Universidad de los Andes

Inteligencia de Negocios

Proyecto 1

Patrones de accidentes de vehículos automotores

Edgar Andrés Margffoy - 201412566

Camila García - 201326493

Índice

1. Selección de los datos a trabajar justificado en términos del problema de negocio que va a resolver y la cantidad de datos disponible…….3

2. Descripción y preparación de los datos usados en los modelos.……4

3. El resultado del análisis de calidad de los datos utilizados en los modelos propuestos, acompañado de un análisis del mismo……...4

4. Descripción de como el requerimiento de negocio es resuelto con el o los requerimientos de minería de datos propuestos, para lo cual debe utilizar la tabla que se presenta a continuación…… 5

5. El resultado de mínimo 2 modelos realizados, que de forma conjunta permitan resolver el requerimiento de negocio……... 5

6. Análisis del resultado de los modelos……....7

7. Las estrategias que la organización debe plantear relacionadas con los resultados obtenidos en los modelos y una justificación de porqué esa información es útil para ellos……10

1. Selección de los datos a trabajar justificado en términos del problema de negocio que va a resolver y la cantidad de datos disponible.

El archivo seleccionado de la página www.datos.gov.co es el Registro nacional de accidented de tránsito. Este archivo consta de más de 170000 registros que incluyen información amplia sobre el tipo de accidente, sus observaciones de causantes, fecha, nivel de riesgo (daños, heridos, muertos), hora, municipio y departamento. Con esta información se pretende encontrar los patrones que permiten agrupar los accidentes en diferentes niveles de riesgo, con el fin de crear políticas, leyes y organizaciones viales que puedan ayudar a prevenirlos y a reaccionar de manera rápida cuando ocurran. En un principio, se desea trabajar solamente sobre un departamento, puesto que la información es muy amplia y su procesamiento podría tomar un largo tiempo, y la implementación de leyes y políticas es más rápida cuando se hace a nivel departamental. La institución encargada de esta implementación es el Ministerio de Transporte. El objetivo de esta entidad, tomado de su página:

> *Objetivo. El Ministerio de Transporte tiene como objetivo primordial la formulación y adopción de las políticas, planes, programas, proyectos y regulación económica en materia de transporte, tránsito e infraestructura de los modos de transporte carretero, marítimo, fluvial, férreo y aéreo y la regulación técnica en materia de transporte y tránsito de los modos carretero, marítimo, fluvial y férreo.*[1]

Asi pues, la fuente asociada al sector se buscó en https://www.mintransporte.gov.co. Información compilada de las estadísticas publicadas en esta página y recopiladas en un informe de Fededesarrollo[2] indican que más del 70% de las personas en las ciudades utilizan vehículos automotores para transportarse en su día a día. Por esto, es primordial que se trate de disminuir los accidentes graves en los que se podrían ver involucrados.

---

[1]Ministerio de Transporte, *Mintransporte,* Consultado Marzo 7 de 2017 https://www.mintransporte.gov.co/Publicaciones/Ministerio/quienes_somos

[2]Fedesarrollo, *Fedesarrollo ,* Consultado Marzo 7 de 2017 http://www.fedesarrollo.org.co/wp-content/uploads/2011/08/Indicadores-del-sector-transporte-en-Colombia-Informe-Consolidado.pdf

2. Descripción y preparación de los datos usados en los modelos.

El archivo original se trata de un archivo CSV que posee 170854 registros. Estos registros tienen información de tipo de accidente, sus observaciones de causantes, fecha, nivel de riesgo (daños, heridos, muertos), hora, municipio y departamento. En el Anexo 1 es posible ver la distribución original de los datos en WEKA. Esta nos ayudó a identificar datos con demasiados valores, que por lo tanto no fueran útiles, y a entender mejor los datos sujetos a análisis. Municipio, por ejemplo, fue eliminado, pues aunque pareciera interesante, el número de diferentes municipios era tan grande que su impacto podría ser negativo en el modelo.

