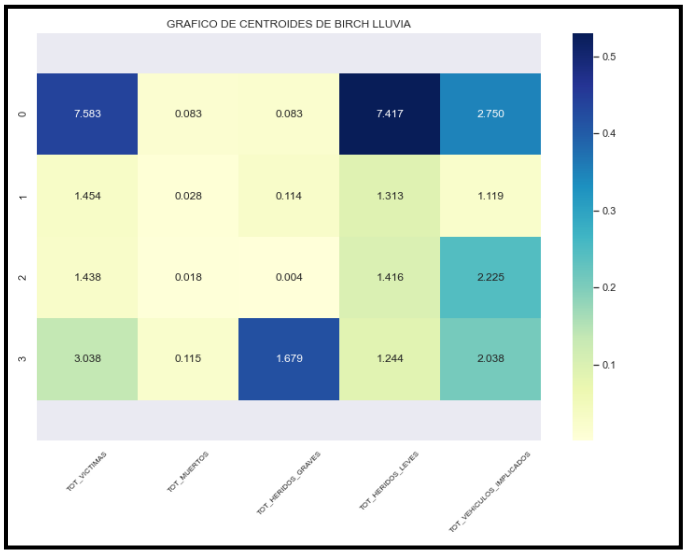


En el algoritmo Birch, vemos cómo sólo en la clase LLUVIA hay un atributo que es determinante: "vehículos implicados", y apenas hay similitud entre ambas clases. En SOLEADO encontramos que hay un clúster donde hay mucha incidencia de víctimas, muertos y heridos graves, y para la clase LLUVIA un alto número de heridos leves o graves, implica un víctimas y vehículos.

Kmeans vs Birch

Para realizar la comparativa usaremos la clase LLUVIA debido a que es nuestro caso de estudio principal y ha sido la más determinante.





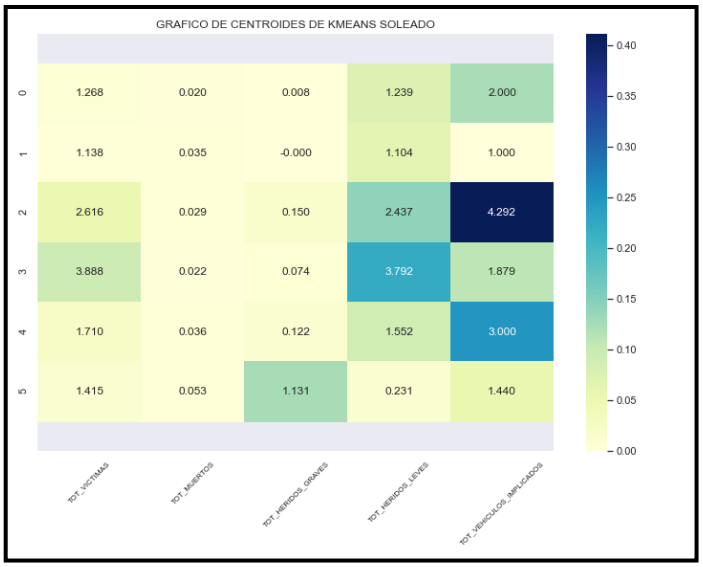
Como son dos algoritmos que tratan los datos y utilizan casuísticas diferentes, no hay un grado alto de similitud entre ellos. Sólo coinciden en que uno de los atributos que determina es el número de vehículos implicados. Para el algoritmo Kmeans son

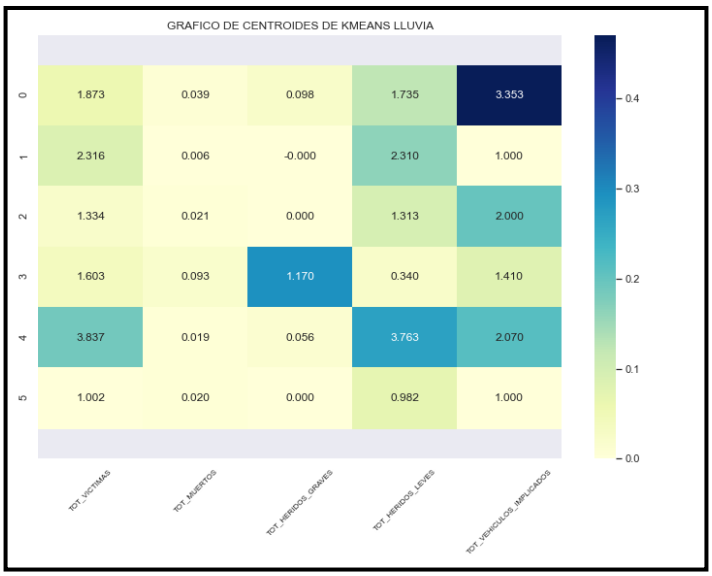
determinantes, en algunos clústeres, los vehículos implicados y los heridos graves y para Birch los heridos leves o graves y el número de víctimas.

Por otro lado, si realizamos una valoración fijándonos en la medida Silhuoete, será más fiable el resultado de Kmeans. Si nos fijamos en los tamaños que tienen cada uno de los clústeres, vemos como tenemos un mejor balanceo en el algoritmo Kmeans y dentro de las clases, LLUVIA presenta una mejor disposición.

K = 6

$Kmeans$

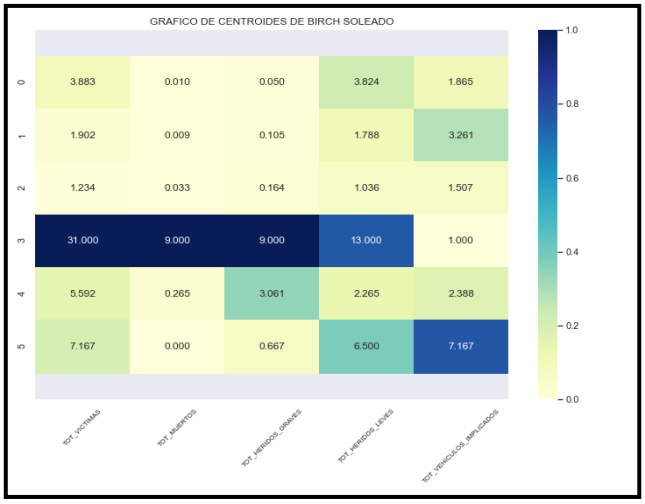


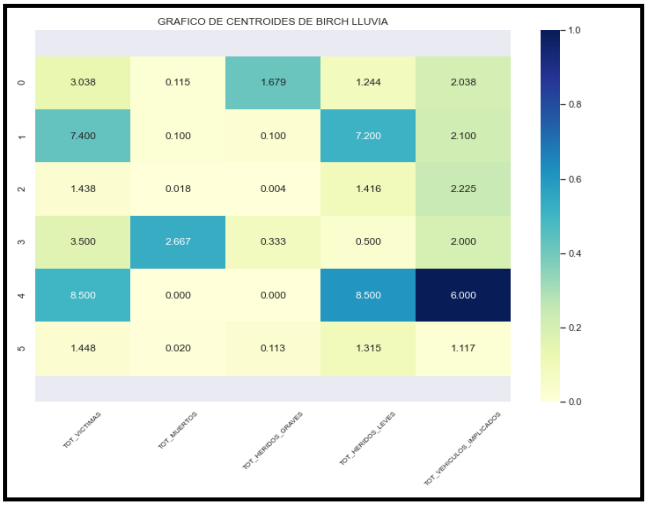


De nuevo, no hay similitudes. Podemos observar cómo hay un patrón en la clase SOLEADO y es que de nuevo los atributos que lo determinan son los vehículos y los heridos leves, luego a la hora de hablar sobre cada clúster no hay nada relevante, ya que es muy similar al anterior. Para la clase LLUVIA se añade el atributo víctimas y vemos que hay nuevas características como que un alto número de heridos leves implica un número alto de víctimas y vehículos o que un elevado número de vehículos implicados conlleva víctimas y heridos leves.

Birch

Para el algoritmo Birch, obtenemos algo similar que para el algoritmo Kmeans.

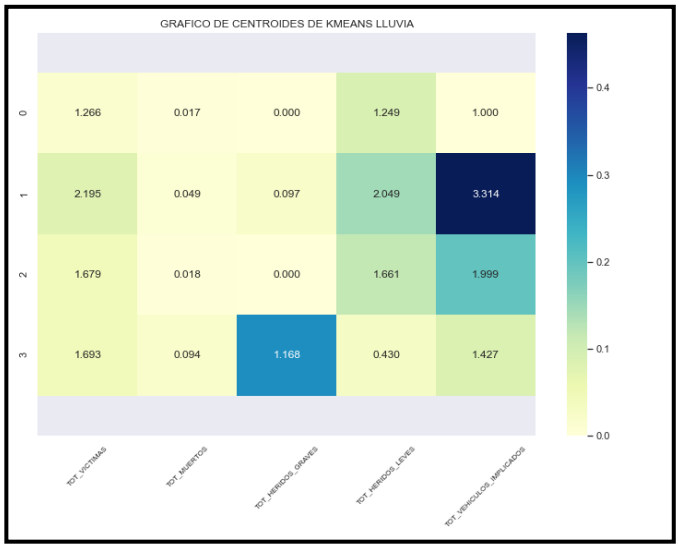


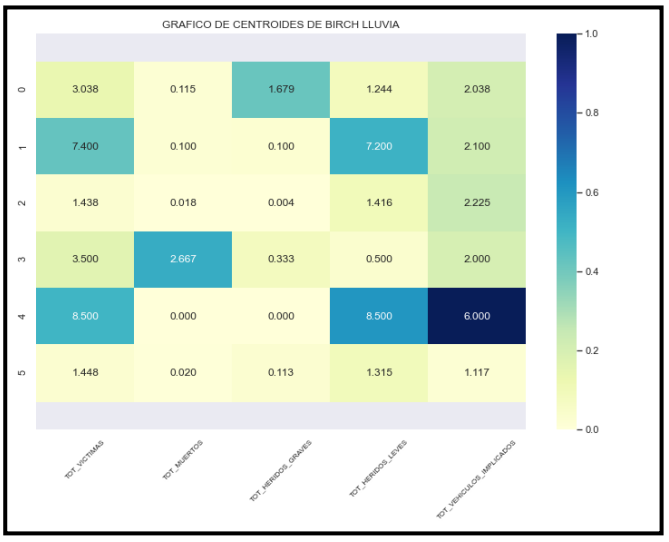


Observamos que algunas de las distribuciones de clústeres también se mantienen. Para este número de clúster no se distinguen de forma tan clara los atributos que destacan, esto se puede deber a que, con el aumento del número de K, disminuye el valor de Silhuoete.

Kmeans vs Birch

Cuando hemos aumentado el número de clústeres, vemos como hay una mayor distribución entre estos. El algoritmo Kmeans nos sigue dando dos atributos que son determinantes, sin embargo, Birch añade victimas como un tercer atributo determinante. No obstante, sigue habiendo diferencias entre ellos.





