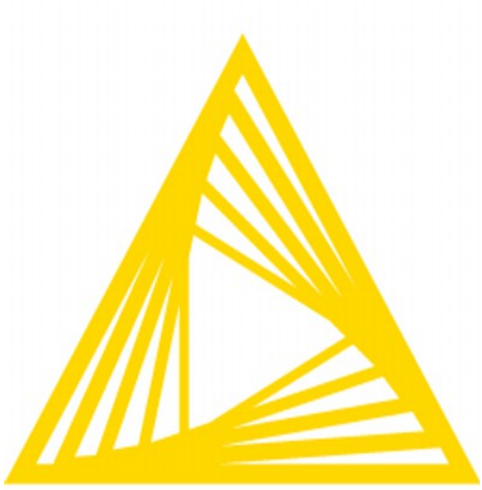
UNIVERSIDAD DE GRANADA
E.T.S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la
Computación e Inteligencia Artificial

Inteligencia de Negocio



PRÁCTICA 3
SEGMENTACIÓN Y ASOCIACIÓN PARA ANÁLISIS EMPRESARIAL

Christian Andrades Molina

ÍNDICE

1. Descripción del problema: accidentes mortales de tráfico.

Una compañía aseguradora trata de comprender la dinámica de los accidentes de tráfico con víctimas mortales. A través de una base de datos con más de 30 variables que comprenden accidentes entre 2008 y 2013, se trata de buscar relaciones de causalidad que expliquen los casos de forma general. Tenemos 11.009 accidentes mortales entre los 522.576 que componen la base de datos.

Las variables que componen la base de datos son las siguientes:

- ID_ACCIDENTE: identificador del accidente. **No es relevante para el análisis.**
- **AÑO:** entre 2008 y 2013.
- **MES:** Entre Enero y Diciembre
- **HORA:** la hora del accidente
- **DIA DE LA SEMANA:** entre Lunes y Domingo
- **PROVINCIA:** cualquier provincia perteneciente a España
- **COMUNIDAD:** cualquier comunidad perteneciente a España
- ISLA: Si estamos en un accidente ocurrido o no en una isla. **No es relevante para el análisis.**
- MUNICIPIO: código postal del municipio. **No es relevante para el análisis.**
- **TOTAL VICTIMAS, MUERTOS, HERIDOS GRAVES, HERIDOS LEVES Y VEHICULOS IMPLICADOS:** número registrado de implicados, tanto personas como vehículos y su gravedad.
- **ZONA:** zona urbana o carretera. **Solo usaremos Zona para el análisis**
- ZONA AGRUPADA: vías urbanas o interurbanas. **No es relevante para el análisis.**
- **CARRETERA:** la carretera donde se produjo el accidente.
- **RED_CARRETERA:** titularidad autonómica, estatal, provincial, municipal u otras.
- **TIPO VIA:** via convencional, autovía, autopista, etc.
- **TRAZADO:** curva suave, recta, etc.
- **TIPO INTERSECCION:** no es intersección, enlace de entrada, etc.
- **CALZADA ACONDICIONADA:** nada especial, otro tipo, carril central de espera, etc. **No es relevante para el análisis.**
- **PRIORIDAD:** señal de stop, ninguna, semáforo, etc.
- **SUPERFICIE CALZADA:** seca y limpia, mojada, umbría, barrillo, etc.
- **LUMINOSIDAD:** pleno día, crepúsculo, noche, etc.
- **FACTORES ATMOSFERICOS:** buen tiempo, viento fuerte, lloviznando, niebla intensa, lluvia fuerte, etc.
- **VISIBILIDAD:** vegetación, edificios, polvo o humo, deslumbramiento, etc.
- **OTRA CIRCUNSTANCIA:** fuerte descenso, obras, baden, escalón, baches, etc.

- **ACERAS:** si hay o no hay.
- **TIPO ACCIDENTES:** todas las posibles variables que existen en un accidente.
- **DENSIDAD CIRCULACIÓN:** fluida, densa o congestionada.
- **MEDIDAS ESPECIALES:** ninguna medida, otra medida, carril reversible o habilitación arcén.

2. Técnicas de Agrupamiento y Reglas de asociación

La práctica a realizar se lleva a cabo sobre técnicas de aprendizaje no supervisado: agrupamiento y reglas de asociación, con el fin de encontrar asociaciones en los datos.

Los métodos utilizados en Agrupamiento se basan en particionamiento y jerarquización: k-means y jerárquico aglomerativo. Los métodos basados en particionamiento se basan en la partición de la base de datos formada por objetos en un conjunto de clusters a elección. En este caso, el método k-means representa por instancias que se van moviendo entre clusters hasta que alcanza el conjunto de clusters seleccionado

Por otro lado, tenemos el jerárquico aglomerativo, que se basa en medir la distancia entre cluster partiendo de un cluster por cada objeto de la BD y fusionando en cada paso los dos más cercanos.