Fue necesario modificar el atributo de hora, puesto que la hora exacta no es útil para encontrar patrones. Así pues, se crearon 4 categorías de hora: mañana, tarde, noche y madrugada. Las horas se clasificaron en estas 4 categorías, de manera que la columna pueda ser utilizada en la identificación de patrones. Además, se descartaron las columnas que tuvieran que ver con identificadores, pues no son útiles para el objetivo que se persigue, se descartó la columna de municipios, como se mencionaba anteriormente, y se descartó la columna de fecha, puesto que se busca encontrar patrones que se puedan aplicar de una manera continua y no solo en ciertas épocas del año. Las distribuciones resultantes se pueden ver en el Anexo 2. Aunque las observaciones también tienen muchos valores, pensamos que pueden ser útiles pues hablan sobre la razón de cada accidente, y su contenido es normalizado.

3. El resultado del análisis de calidad de los datos utilizados en los modelos propuestos, acompañado de un análisis del mismo.

Con el fin de evaluar la calidad de los datos, se procedió a realizar una visualización de los mismos a través de Weka, en este caso se evaluó la consistencia de la información y la distribución de cada una de las categorías presentes en el conjunto de datos. Debido a que todas las entradas se encuentran regularizadas y no cuentan con anomalías de compilación tal como registros con información incompleta, categorías no consistentes o valores nulos, fue posible realizar la importación de los mismos de forma satisfactoria. En general, los valores presentes en el conjunto de datos presentan consistencia en la forma en la cual se encuentran representados, por ejemplo, las observaciones asociadas a cada evento se encuentran categorizadas en valores presentes en un conjunto delimitado.

El conjunto de datos presenta información para cada uno de los departamentos del país, lo cual elimina posibles sesgos al momento de procesar la información. e.g., Si solo se consideran la información proveniente de regiones donde predominan las áreas rurales, el número de accidentes disminuye.

4. Descripción de como el requerimiento de negocio es resuelto con el o los requerimientos de minería de datos propuestos, para lo cual debe utilizar la tabla que se presenta a continuación.

Oportunidad / Problema de negocio		
Conocer y prevenir las causas de accidentalidad que resultan en heridos e individuos muertos.		
Descripción del requerimiento desde el punto de vista de minería de datos		
Identificar las causas o los patrones que permiten caracterizar a un accidente vehicular que presenta individuos muertos o heridos.		
Detalles de la actividad de minería de datos		
Tarea	Técnica	Algoritmo y parámetros utilizados
Clasificación	Supervisada	Árbol de Decisión (C4.5)
Clustering	No Supervisada	K-Means: K = 8

5. El resultado de mínimo 2 modelos realizados, que de forma conjunta permitan resolver el requerimiento de negocio.

Por un lado, se quiso utilizar un modelo de minería de datos no dirigida con el fin de poder observar el comportamiento de los datos en relación a el nivel de riesgo del accidente. Aunque se trata de un modelo no dirigido, tiene una variable objetivo, y según el libro de clase se le llama modelo “semidirigido”. La idea es, entonces, que se descarten los resultados que no nos ayudan a describir la variable objetivo.

La técnica que se utilizó fue una técnica de Clustering, utilizando el algoritmo de SimpleKMeans ofrecido por Weka. Para encontrar el K ideal se hicieron diferentes pruebas, hasta que se llegó a un resultado que no fuera muy general (por ejemplo, un cluster con choque y heridos y otro con choque y no heridos) pero tampoco muy

específico (gran cantidad de clusters con muchas cosas en común). El número con el que se alcanzó este óptimo fue K=8. Por otro lado, se evitó usar las observaciones pues la gran dominancia de “OTRA” causaba que estuviera en todos los clusters. Esto se logró con la opción de “Ignore attributes” que se ofrece en los modelos de clustering de Weka. Por otro lado, se notó que el modelo podía ser ejecutado con gran velocidad, por lo que se pusieron las máximas iteraciones en 1000, en vez de los 500 por defecto, con el fin de crear un modelo más preciso. El resto de parámetros se dejaron en los incluidos por defecto, pues no se vió necesidad de cambiarlos. En el anexo 3 se puede encontrar una captura de pantalla con los parámetros utilizados.

A continuación, con el fin de caracterizar las reglas que cumplen los eventos y su relación con las consecuencias causadas (choques, lesiones, muertes), se procede a realizar una comprensión del problema desde una perspectiva de aprendizaje supervisado, bajo el cual, las variables de entrada corresponden al área (Rural/Urbana) y la hora en la cual suceden los accidentes mientras que la variable objetivo del problema corresponde a la consecuencia evidenciada como parte del evento, en este caso indicar la presencia de Choques simples, lesiones no graves y decesos. Debido a que el modelo de clústering presenta información valiosa relacionada con la distribución geográfica (Departamentos) de cada uno de los eventos y sus consecuencias, es posible descartar esta información con el fin de concentrar los esfuerzos de análisis entorno a los factores que determinan la ocurrencia de un evento en específico y no la distribución espacial de los mismos.