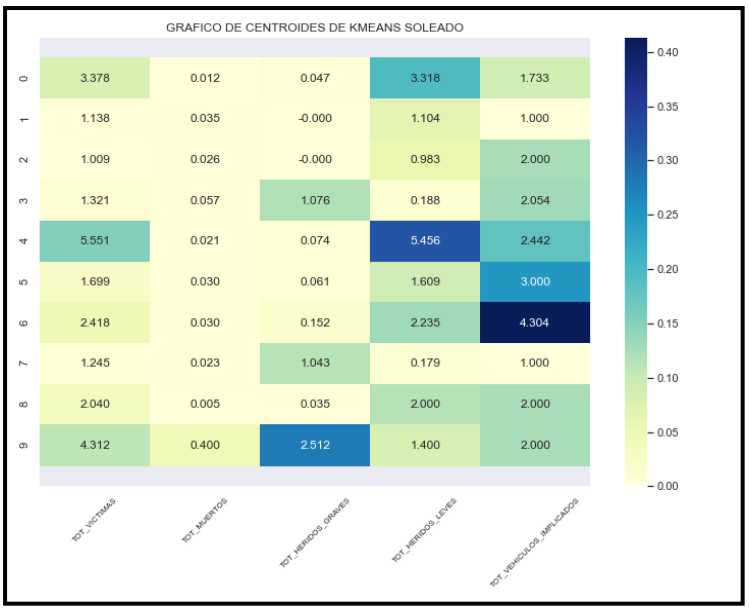
Lo observamos en que Kmeans es muy similar a cuando tenía 4 clústeres y podemos obtener resultados similares, pero para Birch tenemos otros ejemplos: un elevado número de heridos leves nos da un alto número de víctimas o un alto número de vehículos implica un número elevado de heridos leves y víctimas.

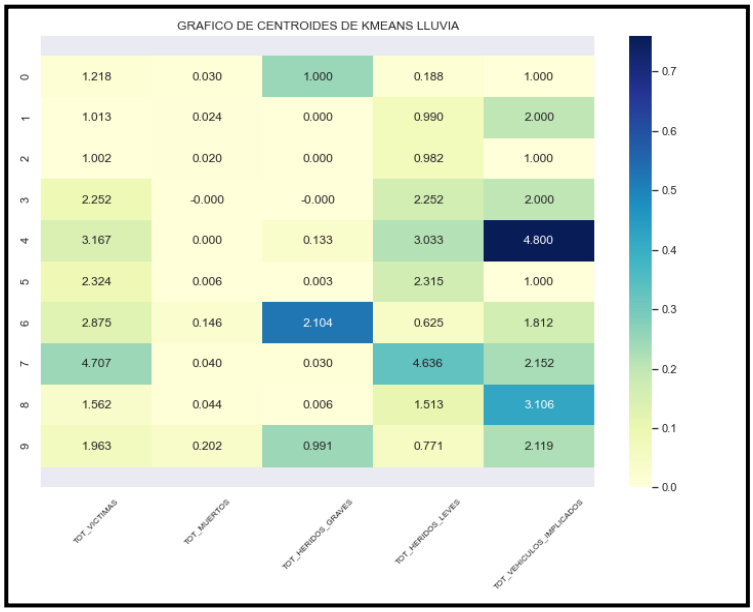
De forma numérica, el algoritmo Birch, empieza a decrementar su valor Silhuoete y obtiene un mal balanceo. Por otro lado, Kmeans aumenta su valor y en el caso de la clase LLUVIA sigue obteniendo un buen balanceo de clústeres.

K = 10

Kmeans

Una vez aumentado de forma considerable el número de clústeres, vemos como en la clase SOLEADO se acentúan los atributos vehículos y heridos leves, como los más determinantes y para la clase LLUVIA se añade el atributo heridos graves que va tomando fuerza para ser considerado como un atributo que pueda determinar.

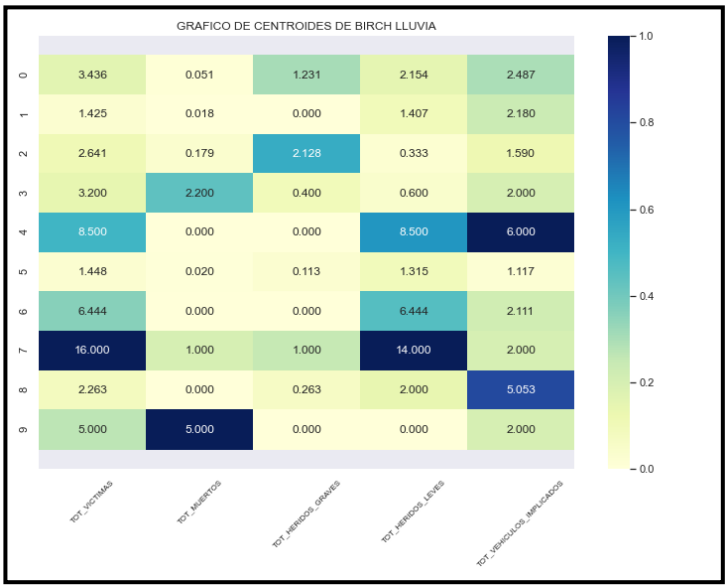


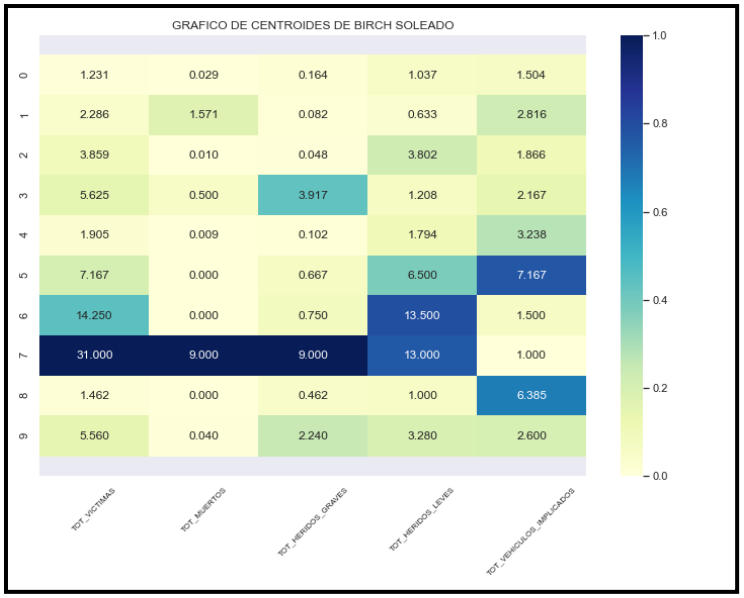


Algunos ejemplos de lo que podemos obtener tras estos gráficos son: Para la clase SOLEADO que un alto número de heridos leves implica vehículos y víctimas o que un alto porcentaje de vehículos involucra heridos leves y para LLUVIA los heridos leves suponen víctimas y vehículos implicados y vehículos implicados conlleva víctimas y heridos leves. De forma similar, encontramos para las dos clases que un número elevado de heridos graves implica víctimas y vehículos implicados.

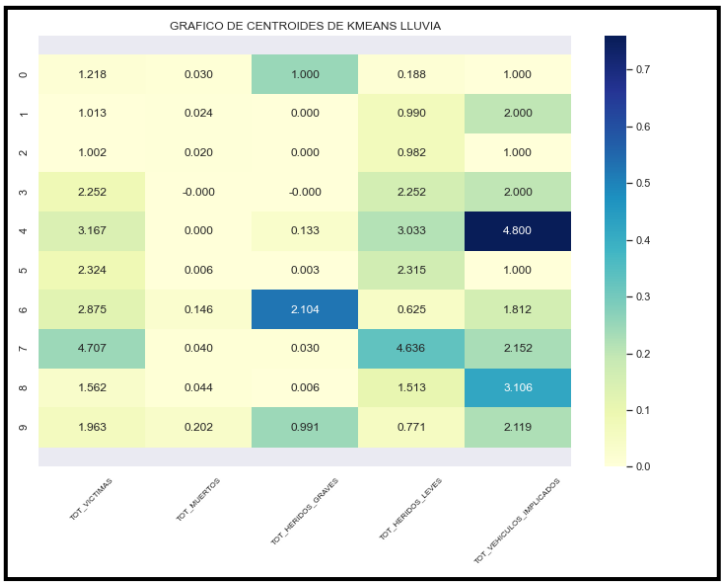
Birch

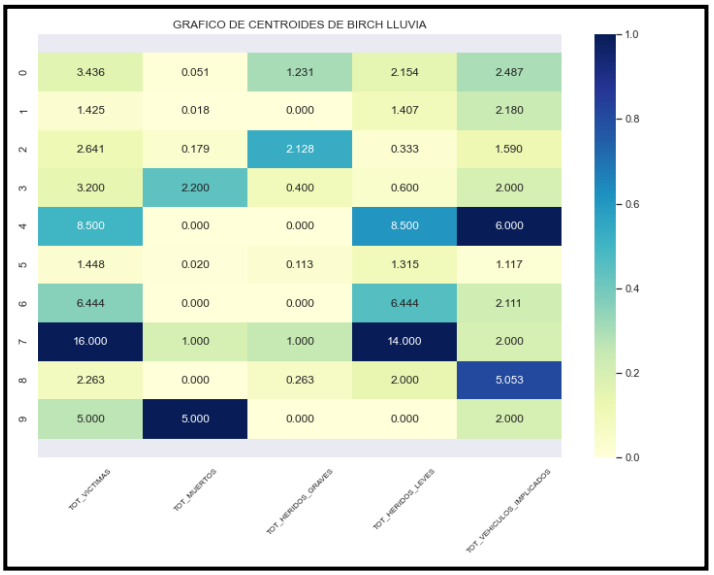
Con el aumento de clústeres, podemos empezar a ver el porqué de los malos resultados de este algoritmo, y es en la cantidad de dispersión que encontramos en los datos. No hay atributos que determinen con claridad ninguna de las dos clases. Al igual que el balanceo de estos, es muy desigual.