3. Agrupamiento:

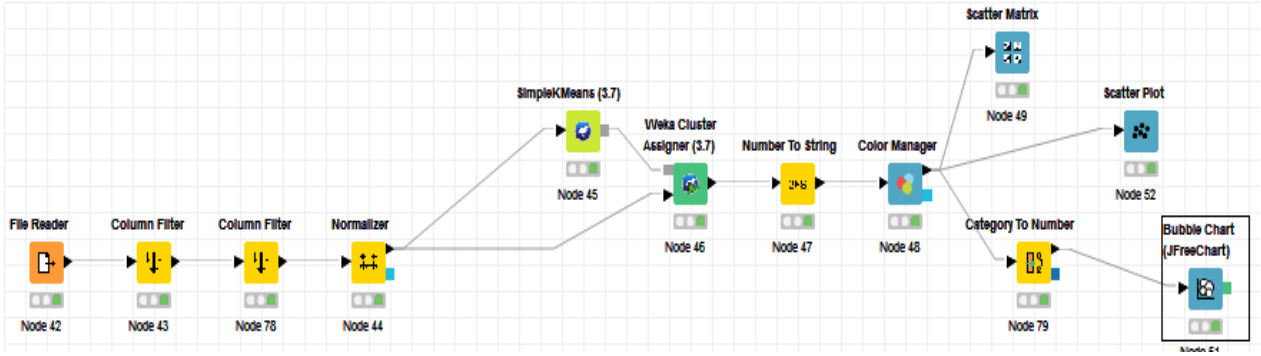
3.1 K-means

Para el análisis de los datos y su correcta visualización, los trataremos en primer lugar. En este caso, mediante un nodo *Column Filter*, eliminaremos varias variables que se eliminaron en clase de prácticas por ser irrelevantes. Estas son: id_accidente, isla, municipio, zona_agrupada, carretera y acond_calzada.

Probaremos con valores en el intervalo [2,5] el número de clusters (k) a elegir, analizando cada uno de ellos con distintas gráficas, seleccionando las variables más oportunas para una correcta visualización y determinando que valor de k es más recomendable.

Para este análisis, en primer lugar realizaremos un segundo filtrado de variables que bajo mi punto de vista son las más necesarias para sacar conclusiones y después aplicar una normalización.

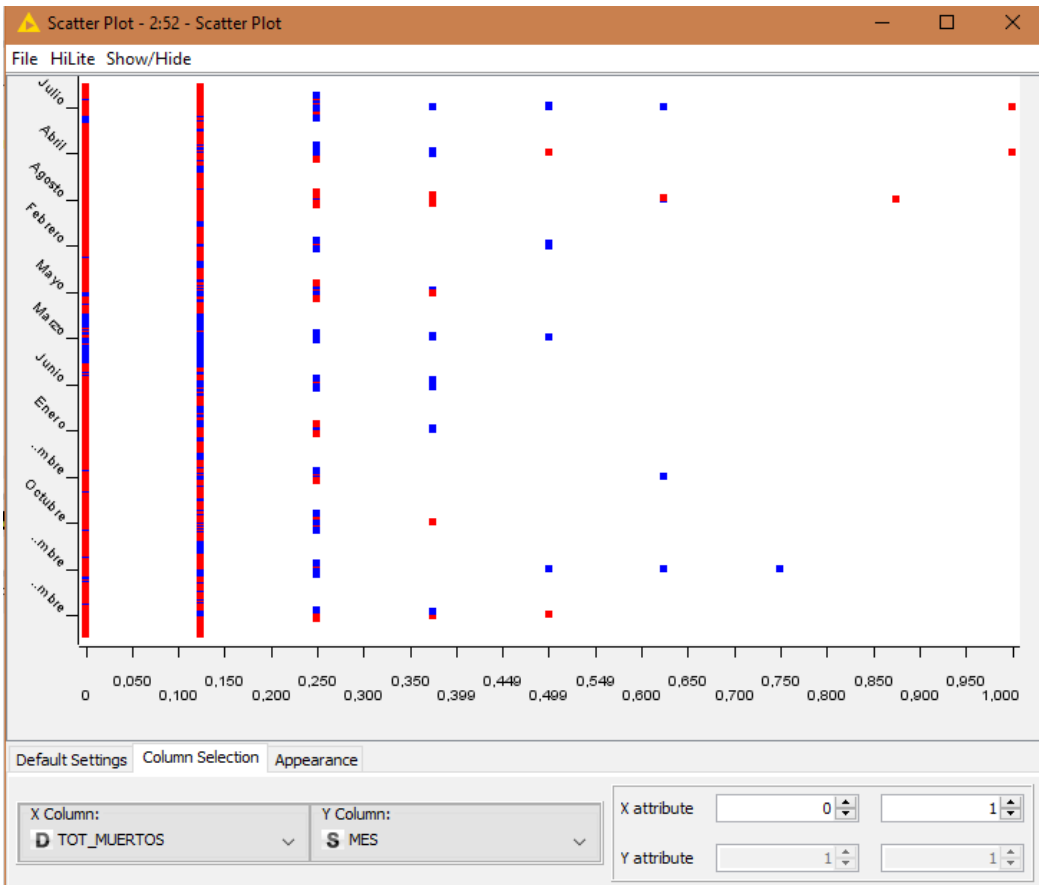
El diagrama de Knime para este primer caso es el siguiente:



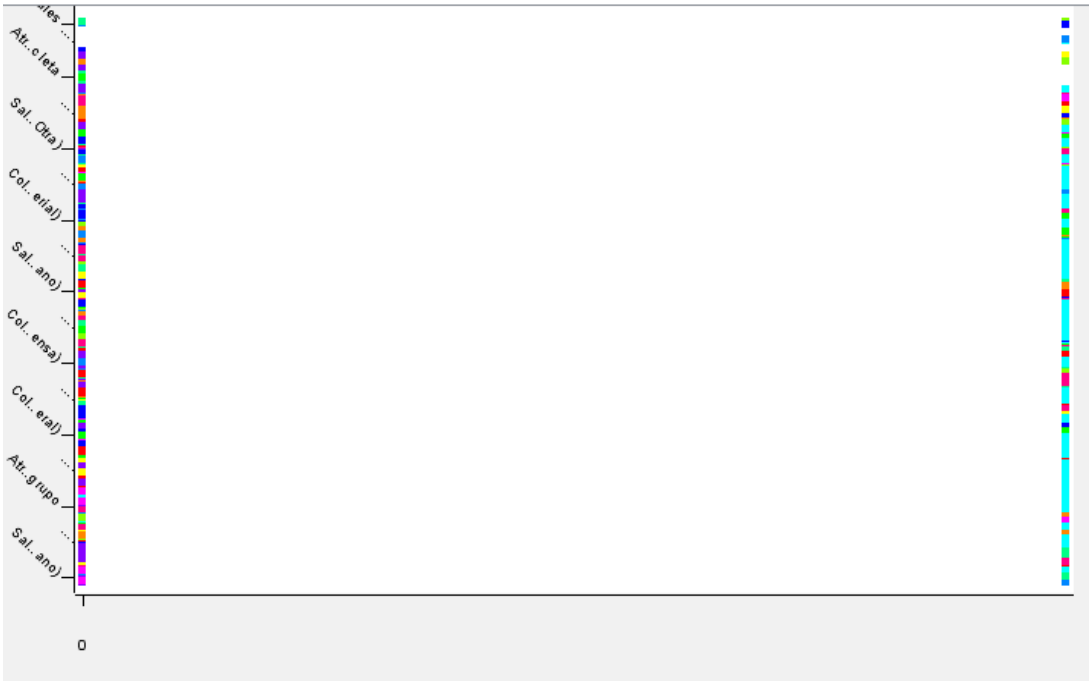
Mediante el nodo File Reader leemos la base de datos ofrecida. El primer Column Filter elimina las variables indicadas en clase y en el 2nd, variables que descarto por no ser importantes para generar resultados interpretables: hora, dia_semana, provincia, heridos graves, heridos leves, red_carretera, tipo_via, trazado no intersec, etc.. Tras ello, normalizamos las variables víctimas, muertos y vehículos implicados.

Clusters: 2

Para dos clusters, analizaremos el total de muertos en cada mes. Los dos clusters se diferencian mediante el color rojo y azul:



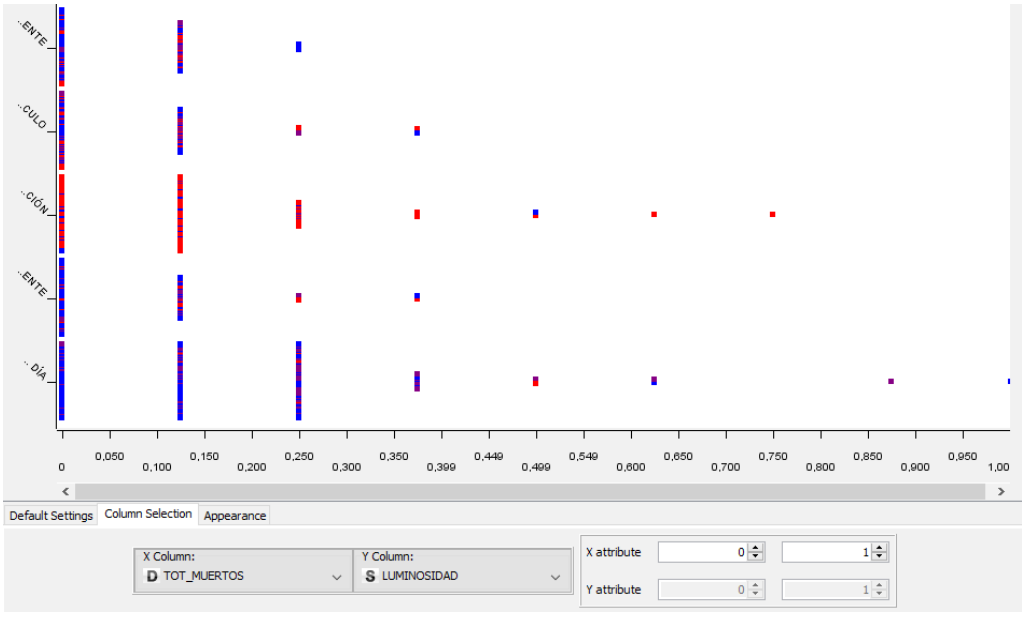
Podemos ver como en la gráfica Scatter Plot, existe una diferencia entre el cluster 1 y 2 en número inferiores de muertes y mayores, mientras que encontramos una barrera difusa en entre 1 y 1,5 donde el número de muertes se solapa en ambos clusters.



Clasificando por accidentes (y en color por meses), se clasifican de forma evidente en ambos clusters, teniendo en uno la mayoría de meses y en el otro Agosto por mayoría y otros como Octubre y Marzo.