Ahora bien, debido a que se desean establecer reglas que deben cumplir los datos de entrada para ser etiquetados como causales de choques, lesiones o muertes, se propone el uso de un Árbol de decisión, en la medida que este es un modelo de clasificación expresivo, bajo el cual, es posible extraer y observar las reglas de decisión tomadas en cada bifurcación, a diferencia de otros modelos de clasificación, bajo los cuales solo es posible determinar un porcentaje de efectividad sobre el conjunto de datos a clasificar.

Con el fin de establecer un conjunto de reglas comprensibles, se procede a la evaluación de diversos modelos de árboles de decisión, bajo los cuales se definieron diversos conjuntos de variables de entrada, no obstante, con el fin de emplear las características expresivas del modelo, se empleó el resultado que presenta el menor número de niveles de profundidad, i.e., Menor número de reglas anidadas. Debido a que el porcentaje de eventos que concluyen con muertes son reducidas con respecto a eventos causales de choques y lesiones, es posible apreciar que el modelo no caracteriza este tipo de accidentes y por el contrario pretende caracterizar las causas de ocurrencia de choques y lesiones, los cuales resultan de interés al momento de emitir campañas viales, en la medida que estos

suceden con mayor frecuencia. En el anexo 4 se presenta el conjunto de parámetros que fueron empleados para entrenar el modelo, es necesario observar que el factor de confianza fue incrementado, con el fin de reducir el número de ramas a podar en el modelo y obtener un conjunto de reglas consistentes.

6. Análisis del resultado de los modelos.

Muestra del resultado del clustering:
(Completo se puede encontrar en los logs incluidos en esta entrega)

Cluster#							
0	1	2	3	4	5	6	7
(67310.0)	(24723.0)	(20287.0)	(616.0)	(26766.0)	(10073.0)	(12112.0)	(8967.0)
ANTIOQUIA SOLO DANOS	ANTIOQUIA CON HERIDOS	ANTIOQUIA CON HERIDOS	ANTIOQUIA SOLO DANOS	BOGOTA D. C. SOLO DANOS	VALLE DEL CAUCA SOLO DANOS	ANTIOQUIA CON HERIDOS	VALLE DEL CAUCA CON HERIDOS
CHOQUE	CHOQUE	CHOQUE	OTRO	CHOQUE	CHOQUE	ATROPELLO	VOLCAMIENTO
Tarde	Noche	Mañana	Noche	Mañana	Noche	Tarde	Tarde

```
=== Model and evaluation on training set ===

Clustered Instances

0       67310 ( 39%)
1       24723 ( 14%)
2       20287 ( 12%)
3         616 (  0%)
4       26766 ( 16%)
5       10073 (  6%)
6       12112 (  7%)
7        8967 (  5%)
```

Se puede ver que aunque un cluster(#3) indica que el departamento de Antioquia se relaciona con accidentes donde solo hubo daños, este cluster es extremadamente pequeño, por lo que no es indicativo de un comportamiento general. Esto se confirma por las estadísticas arrojadas por el modelo: este número es cerca del 0% de los datos. Es posible ver, en cambio, que los clusters 0, 1, 2 y 6 relacionan al departamento de Antioquia con accidentes que han resultado en heridos. Estos clusters representan a un gran número de la población, como se evidencia en las estadísticas (39%, 14%, 12%, 7%)

Aunque podría pensarse que el resultado de los departamentos se debe a que estos tienen una gran cantidad de registros, otros departamentos con cantidades similares (como por ejemplo, cundinamarca y santander) no son protagonistas de los clusters. Esto se debe, seguramente, a que su distribución es menor y no presentan patrones claros de nivel de riesgo del accidente.

Es importante notar que el modelo acabó con alguna concepciones que se tenían antes de iniciar el proyecto. Por ejemplo, esperábamos encontrar que la mayoría de accidentes graves se dan en altas horas de la madrugada, en especial porque estos involucran a personas bajo la influencia del alcohol. Sin embargo, encontramos que

hay una mayor distribución de accidentes con heridos en la mañana y en la noche, posiblemente debido a los altos niveles de tráfico que generan estrés en los conductores.