Kmeans vs Birch





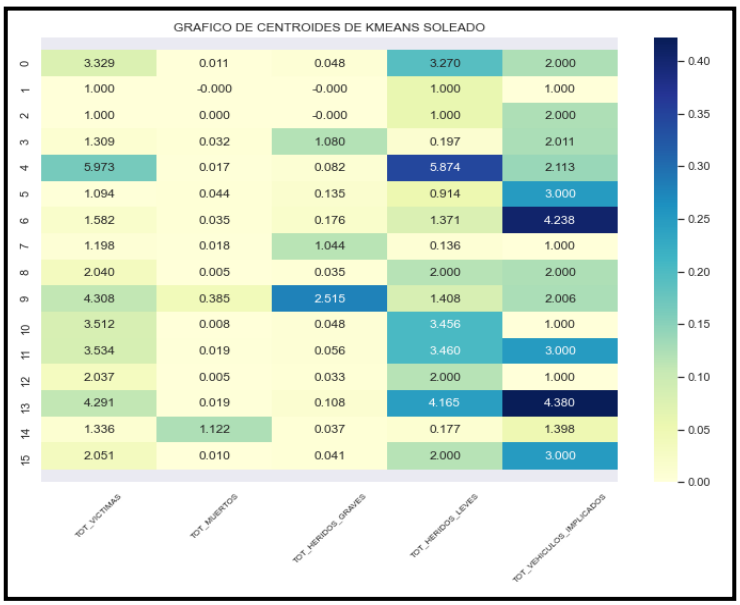
Al aumentar de nuevo el número de clústeres, no obtenemos una similitud entre los dos algoritmos. La principal razón para que ocurra esto, es que al aumentar el número de clúster en donde se puede clasificar la información, cada algoritmo analiza y distribuye los datos con su heurística. No obstante, hay clústeres que son similares entre los dos algoritmos. Por ejemplo, en Kmeans el clúster 4 es similar al clúster 8 en Birch el clúster 7 en Kmeans y el 7 en Birch. Esto nos lleva a la conclusión de que se obtienen datos similares, pero cada algoritmo asigna los clústeres de forma individual y diferente, pero

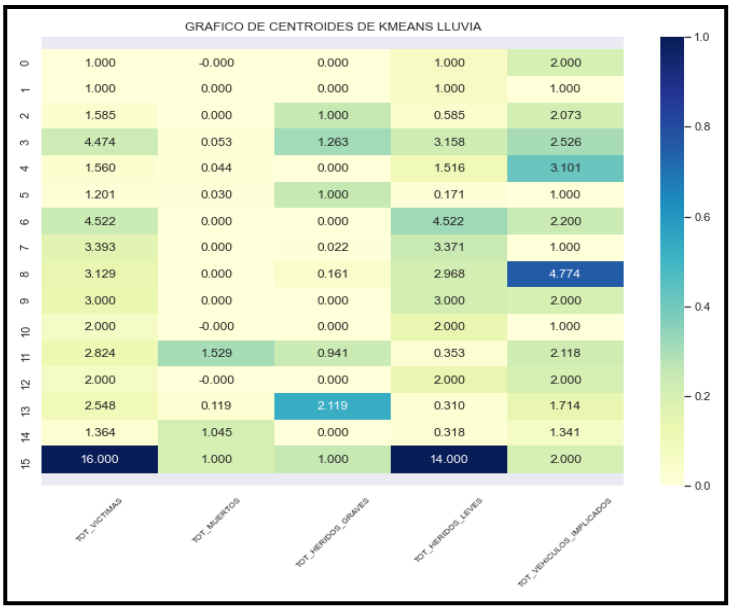
que la clasificación de los datos empieza a ser similar entre ellos. Tras comparar estas dos gráficas, queda patente la distribución que hace Birch. Observamos como en cada atributo hay siempre uno de color muy oscuro, en el clúster 9 un número alto de muertos implica víctimas y vehículos, y en Kmeans destacan los dos atributos determinantes. Nos damos cuenta, de que también observamos esa diferencia de balanceo entre clúster en el que Kmeans hace un reparto más equitativo[5]. De forma numérica, vemos como el algoritmo Kmeans, vuelve a colocarse como el "mejor" algoritmo para abordar este problema.

K = 16

Kmeans

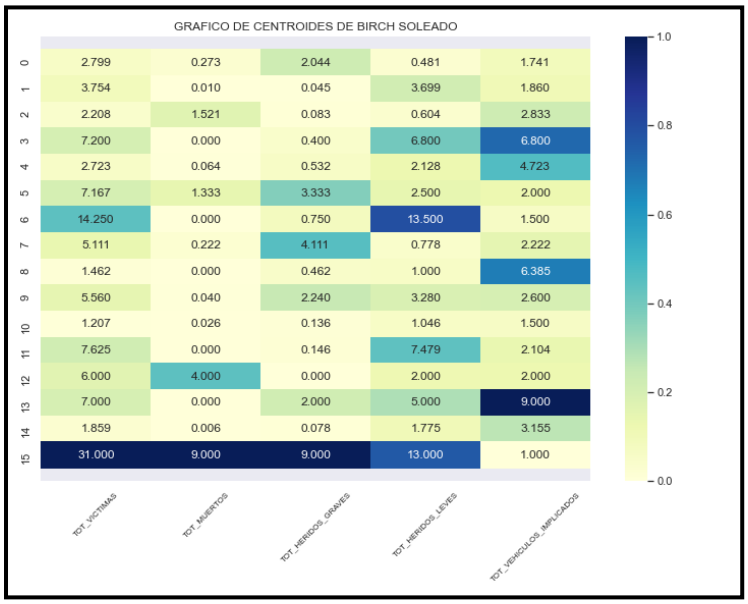
Cuando realizamos el estudio con el mayor número de clúster para Kmeans que tiene sentido utilizar, vemos como es la clase SOLEADO quien acentúa los dos atributos de los que veníamos hablando. La clase LLUVIA realiza una distribución más equitativa y por ello no destaca tanto. Esto también se debe a que para este caso el número de clústeres es muy elevado para el tamaño de datos que tenemos, lo que hace que al repartir los datos entre 16^2 clústeres estén casi vacíos. No obstante, de forma numérica obtenemos mejor resultado con la clase LLUVIA.

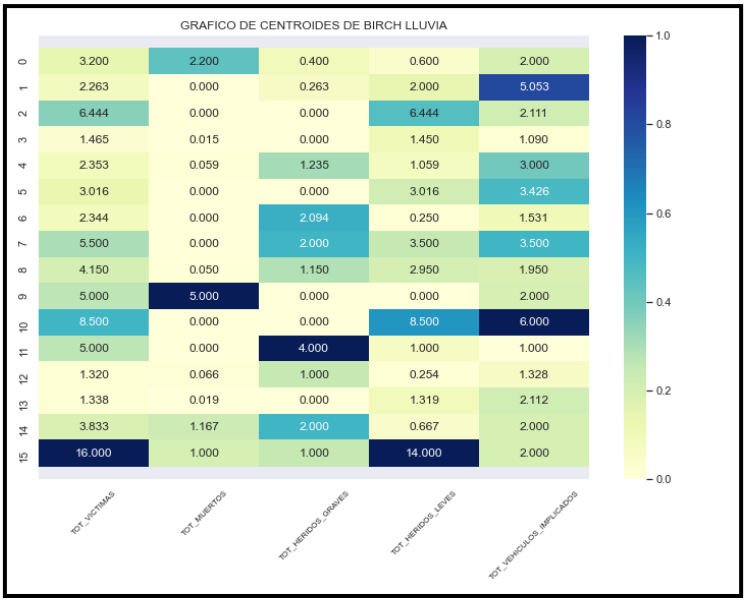




Birch

Para el algoritmo Birch, obtenemos algo muy diferente que para el algoritmo Kmeans.

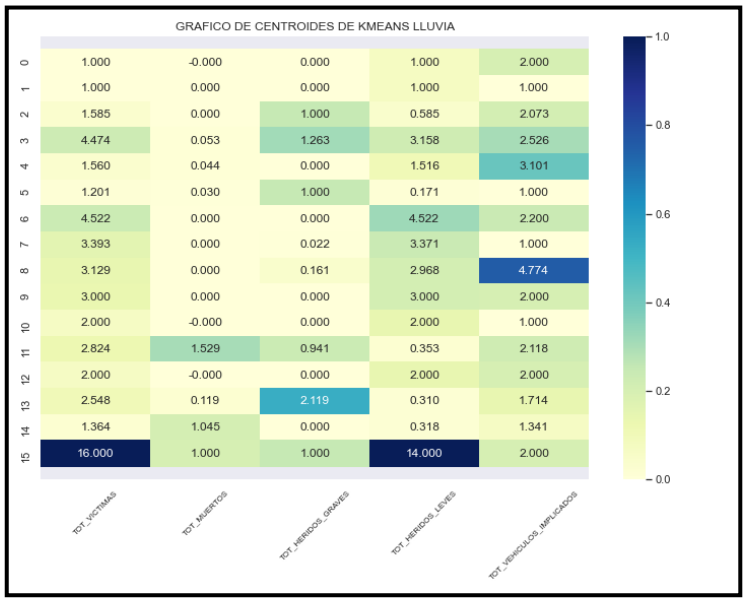


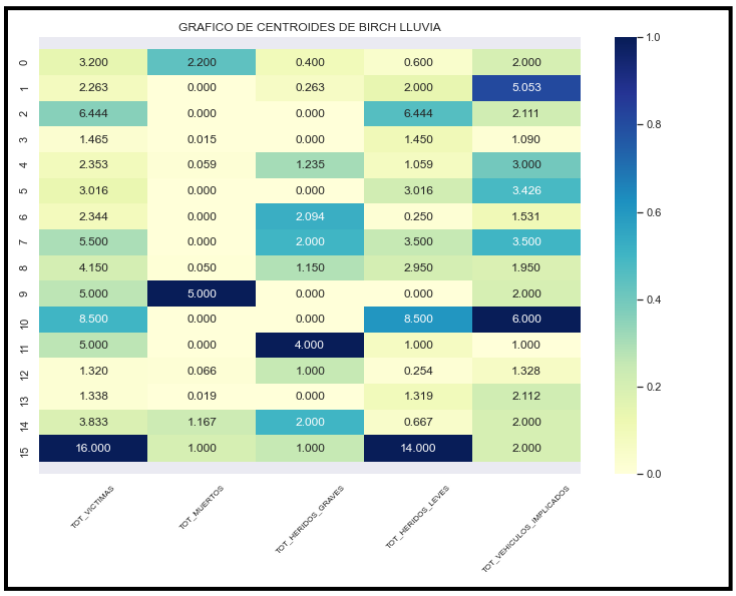


Hemos hablado anteriormente del mal balanceo que realiza el algoritmo Birch, y queda reflejado de forma más impresionante en este número de clústeres y es que a diferencia con Kmeans para la clase LLUVIA no se nota que no hay tantos datos para el número de clúster y que sigue teniendo celdas con un color muy oscuro.

Kmeans vs Birch

Para realizar la comparativa usaremos el gráfico de Andalucía, debido a que es nuestro caso de estudio principal.





Encontramos el mismo patrón que para el anterior punto. Por ejemplo, en Kmeans el clúster 13 es similar al clúster 12 en Birch (un índice elevado en muertos implica víctimas y heridos graves), el clúster 15 en Kmeans y el 15 en Birch (altos porcentajes en heridos graves dan como resultado alto porcentaje de víctimas). Esto nos lleva a la conclusión de que se obtienen datos similares, pero cada algoritmo asigna los clústeres de forma individual y diferente, pero que la clasificación de los datos empieza a ser similar entre ellos. De forma numérica, vemos como el algoritmo Kmeans, vuelve a colocarse como el "mejor" algoritmo para abordar este problema.

Conclusión para el Caso 1

Tras evaluar en varios clústeres y ver las medidas y balanceos podemos decir que en general para este caso abordado hay dos atributos que resultan ser determinantes para el estudio: Vehículos implicados y los heridos leves. Como curiosidad vemos que a pesar de que las carreteras convencionales sean las más peligrosas, sólo es en el sentido de que ocurren muchos accidentes, y, por suerte, no hay un número elevado de muertes.