- Clusters: 3

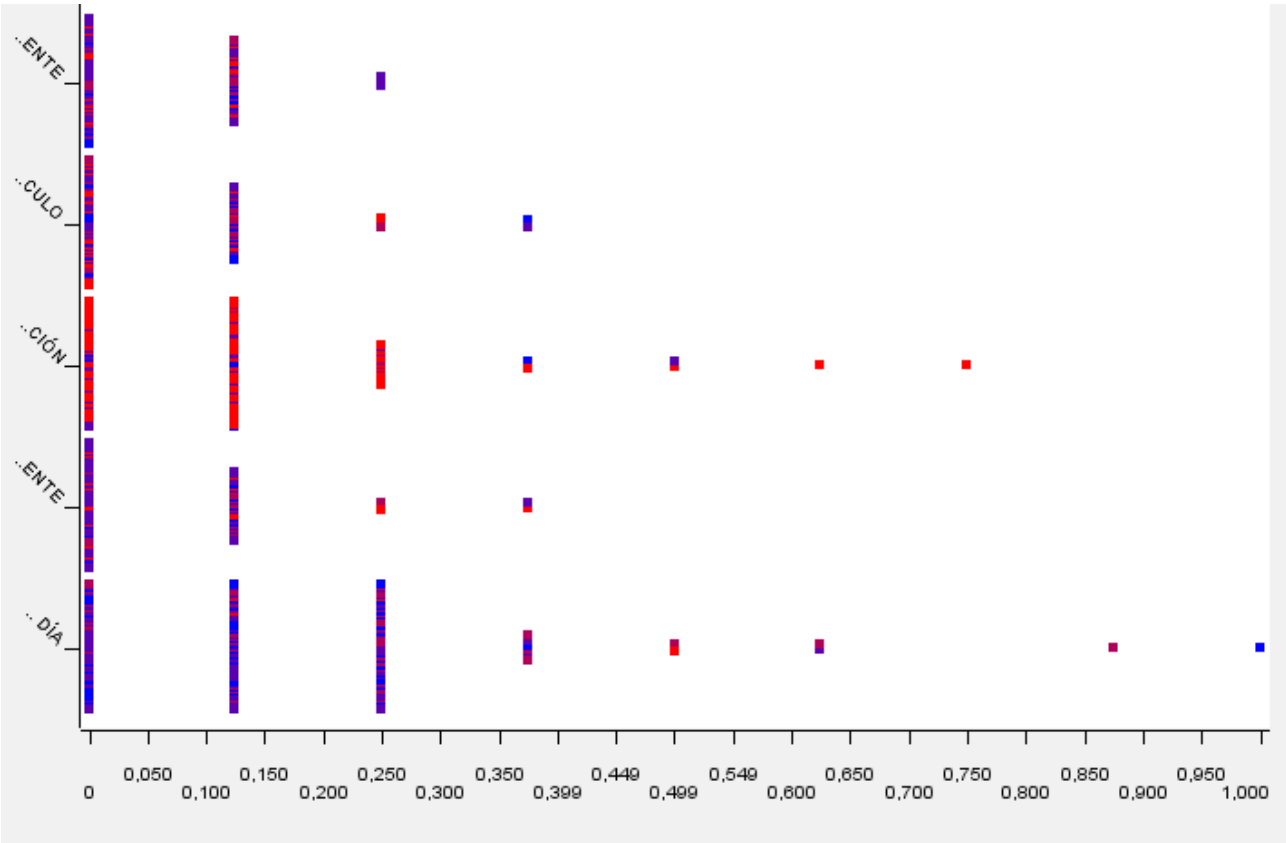
Para tres clusters, analizaremos el total de muertos según la luminosidad. Los tres clusters se diferencian entre el color rojo y azul:



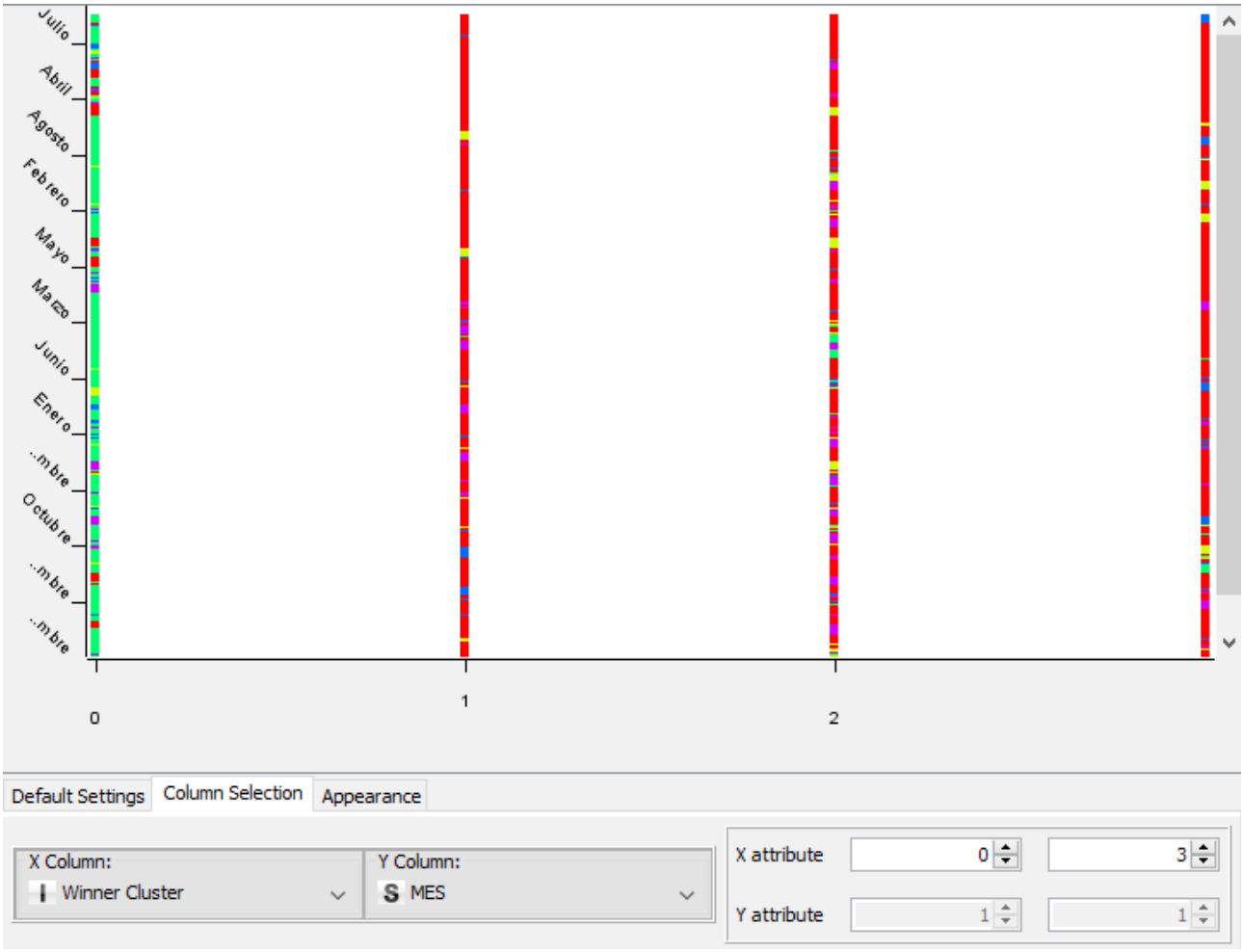
En este caso, las muertes producidas sobre “Noche: sin iluminación” se encuentran en el primer cluster, mientras que el resto de situaciones las encontramos en los restantes. Los datos que nos ofrecen son evidentes, ya que vemos un aumento significativo de muertes durante el día, siendo las horas con más tráfico y riesgo frente a la madrugada.