Por otro lado, también se encontraron patrones que no tienen uso para el objetivo de negocio, como por ejemplo, el hecho de que los accidentes de tipo atropello estén de la mano con el nivel de riesgo de heridos. Dado que el atropello involucra un vehículo motorizado y un peatón, es de esperarse que haya heridos en la escena. Otro patrón que no concurre en un uso claro es el de los choques, que aparece en el cluster más grande (#0) junto con el nivel de riesgo de solo daños, pero en otros clusters más pequeños (1, 2, 5) aparece junto al nivel de riesgos de heridos. Con el fin de aclarar esto, se hace uso del árbol de decisión.

Con respecto a los resultados obtenidos tras el entrenamiento del árbol de decisión J48 descrito a lo largo de la sección previa, es posible concluir que la mayoría de condiciones sobre las cuales suceden los eventos de accidentes, tienen como consecuencia, la presencia de heridos en la escena. Sin embargo, otros eventos tal como los choques e incendios cuentan con daños sobre la propiedad privada de cada una de las partes involucradas, este comportamiento es esperado, en la medida que ante la ocurrencia de un choque o un incendio, las personas involucradas en el accidente tienden a abandonar los vehículos en cuestión, sin embargo, esta hipótesis resulta ser incompleta en la medida que el modelo descarta muertes debido a estas causas.

Ahora bien, es posible apreciar que en el caso de los accidentes que suceden debido al volcamiento del vehículo, los factores medioambientales, así como de iluminación influyen en el resultado final. En el presente caso, es posible apreciar que las condiciones de iluminación influyen más cuando el evento sucede en áreas rurales con respecto a las áreas urbanas, esto se debe a que muchas rutas y carreteras que se encuentran en estas zonas cuentan con iluminación limitada, lo cual causa que los accidentes en la madrugada resulten ser de menor gravedad con respecto a los eventos presentados a lo largo del día, en este caso se espera que los vehículos transiten a menores velocidades, lo cual reduce la probabilidad de ocurrencia de un volcamiento de un vehículo. A nivel urbano, se espera que el accidente incremente el número de heridos, debido a la existencia de otros vehículos o peatones en las mallas viales de las ciudades y perímetros urbanos del país.

Adicionalmente, a partir de los resultados presentados, es posible concluir y confirmar hipótesis preconcebidas o intuitivas previas al entrenamiento del modelo, por ejemplo, si un accidente involucra el atropello de otra persona, entonces esta es

herida, en este caso, es posible observar que este tipo de reglas no aportan información al objetivo de negocio a resolver.

Una vez más, debido a que el modelo descarta o no generaliza ningún evento relacionado con decesos y muertes, como es posible apreciar en la matriz de confusión presentada a continuación, los resultados presentados previamente responden exclusivamente a las causas de accidentes que incurren en daños a propiedad privada y en heridas a individuos, y por lo tanto, este modelo solo debe ser empleado en el análisis y promoción de campañas que reduzcan el número de accidentes relacionado con este tipo de eventualidades.

Finalmente, conforme a estos resultados, es posible apreciar que las causas de accidentalidad relacionadas con daños y presencia de heridos se encuentran estrechamente relacionadas y son similares, es posible concluir esto a partir del número de falsos positivos reportados para la categoría Accidentes con heridos, bajo la cual, alrededor del 50% de los casos es clasificado como un accidente que ocasiona daños, sin embargo, es posible apreciar que solo el 1.9% de los casos en los cuales el accidente deriva en daños es clasificado incorrectamente como un accidente con heridos, lo cual permite concluir que los accidentes que presentan heridos también presentan daños, sin embargo, los accidentes que causan daños de forma exclusiva, no presentan heridos, y por lo tanto presentan características libres de ambigüedad o posible sesgo.

```
J48 pruned tree
------------------

tipoafectacion = CHOQUE: SOLO DANOS (131246.0/39171.0)
tipoafectacion = ATROPELLO: CON HERIDOS (19346.0/718.0)
tipoafectacion = OTRO: CON HERIDOS (5797.0/728.0)
tipoafectacion = VOLCAMIENTO
|   area = URBANO: CON HERIDOS (6131.0/481.0)
|   area = RURAL
|   |   hora = Mañana: CON HERIDOS (474.0/227.0)
|   |   hora = Tarde: CON HERIDOS (469.0/212.0)
|   |   hora = Noche: CON HERIDOS (381.0/153.0)
|   |   hora = Madrugada: SOLO DANOS (323.0/162.0)
tipoafectacion = CAIDA OCUPANTE: CON HERIDOS (6625.0/99.0)
tipoafectacion = INCENDIO: SOLO DANOS (62.0/10.0)
```