Basándonos en las medidas obtenidas, vemos como la clase LLUVIA es la que mejor resultados ha conseguido en ambos algoritmos. Para el caso de Kmeans se llega a

rozar el 90% de similitud y para el Birch, a excepción del último caso, se ronda un máximo de un 50% de similitud. Otro estudio clave, es el balanceo de los clústeres, en los que Birch acumula en solo dos clústeres la totalidad de los datos de estudios, lo que nos lleva a la conclusión de que las características de esos casos abarcan la totalidad y por ello realiza un mal balanceo. Para el caso de la clase SOLEADO en la que tenemos 23.000 datos, este es el balanceo de la clase Birch:

10: 19133, 1: 1716, 14: 1681, 0: 293, 4: 94, 11: 48, 2: 48, 9: 25, 7: 18, 8: 13, 5: 6, 3: 5, 6: 4, 15: 1, 13: 1, 12: 1

Como observamos, en el clúster número 10 se lleva prácticamente el 90% de los datos, y el resto, 15 clústeres, se tienen que repartir los 4.000 datos restantes. Esta es una clara diferencia entre los algoritmos, si nos fijamos ahora en el de Kmeans:

1: 6721, 2: 6254, 8: 2166, 7: 1537, 12: 1105, 3: 1101, 0: 1036, 5: 746, 14: 435, 15: 411, 10: 377, 11: 322, 4: 293, 6: 256, 9: 169, 13: 158

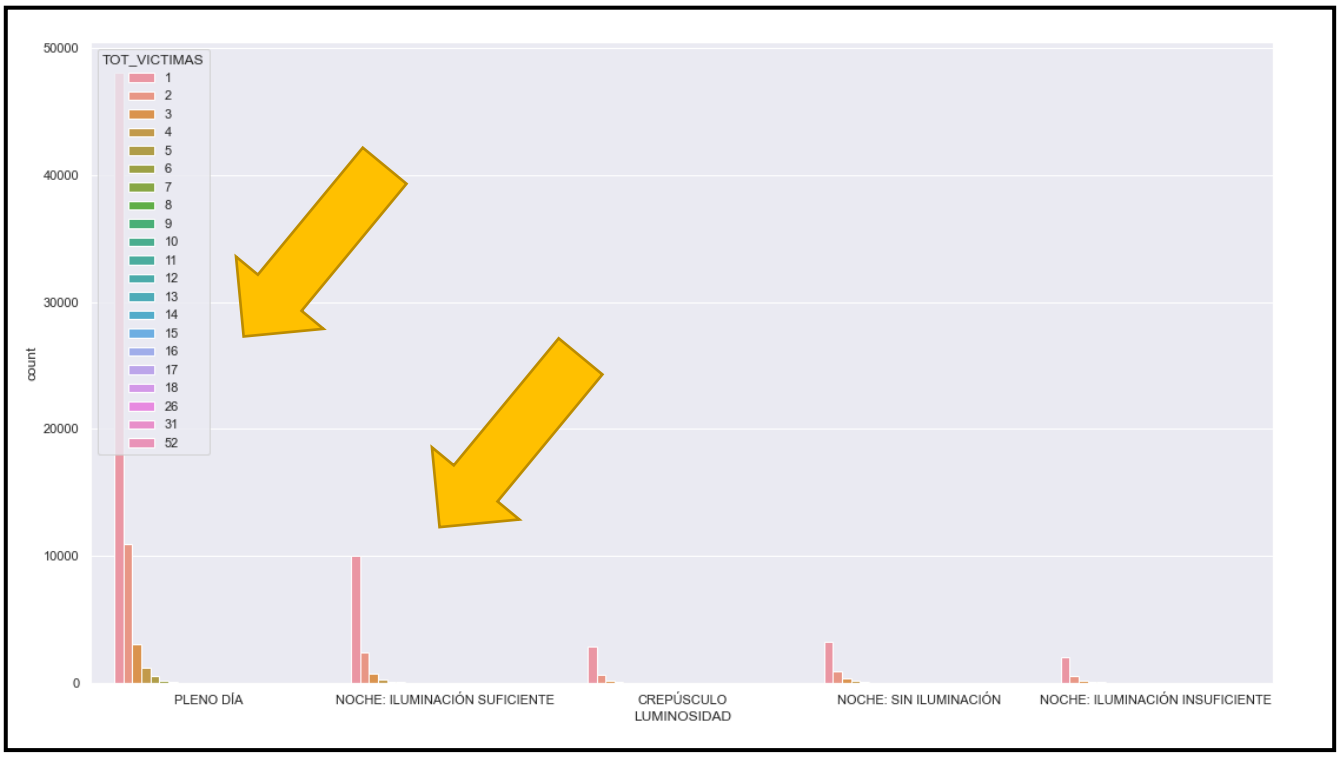
Vemos como tiene una mejor distribución y por eso es por lo que obtenemos unas mejores mediciones y unos mejores mapas de calor y por lo que podríamos decir que, tras probar sólo con dos algoritmos, que usar Kmeans para este problema podría ser muy beneficioso.

Caso 2

Para este segundo estudio nos vamos a basar en otro informe de la DGT. Esta vez intentamos comprobar si hay un alto porcentaje de accidentes en relación con el tramo horario. Según los datos [6]recogidos, la mayor parte de los accidentes ocurren durante la entrada y salida del trabajo y, por tanto, ocurren durante el día, sin embargo, encontramos otros informes que nos dicen que ocurren más accidentes de noche que de día.

Para comparar, voy a escoger un tipo de zona que tiene alto porcentaje de accidentes como es la zona urbana, esto es debido a que, en ocasiones, las luces, intersecciones o semáforos producen más distracciones que en una carretera.

Justificación del tramo horario

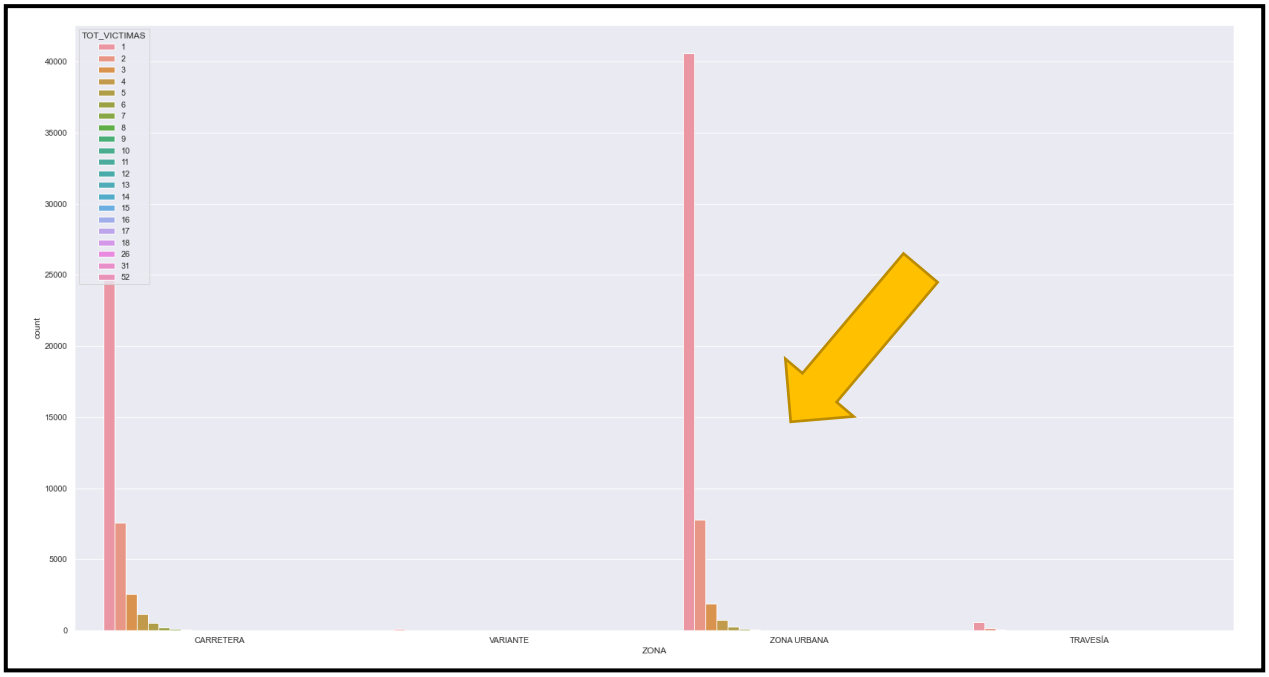


Como podemos observar, los tramos horarios que más datos tienen son a "Pleno día" y "noche con iluminación" que serán las clases que vamos a estudiar. Para ello también nos vamos a apoyar en el informe[7].

[6] https://www.caranddriver.com/es/coches/planeta-motor/a33542174/accidentes-trafico-causas-habituales/

[7] http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/accidentes-urban/accidentes_trafico005.pdf

Justificación de la Zona



Como observamos, donde más víctimas hay es en la zona urbana, tal y como se recoge en el informe citado anteriormente[6] de la DGT.

Interpretación de la segmentación

Relación y tamaño de los clústeres

	Comunidad	Día	Noche
	N.º Clústeres	Tamaño de cada clúster (Balanceo)[8]	
Kmeans	4	0: 19194, 2: 11010, 1: 6286, 3: 732	1: 5194, 3: 3355, 0: 2228, 2: 371
	6	0: 17758, 2: 9255, 3: 6261, 5: 3035, 1: 727, 4: 186	4: 4746, 1: 2751, 5: 2214, 0: 979, 2: 369, 3: 89
	10	1: 16426, 2: 9255, 0: 4926, 9: 1624, 7: 1437, 4: 1425, 8: 1230, 5: 583, 3: 186, 6: 130	2: 4647, 0: 2751, 1: 1319, 3: 668, 4: 541, 7: 436, 9: 392, 6: 164, 8: 141, 5: 89
	16	1: 16426, 2: 9255, 8: 3480, 5: 1502, 7: 1222, 11: 1107, 9: 934, 0: 900, 15: 606, 12: 393, 14: 368, 6: 323, 3: 258, 10: 241, 4: 186, 13: 21	4: 4336, 1: 2751, 2: 1159, 7: 471, 3: 455, 13: 383, 8: 368, 0: 329, 11: 200, 15: 180, 14: 144, 5: 128, 6: 89, 12: 63, 10: 56, 9: 36
Birch	4	0: 37000, 2: 187, 3: 21, 1: 14	0: 10883, 1: 161, 2: 89, 3: 15
	6	0: 36999, 5: 184, 3: 21, 1: 14, 2: 3, 4: 1	5: 9974, 2: 861, 1: 161, 0: 89, 4: 48, 3: 15
	10	2: 36213, 1: 786, 0: 184, 5: 19, 8: 13, 3: 2, 6: 2, 9: 1, 4: 1, 7: 1	5: 9974, 8: 854, 1: 158, 0: 85, 4: 48, 6: 10, 9: 7, 3: 5, 2: 4, 7: 3
	16	1: 36191, 2: 675, 14: 167, 5: 111, 12: 22, 4: 18, 3: 15, 8: 13, 0: 2, 6: 2, 13: 1, 9: 1, 10: 1, 15: 1, 11: 1, 7: 1	7: 9022, 6: 952, 3: 854, 12: 103, 0: 84, 5: 55, 2: 46, 13: 8, 4: 7, 1: 5, 15: 3, 9: 3, 14: 2, 8: 2, 11: 1, 10: 1

[8] A esto nos referiremos cuando hablamos del balanceo de clústeres.

El tamaño para el DF Día es de 37.222 y para Noche 11.148.