- Clusters: 4

Para cuatro clusters, analizaremos el total de muertos según la luminosidad. Los cuatro clusters se diferencian entre el color rojo y azul:



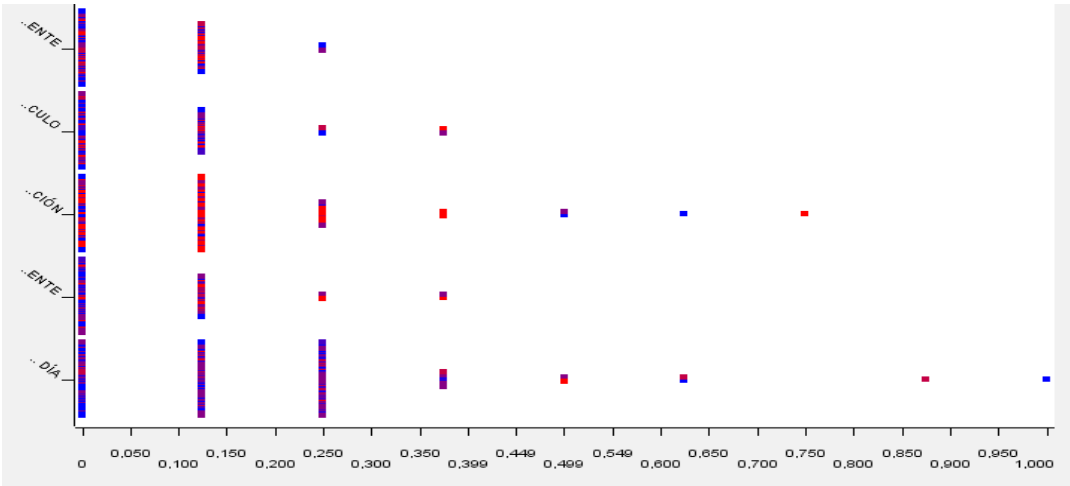
La diferencia entre 3 clusters y 4 es prácticamente inapreciable. En el primer cluster se siguen observando los casos de noche y en el resto los restantes. No hay ninguna razón para diferenciarlos



Para esta gráfica, los clusters se clasifican de forma muy precisa para el caso de Noche sin luminación, encontrandose la mayor parte en el cluster 1.

- Clusters: 5

Para cinco clusters, analizaremos el total de muertos según la luminosidad. Los cinco clusters se diferencian entre el color rojo y azul:

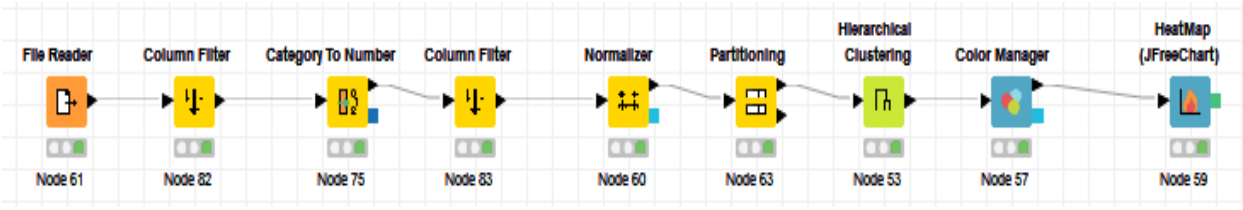


La diferencia entre los clusters anteriores sigue siendo inapreciable, aunque nos podemos atener a una diferencia más difusa en noches sin iluminación

Teniendo en cuenta las gráficas mostradas, siendo las únicas donde conseguía ver una diferencia más precisa en la clasificación de casos en los clusters correspondientes (mediante la iluminación), determinando k=3, se consigue unas fronteras menos difusas y facilitan mejor el análisis.

3.2 Jerárquico aglomerativo:

Por otro lado, tenemos el jerarquico aglomerativo. Para este análisis hemos utilizado esta secuencia:

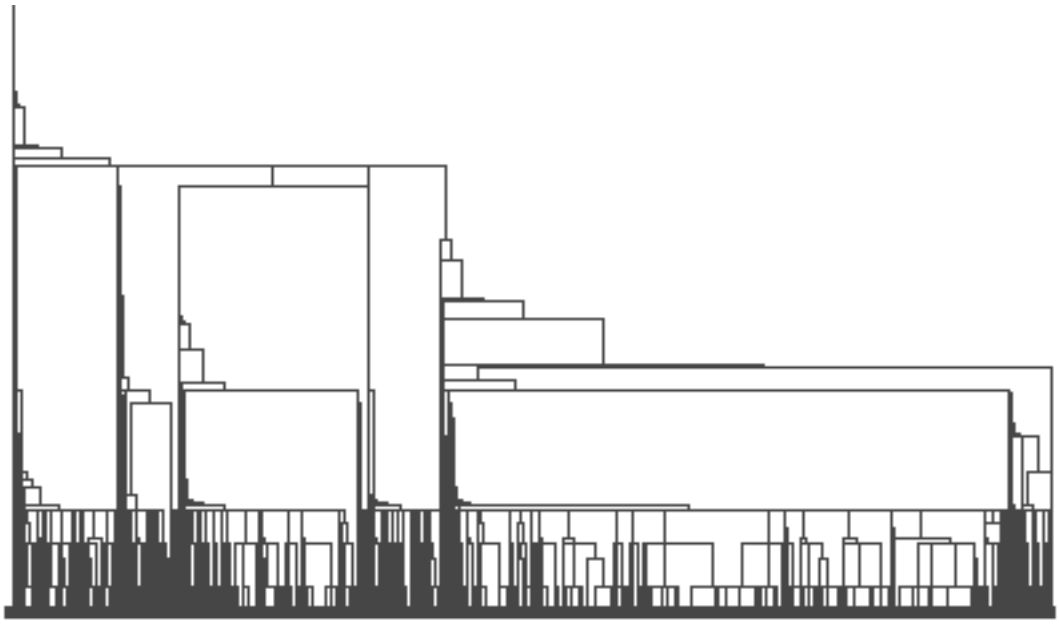


Mediante el nodo File Reader hemos leído los datos ofrecidos en la práctica sobre los accidentes mortales. En el nodo Column Filter, filtramos el número de variables, quedándonos con 6 para reducir el tiempo de ejecución del algoritmo 'Jerárquico Aglomerativo' (muy alto si está condicionado por muchas opciones) y conseguir unos resultados más interpretables.

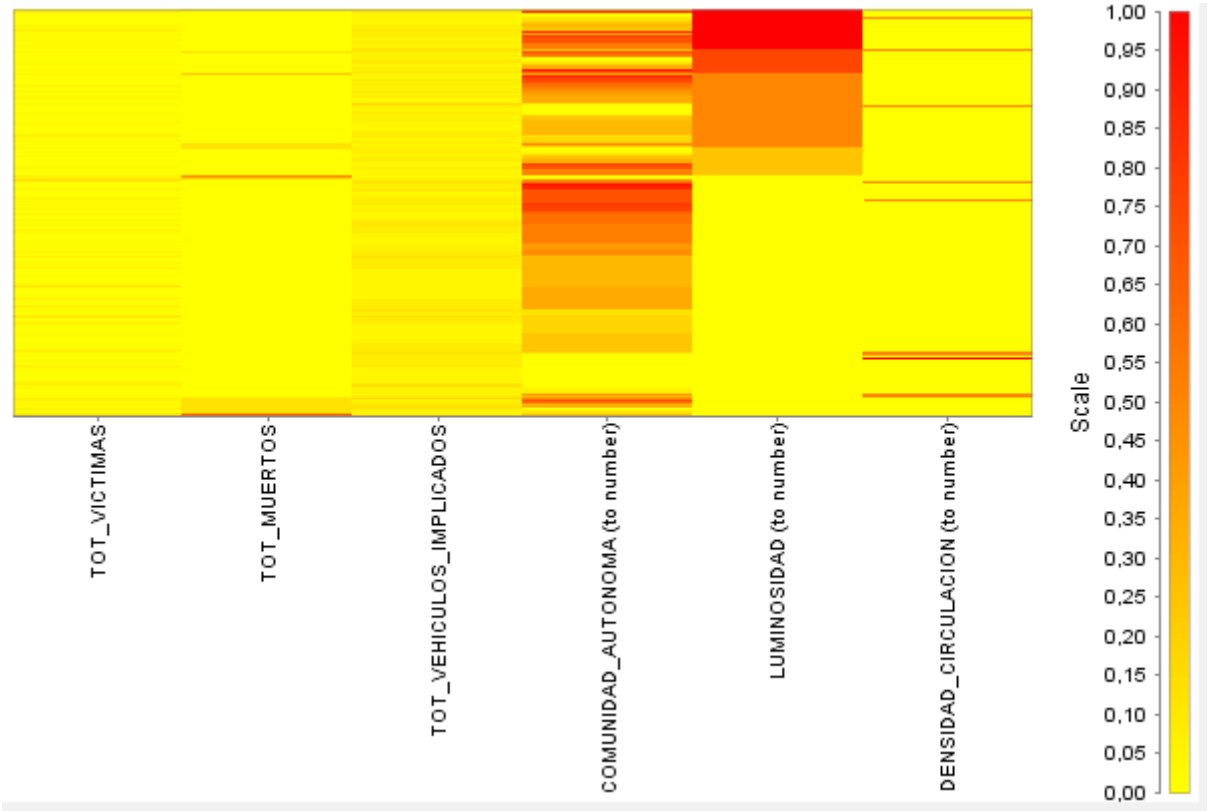
Estas variables son: comunidad_autonoma, total_victimas, total_muertos, total_vehículos_implicados, luminosidad y densidad_circulación. Transformamos las variables de formato category a formato number para que el algoritmo las acepte.

Las variables originales transformadas, las eliminamos con otro nodo Column Filter y las restantes las normalizamos para mejorar la visualizacion del resultado. Hemos particionado al 60% ya que un porcentaje mayor bloqueaba el ordenador donde ejecutamos el algoritmo, siendo imposible avanzar. El dendograma resultante es el siguiente:

Nos apoyaremos en la visualización tanto del dendograma derivado del agrupamiento como mapas de temperatura (heat map).



Mediante el siguiente gráfico podemos visualizar las distintas relaciones de agrupación que hay entre los datos y la existente entre los propios grupos que se desarrollan. La clasificación en los grupos finales es imposible de apreciar debido a la cantidad de desglosamientos pero si en los primeros, obteniendo un grupo que se divide en dos en cada fila hasta clasificarse hasta en 4.