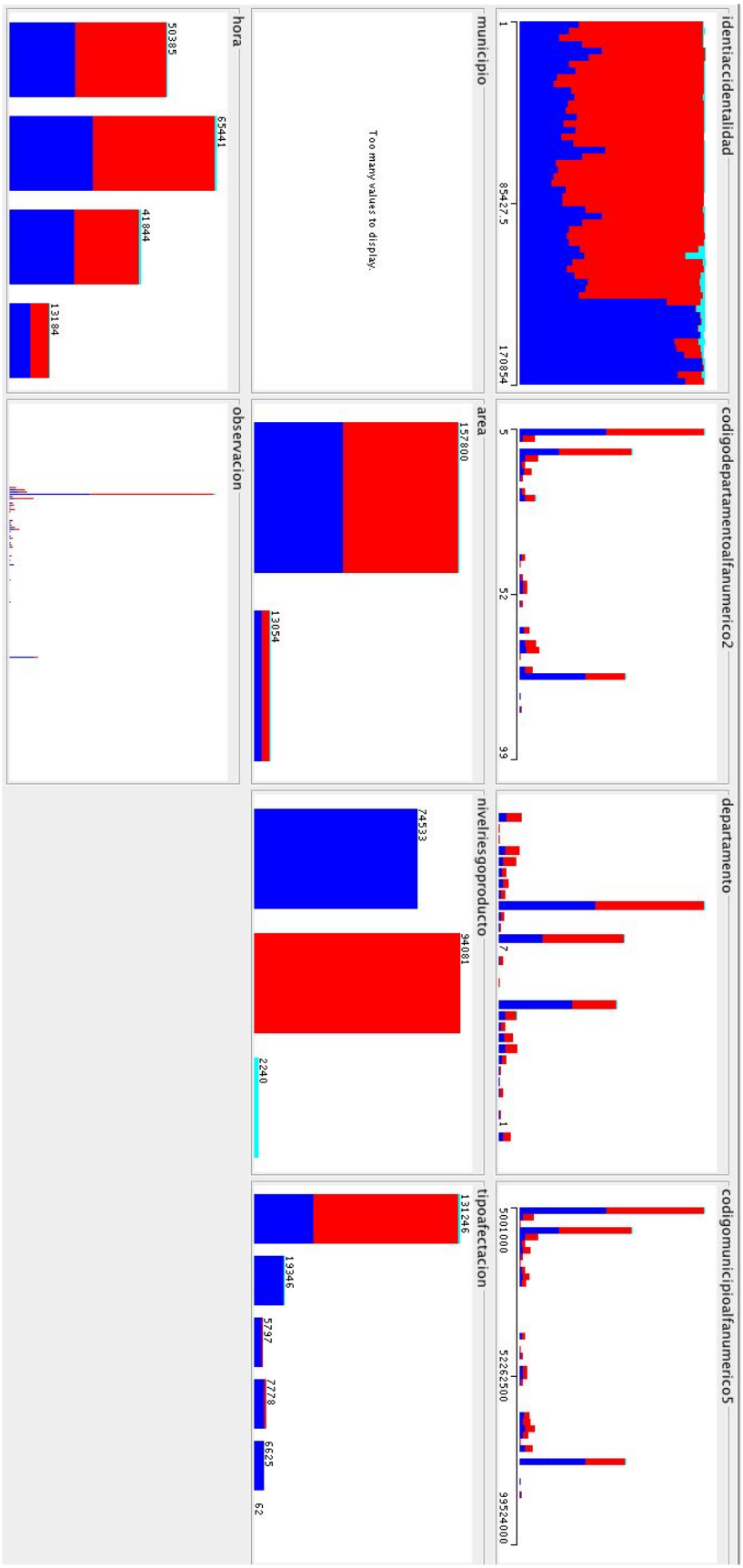
```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.491    0.027    0.933      0.491    0.643      0.547    0.734     0.754     CON HERIDOS
                 0.981    0.513    0.701      0.981    0.818      0.554    0.736     0.699     SOLO DANOS
                 0.000    0.000    0.000      0.000    0.000      0.000    0.602     0.025     CON MUERTOS
Weighted Avg.    0.754    0.294    0.793      0.754    0.731      0.544    0.733     0.714

=== Confusion Matrix ===

     a     b     c   <-- classified as
 36597 37936     0 |     a = CON HERIDOS
  1792 92289     0 |     b = SOLO DANOS
   824  1416     0 |     c = CON MUERTOS
```