Comparativa de mediciones

Silhuoete

Comunidad	Día				Noche			
Nº Clusters	4	6	10	16	4	6	10	16
Kmeans	0.669	0.805	0.891	0.959	0.608	0.753	0.831	0.93
Birch	0.756	0.775	0.730	0.686	0.711	0.592	0.575	0.528

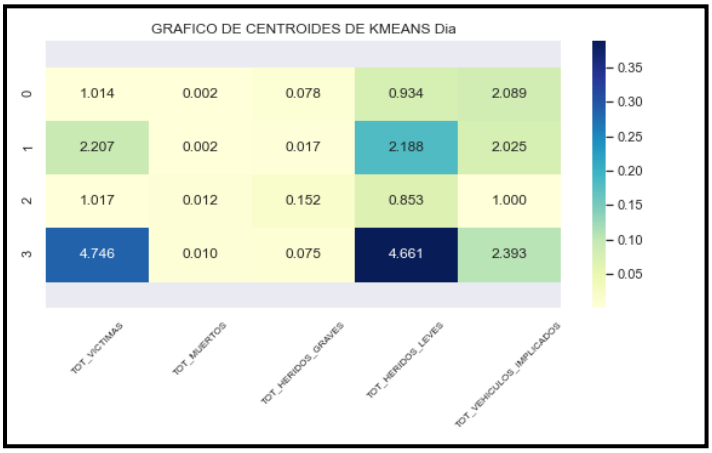
Calinsky

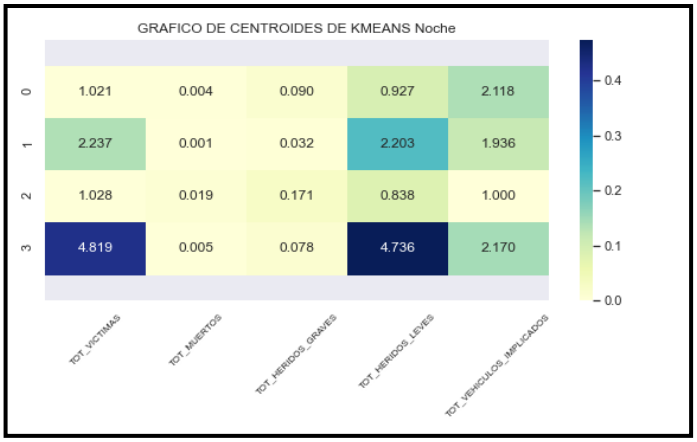
Comunidad	Día				Noche			
Nº Clusters	4	6	10	16	4	6	10	16
Kmeans	3311.955	34019.83	42264.31	57394.38	3432.337	7957.69	9780.64	13999.9
Birch	2402.52	1484.903	855.404	1869.66	1684.74	2732.23	1609.70	1625.067

Gráficas de centroides

K = 4

Kmeans

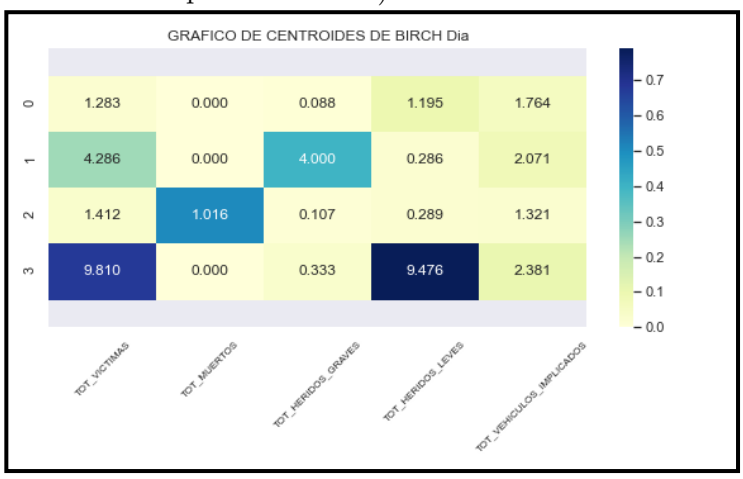


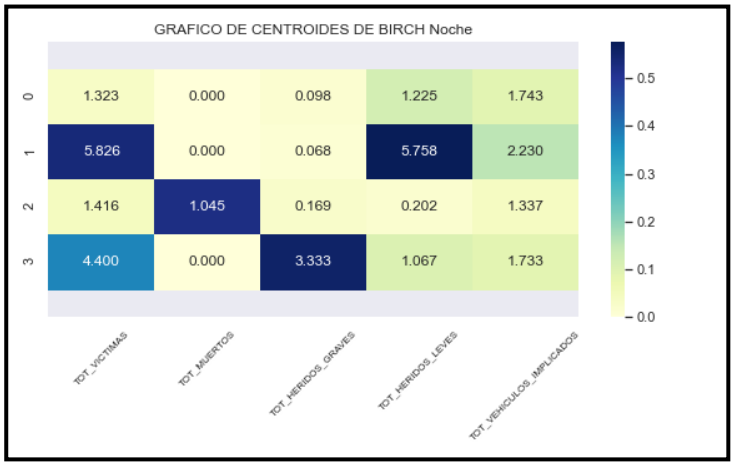


Para un número de clústeres pequeños, tenemos el problema de que queremos incluir en solo 4 clústeres muchos datos con una cantidad elevada de variables. Al representar el gráfico de los centroides, vemos como encontramos similitudes para las dos comunidades. A priori, vemos como hay dos atributos que determinan las clases: "*heridos_leves*" y "*vehículos_implicados*". Tenemos dos gráficas similares, a pesar de la diferencia de datos. Un alto número de heridos leves implica víctimas y vehículos.

Birch

Para este caso volvemos a tener el mismo problema con el algoritmo Birch, que es el mal balanceo de los clústeres. Y tienen similitud entre los gráficos como también sucede en el apartado anterior. Podemos decir que es determinante (aunque por muy poca diferencia) el atributo heridos leves.

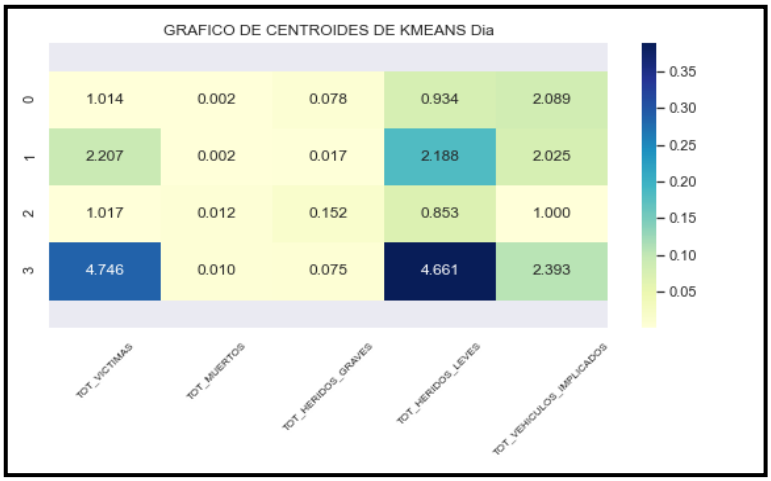


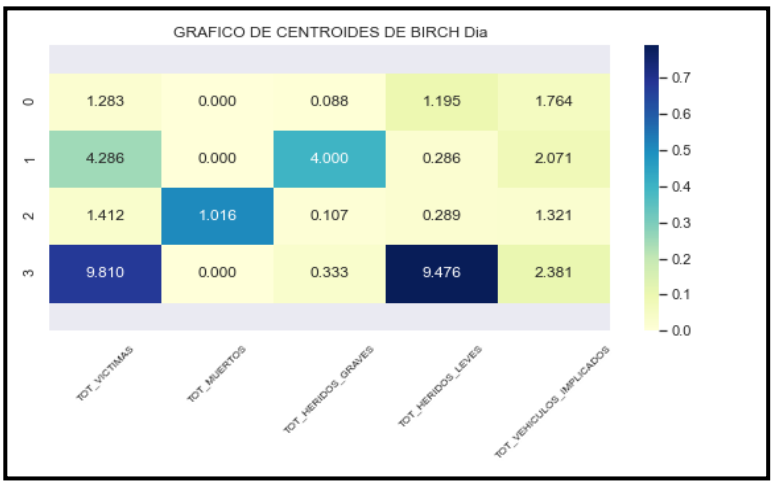


Encontramos, que, para un alto porcentaje de heridos leves, hay un alto número de víctimas y un alto porcentaje de heridos leves o graves implica un alto nivel de víctimas.

Kmeans vs Birch

Para realizar las comparaciones entre los dos algoritmos utilizaremos la clase DÍA que es con la que mejores datos obtenemos.



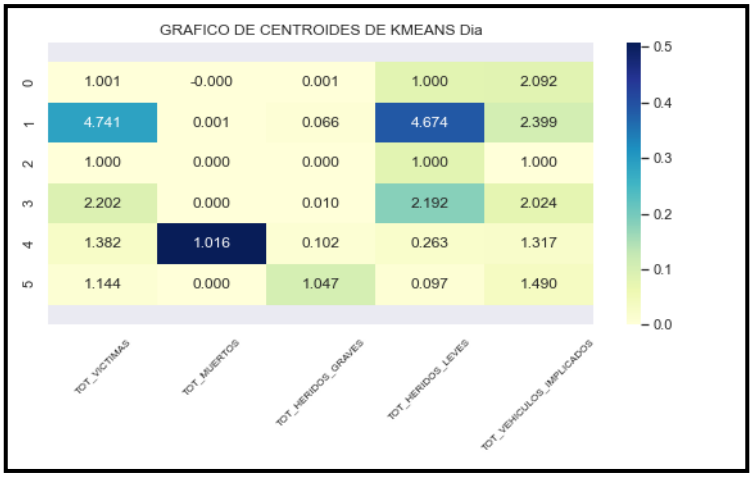


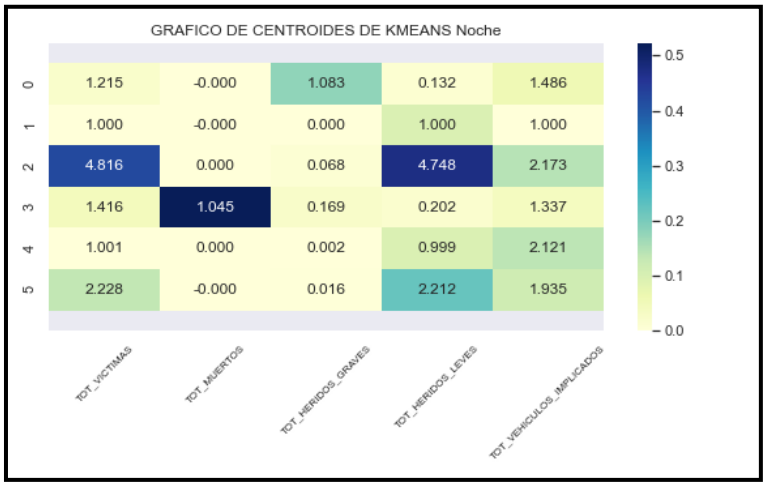
Para este número de clústeres, tenemos el mismo inconveniente que para el caso 1, en el que queríamos agrupar muchos datos en tan solo 4 clústeres. Para el algoritmo Kmeans vemos como hay dos atributos que pueden determinar la clase (Heridos leves y vehículos implicados) mientras que para Birch sigue habiendo dispersión de los datos.

Por otro lado, si realizamos una valoración fijándonos en la medida Silhuoete, el algoritmo Birch, para este "k", tiene un mejor resultado, pero en relación con la medida Calinsky, será más fiable el resultado de Kmeans.