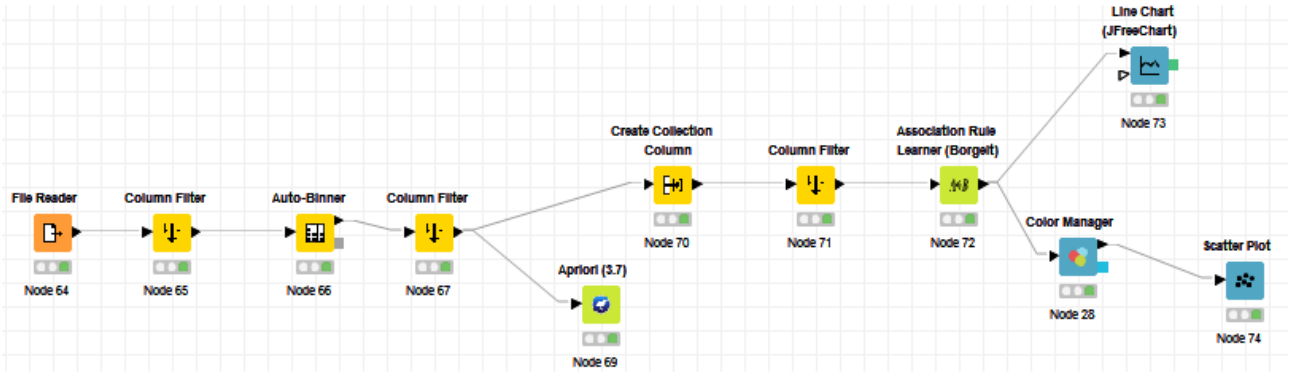


La gráfica siguiente es el mapa de calor. En ella podemos ver las 6 variables:

tot_muertos, tot_victias, tot_vehículos_implicados, comunidad_autonoma, luminosidad y densidad_circulación. Entre todas, la más significativa es comunidad autonoma, seguida de la luminosidad. La densidad de circulación y el total muertos muestran una discreta relación, con un mapa de calor similar, siendo total vehículos implicados, la variable más difusa con respecto a su representación.

4. Reglas de asociación.

En este segundo apartado, aplicar esta técnica de aprendizaje no supervisado (mediante reglas de asociación), implica este diagrama en Knime:



En primer lugar, mediante el nodo File Reader leeremos el archivo con los datos a analizar con este algoritmo. En Column Filter eliminamos nuevamente las variables descartadas por no ser útiles: id_accidente, isla, municipio, zona_agrupada, carretera y acond_calzada. La discretización de las variables continuas se llevará a cabo con el nodo Auto-Binner, en las que incluiremos: año, hora, victimas, muertos, heridos_graves, heridos_leves y vehículos implicados. Una vez generadas las nuevas variables discretizadas, eliminamos con un Column Filter las originales.

Dividimos el trabajo en dos ramas: para el algoritmo Apriori y para el algoritmo Association Rule Leaner. Para el primero su ejecución es trivial y para el segundo, empezamos con el nodo Create Collection Column donde creamos una columna con las variables a elegir en las reglas que se van a generar. Con un Column Filter eliminamos las variables restantes quedándonos con AggregatedValues. Por último visualizamos lo pedido con las gráficas “Line Chart” y “Scatter Plot”.

Una vez explicado el procedimiento, analizamos las reglas generadas por Apriori y Association Rule Leaner. En el primero hemos limitado a 20 reglas, de las que seleccionamos ...

8. FACTORES_ATMOSFERICOS=BUENTIEMPO TOT_MUERTOSBinned=val1 8609 ==> SUPERFICIE_CALZADA=SECAYLIMPIA 8120 <conf:(0.94)> lift:(1.13) lev:(0.09) [937] conv:(2.91)
12. TOT_HERIDOS_GRAVESBinned=val2 8365 ==> DENSIDAD_CIRCULACION=FLUIDA 7753 <conf:(0.93)> lift:(1) lev:(0) [-6] conv:(0.99)
14. SUPERFICIE_CALZADA=SECAYLIMPIA TOT_MUERTOSBinned=val1 8472 ==> DENSIDAD_CIRCULACION=FLUIDA 7841 <conf:(0.93)> lift:(1) lev:(0) [-17] conv:(0.97)

15. FACTORES_ATMOSFERICOS=BUENTIEMPO DENSIDAD_CIRCULACION=FLUIDA 8632 ==> TOT_MUERTOSBinned=val1 7984 <conf:(0.92)> lift:(1) lev:(0) [35] conv:(1.05)
16. SUPERFICIE_CALZADA=SECAYLIMPIA DENSIDAD_CIRCULACION=FLUIDA 8486 ==> TOT_MUERTOSBinned=val1 7841 <conf:(0.92)> lift:(1) lev:(0) [27] conv:(1.04)
19. FACTORES_ATMOSFERICOS=BUENTIEMPO 9326 ==> TOT_MUERTOSBinned=val1 8609 <conf:(0.92)> lift:(1) lev:(0) [21] conv:(1.03)
20. SUPERFICIE_CALZADA=SECAYLIMPIA 9185 ==> TOT_MUERTOSBinned=val1 8472 <conf:(0.92)> lift:(1) lev:(0) [14] conv:(1.02)

En la regla número 8, si el accidente ocurrió con buen tiempo y el número de muertos ronda entre 1 y 8, la calzada estaba seca. En la 16, si la calzada está seca y la circulación es fluida, el número de muertos está alrededor de 1 a 7.