7. Las estrategias que la organización debe plantear relacionadas con los resultados obtenidos en los modelos y una justificación de porqué esa información es útil para ellos.

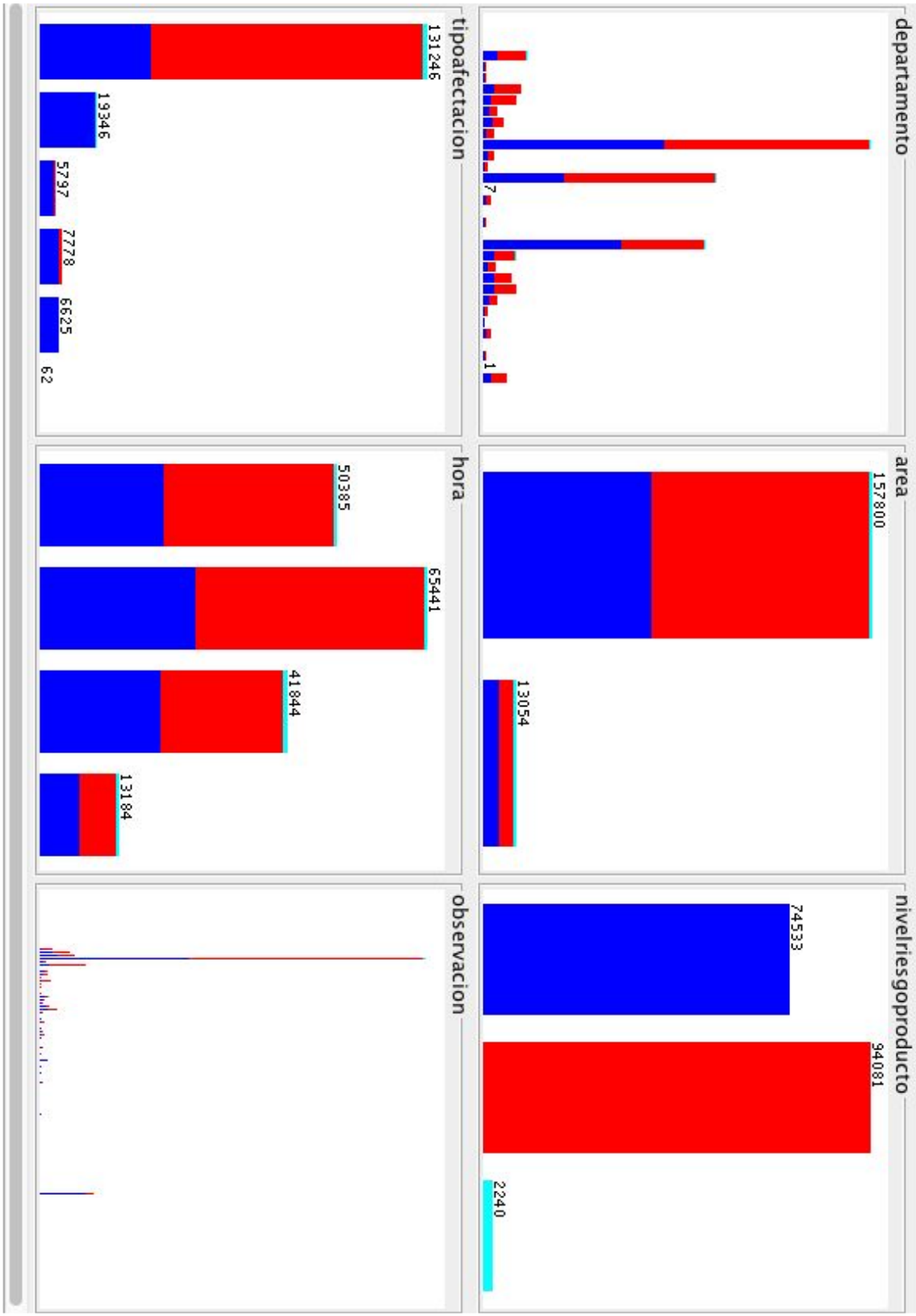
Se concluyó entonces que aunque es común pensar que los accidentes ocurren con mayor regularidad en la noche, este no es el caso. Por eso, aumentar la cantidad de retenes y agentes de tránsito a esta hora podría ser una distribución inadecuada de los recursos. Podría ser beneficioso que los recursos se destinarán a horas como la mañana y la tarde donde se tiene más tráfico y por consecuente ocurren más accidentes. Esto es útil para el ministerio de transporte puesto que pueden reorganizar sus recursos y así cumplir uno de sus objetivos, que es velar por un transporte más seguro y una mejor respuesta a los accidentes. Así, los momentos del día donde se concentran más accidentes podrán tener más agentes presentes en las cercanía y así una respuesta más rápida a los mismos, en especial en los momentos donde hay heridos, de manera que se pueda responder más fácil a sus necesidades médicas.