K = 6

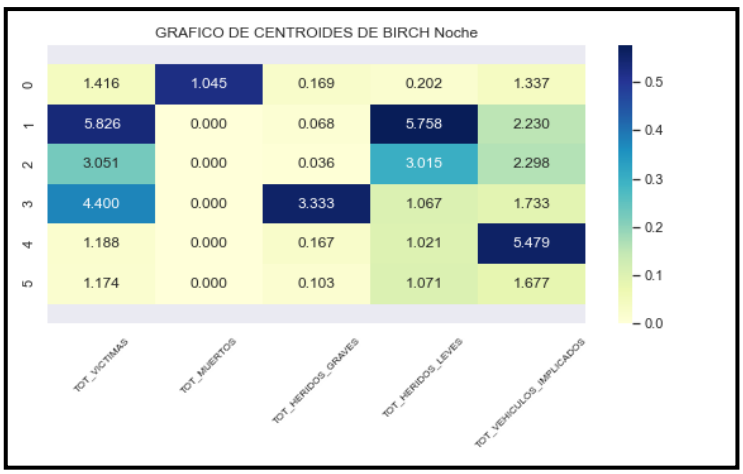
Kmeans

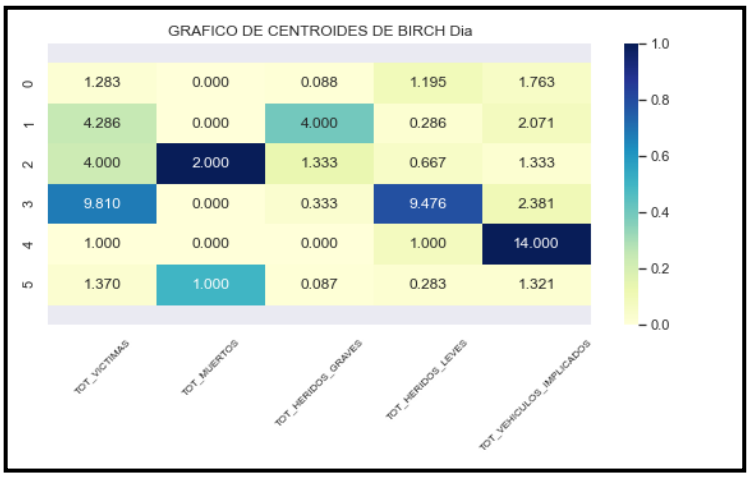




Al aumentar el número de clústeres nos fijamos como prácticamente es el mismo gráfico, pero desordenado. Para la clase DÍA el clúster 2 es el mismo que para NOCHE el clúster 3, el 4 con el 3, el 3 con el 5 el 5 con el 0... Podremos decir que los datos de estos accidentes son similares, aunque para la clase DÍA haya 20.000 datos más. En la medida Silhuoete es la primera clase la que obtiene mejor medida.

Birch

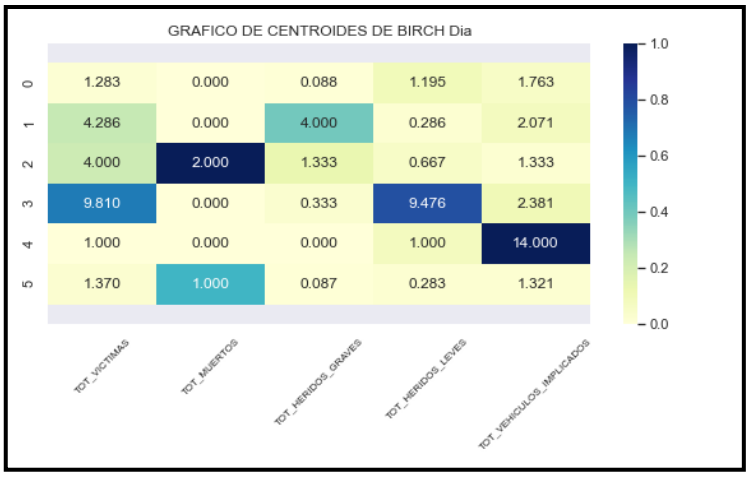


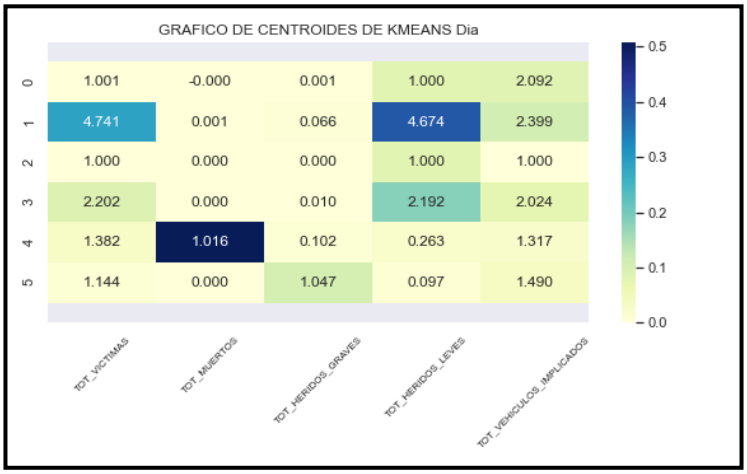


Para este algoritmo, seguimos sin poder verificar si hay algún atributo determinante para alguna de las clases. A diferencia de Kmeans, las gráficas obtenidas no son similares. Sólo consigue una agrupación igual entre el clúster 1 y el clúster 3

respectivamente. Para un alto porcentaje de heridos leves, hay un alto número de víctimas.

Kmeans vs Birch





Para este número de clústeres, si obtenemos algunas similitudes entre los dos algoritmos. Por ejemplo, en el clúster 3 una tasa elevada de heridos leves, nos indica que también hay víctimas y vehículos implicados.

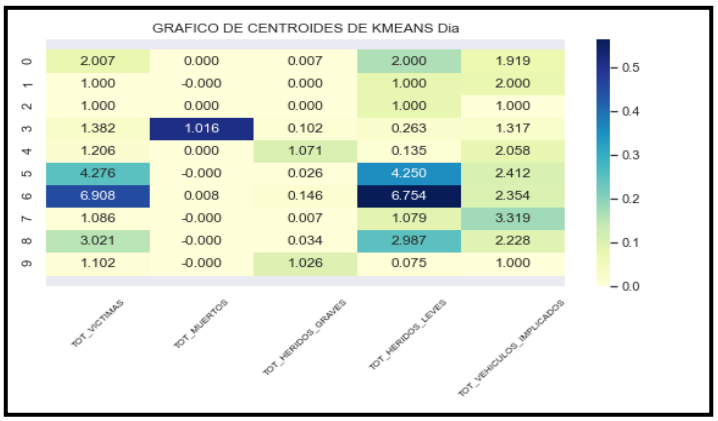
De esta forma, podemos observar que, al aumentar el número de clústeres para los algoritmos, aumenta la similitud y las mediciones salen más altas. ¿A qué puede ser debido? A que ahora, la cantidad elevada de datos que tiene que clasificar dentro de un clúster, puede estar más separada y distribuida. Sin embargo, no hay un atributo que determine la clase y parece haber una distribución más concreta.

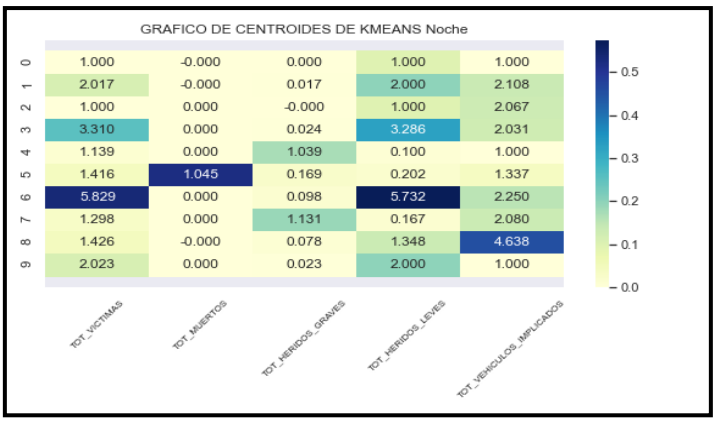
De forma numérica, vemos como el algoritmo Kmeans, ya se coloca como el "mejor" algoritmo para abordar este problema.

K = 10

Kmeans

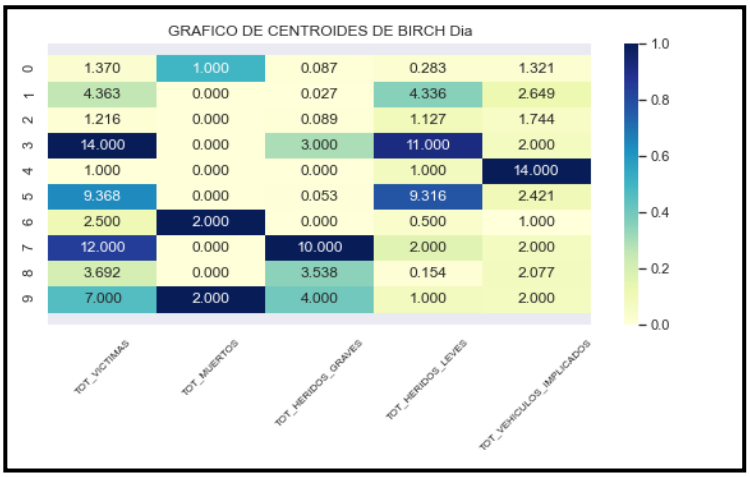
Finalmente, tras aumentar de forma considerable el número de clústeres, vemos como el atributo heridos leves, empieza a ser relevante para ambas clases, y volvemos a tener, aunque esta vez no de forma tan directa, una alta similitud entre los dos gráficos.