Para el algoritmo Association Rule Leaner tenemos:

Association Rules - 3:72 - Association Rule Learner (Borgelt)

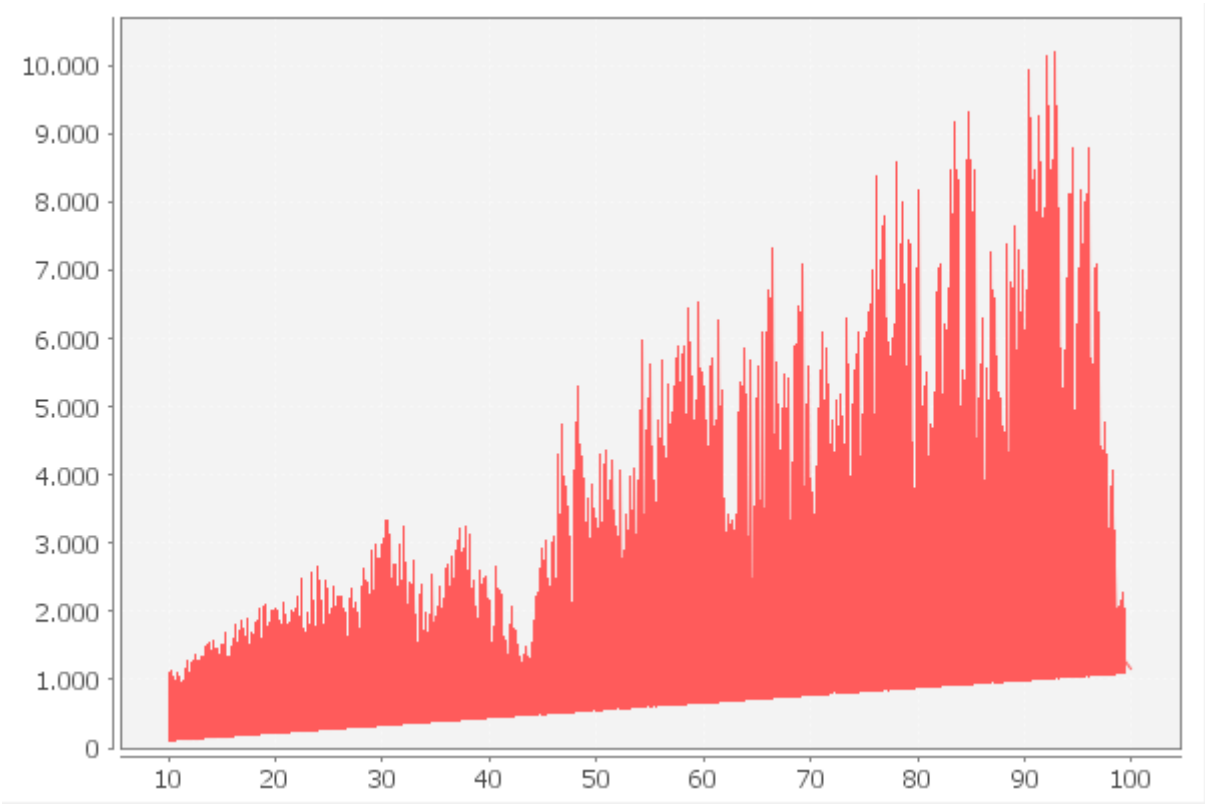
File

Table "default" - Rows: 130389 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	Consequent	Antecedent	ItemSe...	Relativ...	R
Row0	Salamanca	[Castilla y León,FLUIDA]	117	1.063	10.6
Row1	Salamanca	[Castilla y León]	122	1.108	10.8
Row2	Salida de la vía por la derecha sin colisión (En llano)	[Cataluña,CARRETERA]	122	1.108	10.7
Row3	Valladolid	[Castilla y León,FLUIDA]	150	1.363	13.6
Row4	Valladolid	[Castilla y León]	155	1.408	13.7
Row5	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),CARRETERA,...]	118	1.072	10.7
Row6	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),CARRETERA,...]	128	1.163	10.7
Row7	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),CARRETERA]	130	1.181	10.6
Row8	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),SECA Y LIMPIA...	116	1.054	10
Row9	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),BUEN TIEMPO,...]	122	1.108	10.4
Row10	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),BUEN TIEMPO]	124	1.126	10.2
Row11	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[0,0],...]	125	1.135	11.1
Row12	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[0,0],...]	127	1.154	11
Row13	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[0,0],...]	135	1.226	10.9
Row14	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[0,0]]	137	1.244	10.8
Row15	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[1,1],...]	138	1.254	11.1
Row16	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),[1,1]]	140	1.272	10.9
Row17	Córdoba	[Andalucía,NINGUNA (SOLO NORMA),FLUIDA]	152	1.381	11
Row18	Córdoba	[Andalucía,NINGUNA (SOLO NORMA)]	154	1.399	10.9
Row19	Córdoba	[Andalucía,[0,0],[1,1],...]	150	1.363	10.4
Row20	Córdoba	[Andalucía,[0,0],[1,1]]	152	1.381	10.2
Row21	Córdoba	[Andalucía,[0,0],FLUIDA]	163	1.481	10.4
Row22	Córdoba	[Andalucía,[0,0]]	165	1.499	10.2
Row23	Córdoba	[Andalucía,[1,1],FLUIDA]	166	1.508	10.5
Row24	Córdoba	[Andalucía,[1,1]]	168	1.526	10.3

Las gráficas utilizadas para representar el conocimiento resultante son el diagrama de líneas (Line Chart) donde mostramos el numero de reglas obtenidas en función del valor de soporte ...

... y otra gráfica (nube de puntos) que muestra el valor de soporte y confianza de cada regla:


10.000
9.000
8.000
7.000
6.000
5.000
4.000
3.000
2.000
1.000
0
10
20
30
40
50
60
70
80
90
100