Además, se encontró que los departamentos con mayor número de accidentes que involucran heridos son el departamento de Antioquia y Bogotá D.C. Dado que el ministerio también tiene la responsabilidad de crear campañas y políticas con el fin de aumentar la seguridad del transporte del país, esta información podría serle muy útil para saber que debe concentrar sus esfuerzos de campañas de seguridad víal en estos dos departamentos.

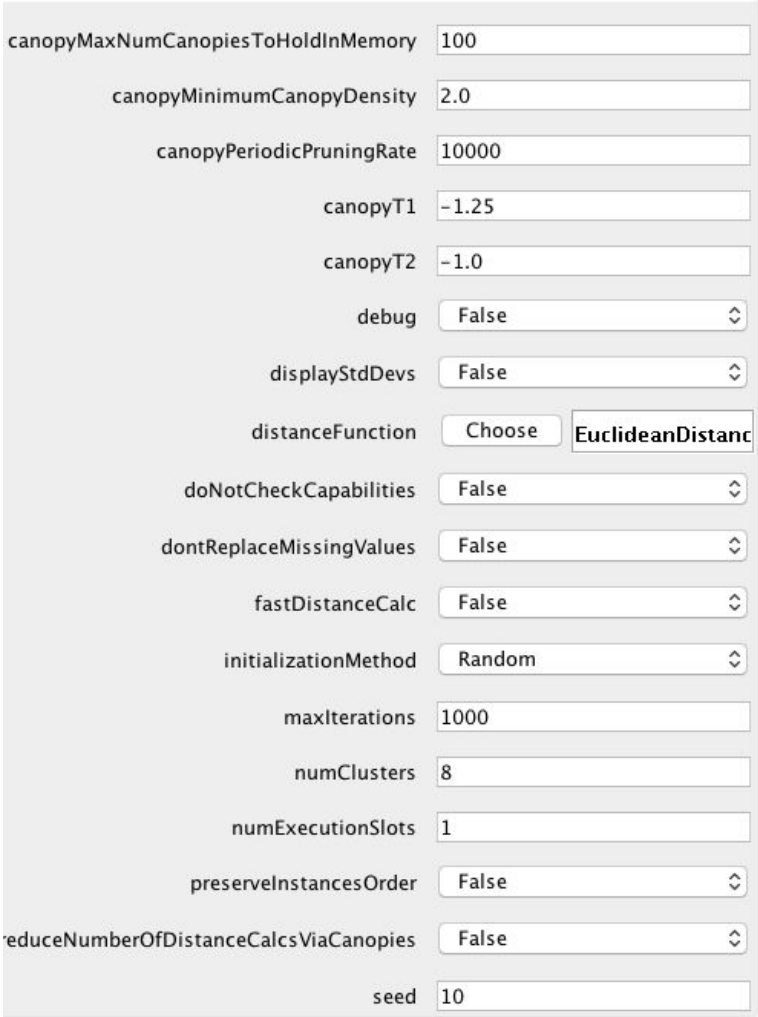
Anexo 1.