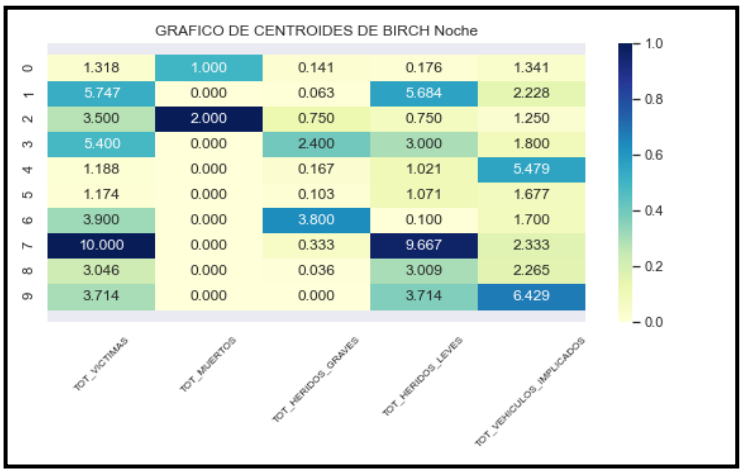




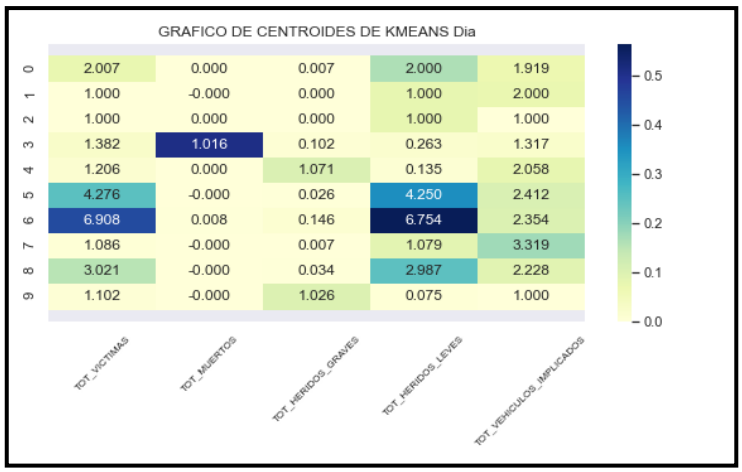
Birch

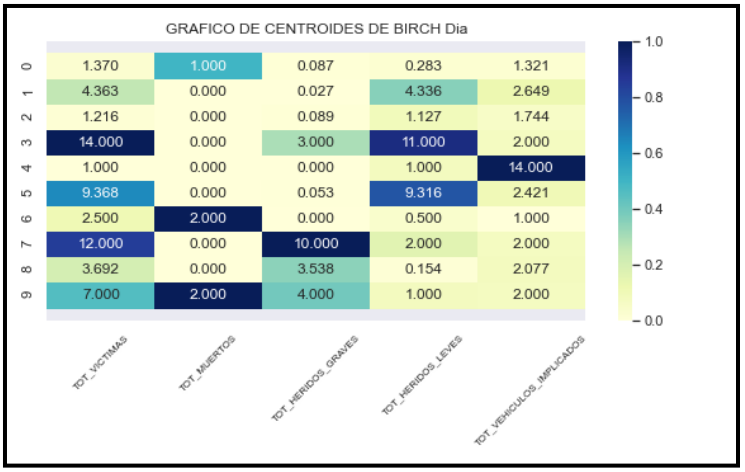
Para el algoritmo Birch, obtenemos algo similar que para el algoritmo Kmeans. El atributo que determina es el número de víctimas y para la clase DÍA, que hay más datos, obtenemos una mayor distribución de los datos.





Kmeans vs Birch





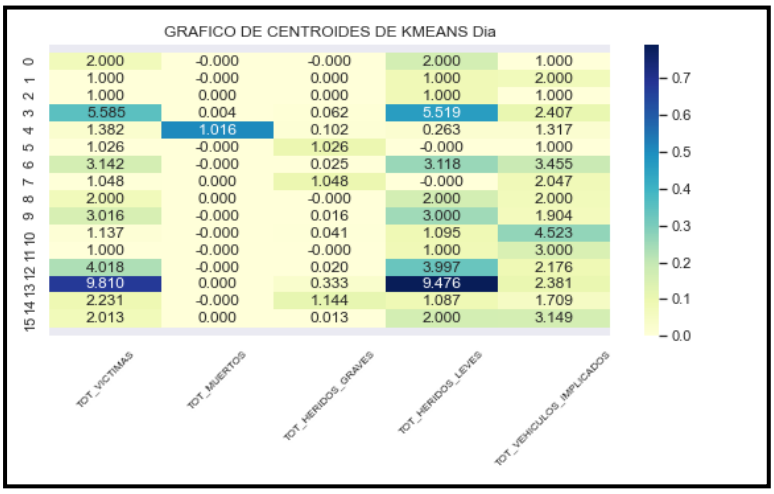
Al aumentar de nuevo el número de clústeres, no obtenemos una similitud entre los dos algoritmos. La principal razón para que ocurra esto, es que al aumentar el número de clúster en donde se puede clasificar la información, cada algoritmo analiza y distribuye los datos con su heurística. No obstante, hay clústeres que son similares entre los dos algoritmos. Por ejemplo, en Kmeans el clúster 3 es similar al clúster 6 en Birch (un índice elevado en muertos implica víctimas y heridos leves), el clúster 6 en Kmeans y el 3 en Birch (altos porcentajes en víctimas y heridos leves dan como resultado vehículos implicados) y el clúster 0 en Kmeans y el 3 en Birch (índice alto en heridos leves, nos da un índice en víctimas). Esto nos lleva a la conclusión de que se obtienen datos similares, pero cada algoritmo asigna los clústeres de forma individual y diferente, pero que la clasificación de los datos empieza a ser similar entre ellos.

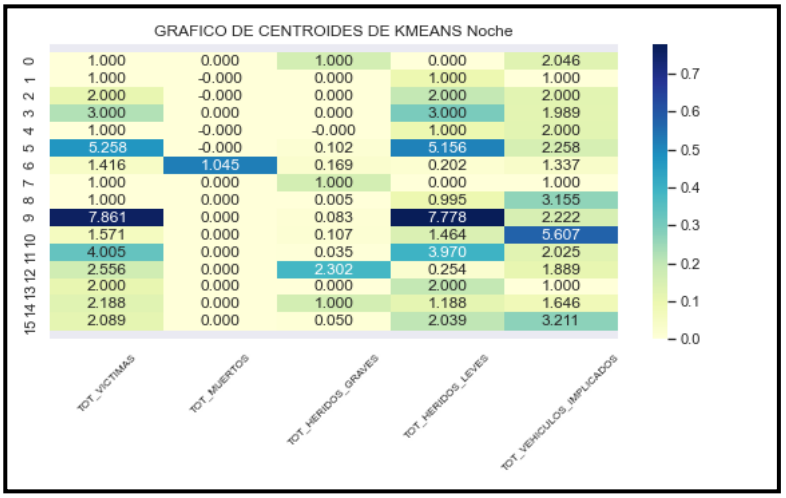
De forma numérica, vemos como el algoritmo Kmeans, vuelve a colocarse como el "mejor" algoritmo para abordar este problema.

K = 16

Kmeans

Cuando ponemos el último número de clúster, vemos como obtenemos un peor resultado en el gráfico. Aunque la cantidad de datos es de cerca de 37.000 datos, los que

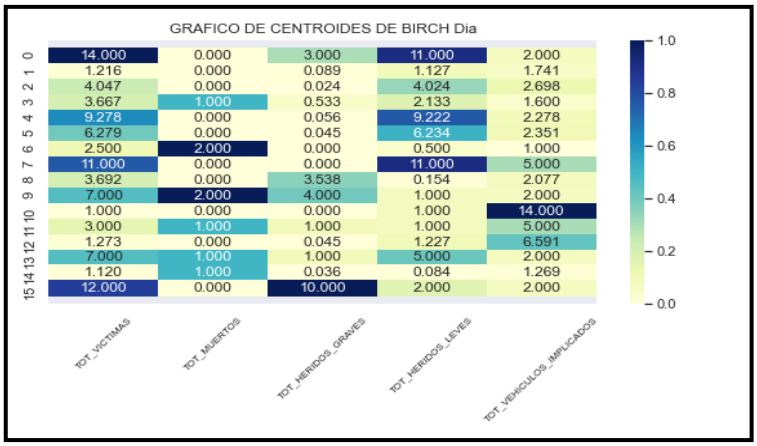


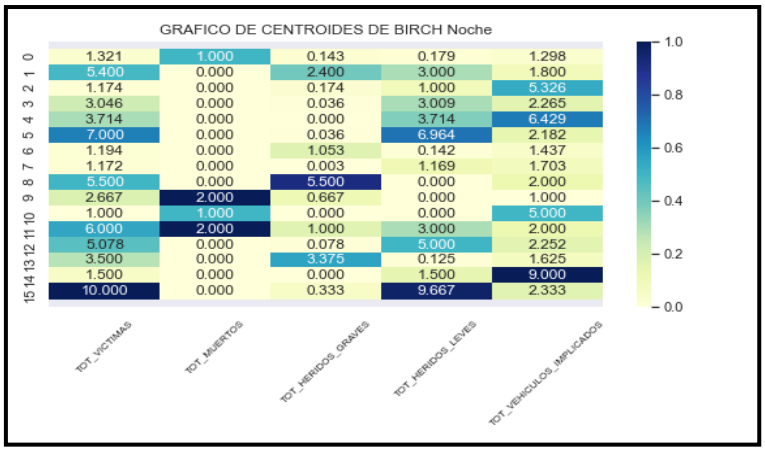


se evalúan se clasifican en varios grupos muy claros y se quedan vacíos una gran cantidad de bloques. El atributo herido leves, finalmente, si se queda como atributo determinante. Esta vez, el grado de similitud entre ambas clases se vuelve a reducir con lo que podemos comprobar que no eran tan similares, no obstante, habrá grupos de accidentes y características que si lo sean.

Birch

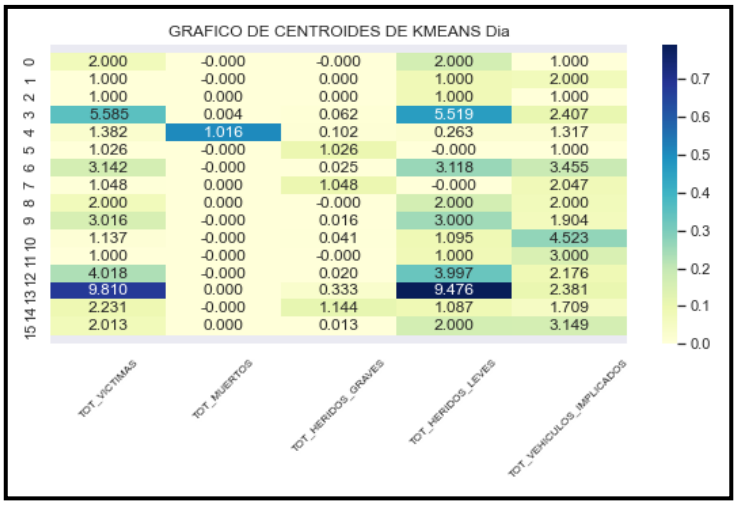
Para el algoritmo Birch, obtenemos algo similar que para el algoritmo Kmeans. Se mantiene el atributo victimas como determinante y la agrupación de clústeres no es muy clara, esto también se debe al mal balanceo.

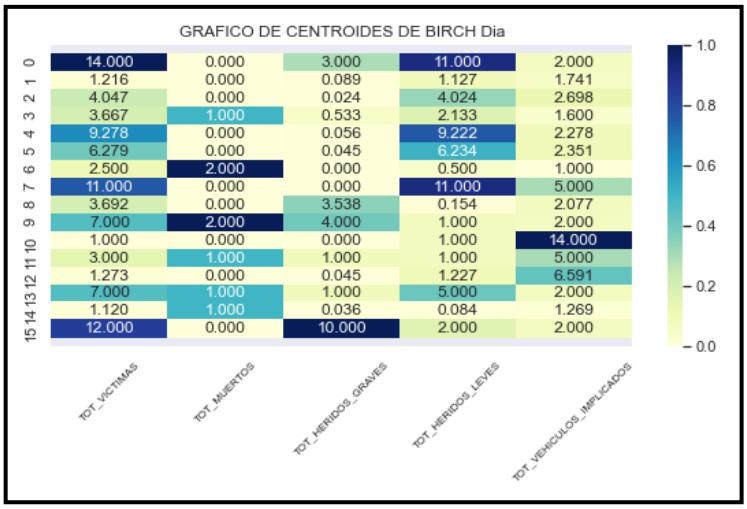




Kmeans vs Birch

Para realizar la comparativa usaremos la clase DÍA, debido a que es nuestro caso de estudio principal.