identiaccidentalidad
1
85427.5
170854
codigodepartamentoalfanumerico2
5
52
99
departamento
7
1
codigomunicipioalfanumerico5
5001000
52262500
99524000
municipio
Too many values to display.
area
157800
13054
nivelriesgoproducto
74533
94081
2240
tipoafectacion
131246
13346
5797
2778
6625
62
hora
50385
65441
41844
13184
observacion

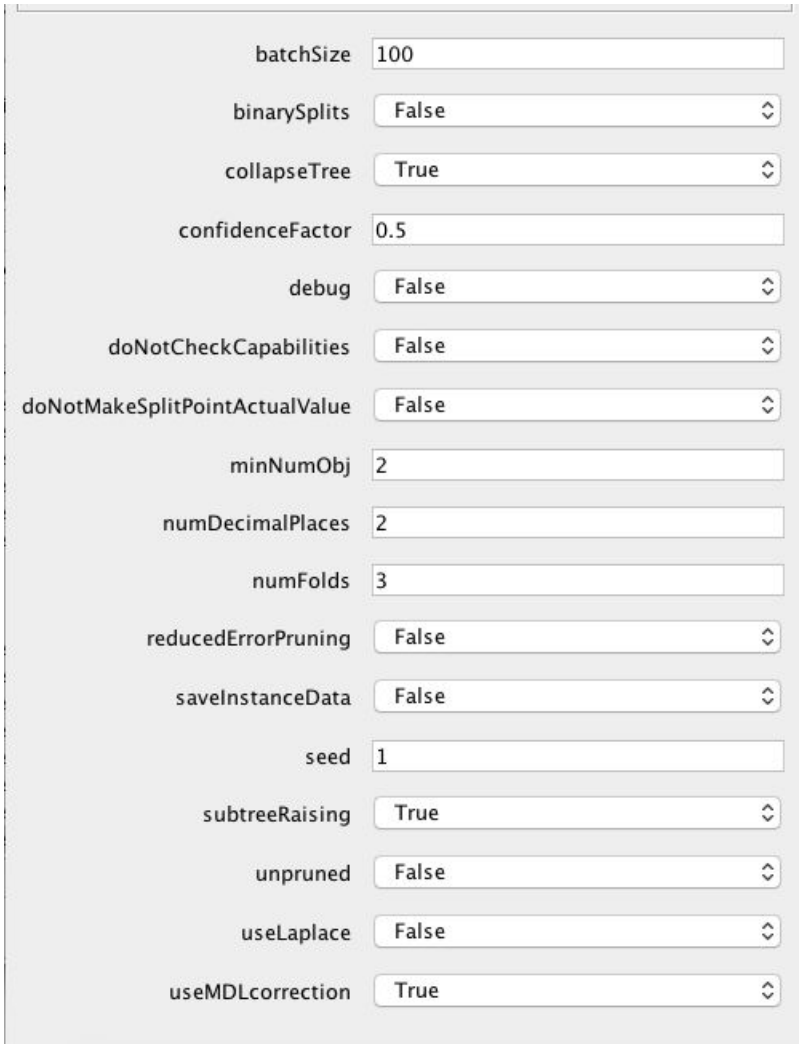
Anexo 2


departamento
area
nivelriesgoproducto
tipoafectacion
hora
observacion
157800
13054
74533
94081
2240
131246
19346
5797
7778
6625
62
50385
65441
41844
13184

Anexo 3


canopyMaxNumCanopiesToHoldInMemory
100
canopyMinimumCanopyDensity
2.0
canopyPeriodicPruningRate
10000
canopyT1
-1.25
canopyT2
-1.0
debug
False
displayStdDevs
False
distanceFunction
Choose
EuclideanDistanc
doNotCheckCapabilities
False
dontReplaceMissingValues
False
fastDistanceCalc
False
initializationMethod
Random
maxIterations
1000
numClusters
8
numExecutionSlots
1
preserveInstancesOrder
False
reduceNumberOfDistanceCalcsViaCanopies
False
seed
10

Anexo 4


batchSize 100
binarySplits False
collapseTree True
confidenceFactor 0.5
debug False
doNotCheckCapabilities False
doNotMakeSplitPointActualValue False
minNumObj 2
numDecimalPlaces 2
numFolds 3
reducedErrorPruning False
saveInstanceData False
seed 1
subtreeRaising True
unpruned False
useLaplace False
useMDLcorrection True