Encontramos el mismo patrón que para el anterior punto. Por ejemplo, en Kmeans el clúster 13 es similar al clúster 7 en Birch (un índice elevado de víctimas implica heridos graves), el clúster 3 en Kmeans y el 4 en Birch (altos porcentajes en heridos leves dan como resultado alto porcentaje de víctimas) y para los clústeres 1 (índice alto en heridos leves, nos da un índice en víctimas). De forma numérica, vemos como el algoritmo Kmeans, vuelve a colocarse como el "mejor" algoritmo para abordar este problema.

Conclusión para el Caso 2

Tras evaluar en varios clústeres y ver las medidas y balanceos podemos decir que en general para este caso abordado no hay atributos en común que determinen las clases. Para el algoritmo Kmeans el atributo más relevante ha sido el de Heridos leves y para el algoritmo Birch, ha sido el número de víctimas.

Podemos observar cómo en ambas clases, a pesar de que es una conducción por zona urbana, cuando ocurre un accidente durante el día, este provoca un alto número de heridos leves, y que durante la noche aumenta el número de víctimas. No obstante, hay menos datos para la noche que para el día, debido a que el número de desplazamientos durante el día es mayor a que durante la noche.

Basándonos en las medidas obtenidas, vemos como la clase DÍA es con la que mejor resultados hemos conseguido en ambos algoritmos. Para el caso de Kmeans se llega a rozar el 90% de similitud y para el Birch, a excepción del último caso, se ronda un máximo del 70% de similitud. Cabe señalar que, a pesar de los bajos porcentajes para el algoritmo Birch, en el primer número de clúster es con el que obtenemos una mejor medida de Silhuoete.

Otro estudio clave, es el balanceo de los clústeres, en los que Birch acumula en un solo clúster la totalidad de los datos de estudios, lo que nos lleva a la conclusión de

que las características de esos casos abarcan la totalidad y por ello realiza un mal balanceo. Para el caso de la clase DÍA en la que tenemos 37.000 datos, este es el balanceo de la clase Birch:

1: 36191, 2: 675, 14: 167, 5: 111, 12: 22, 4: 18, 3: 15, 8: 13, 0: 2, 6: 2, 13: 1, 9: 1, 10: 1, 15: 1, 11: 1, 7: 1

Como observamos, en el clúster número 1 se lleva prácticamente el 95% de los datos, y el resto, 15 clústeres, se tienen que repartir los 1.000 datos restantes. Esta es una clara diferencia entre los algoritmos, y si nos fijamos ahora en el de Kmeans:

1: 16426, 2: 9255, 8: 3480, 5: 1502, 7: 1222, 11: 1107, 9: 934, 0: 900, 15: 606, 12: 393, 14: 368, 6: 323, 3: 258, 10: 241, 4: 186, 13: 21

Vemos como tiene una mejor distribución y por eso es por lo que obtenemos unas mejores mediciones y unos mejores mapas de calor y por lo que podríamos decir que, tras probar sólo con dos algoritmos, que usar Kmeans para este problema podría ser muy beneficioso.

Conclusión final

Para finalizar el informe sobre clustering, voy a realizar una visión global destacando los aspectos más importantes. En mi caso de estudio he realizado las medidas para los clústeres basándome en la medida WCSS, pero hay que comentar que si hubiese añadido a Kmeans un caso de estudio con un valor K=50, el valor de Silhuoete sería muy cercano a 1, esto es debido a que al aumentar para Kmeans el número de clústeres se aumenta también su efectividad a la hora de la clasificación. Haber realizado este tipo de mediciones hubiese sido perjudicial para el estudio, puesto que al no tener una cantidad muy grande de datos y conseguir clasificar los datos en clústeres correctos, hubiesen quedado muchos de ellos vacíos y prácticamente inutilizando el uso de tantos clústeres.

Para el algoritmo Birch, ha quedado demostrado que tanto en las medidas de evaluación obtenidas como a la hora de realizar el HeatMap y visualizar el balanceo, no ha sido un algoritmo que se pueda realizar para estudios desde la DGT. Este algoritmo ha tenido el gran inconveniente de que siempre ha asignado una totalidad de datos a un solo clúster y eso le ha perjudicado para el resto.

Por tanto, podemos concluir que el uso de Kmeans para abordar este problema ha sido probablemente la mejor opción, ya que es un algoritmo que nos ofrece un buen balanceo de clases y atributos al igual que un gran valor Silhuoete y Calinsky.

Material adicional

Voy a visualizar de forma adicional algunas gráficas en la que podemos ver distintas clasificaciones de los clústeres obtenidos.

Para el caso uno voy a representar el gráfico de Scatter_Matrix

- Lluvia:

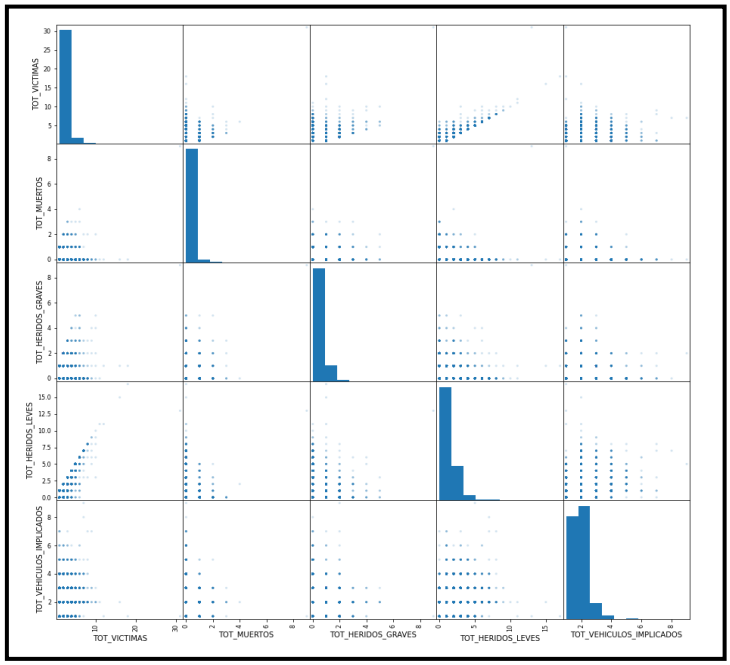


Gráfico Scatter_Matrix de la clase Lluvia

- Sol:

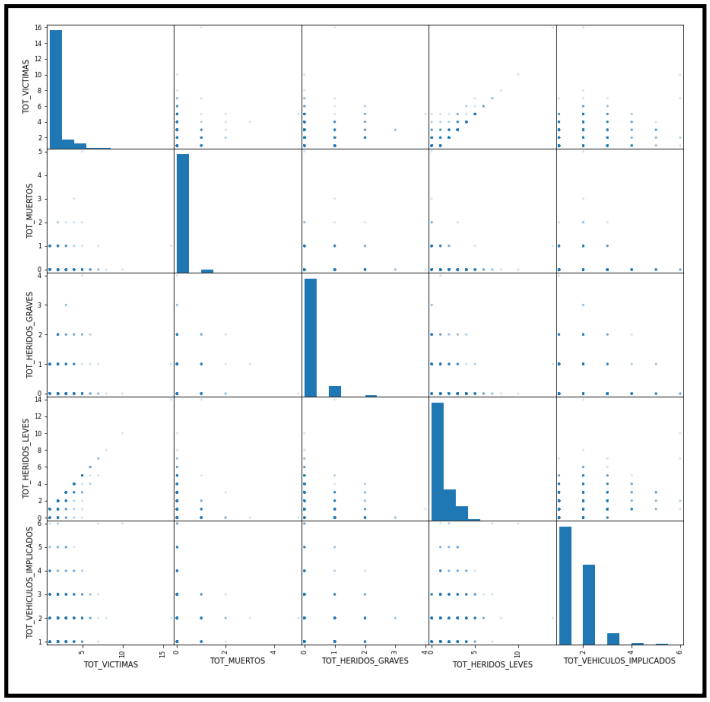
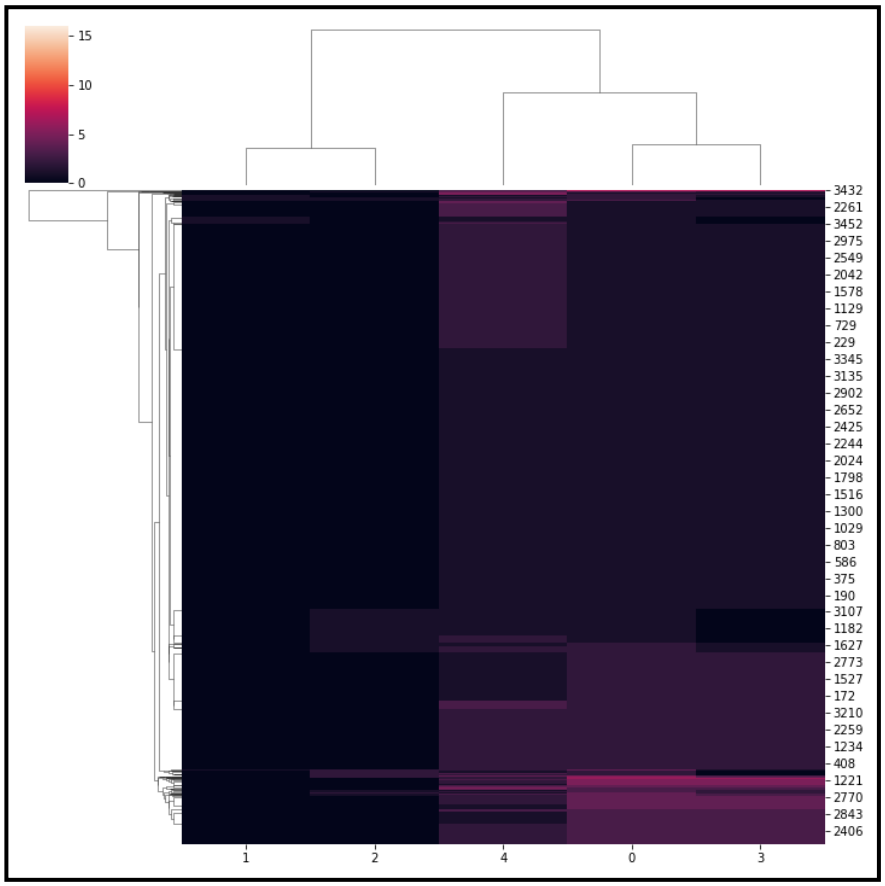


Gráfico Scatter_Matrix de la clase Sol

Se puede observar, como los atributos Tot_Vehículos_implicados y Heridos_Leves son los que más relevancia tienen. Esto ya se había mencionado en los estudios realizados durante el informe.

Finalmente, visualizaré un gráfico dendograma para ver la incidencia en la clase de los atributos determinantes. Los valores son 0: Víctimas, 1: Muertos, 2: Graves, 3: Leves, 4: Vehículos.

- Caso 1 clase Lluvia:



Dendograma para la clase LLUVIA

Vemos como los colores más claros corresponden a los atributos Vehículos implicados y víctimas.