



UNIVERSIDAD DE GRANADA

PRÁCTICA 2

PREPARACIÓN DE DATOS CON KNIME

TRATAMIENTO INTELIGENTE DE DATOS

MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

AUTOR

Pablo Valenzuela Álvarez (pvalenzuela@correo.ugr.es)



CONTENIDO

Ejercicio 1. Factores de riesgo para el cáncer de cuello uterino	3
Análisis de datos.....	3
Preprocesado	4
Valores perdidos	5
Visualización	6
Eliminación de features.....	7
CAIM Binner.....	7
Auto Binner.....	8
Correlación Lineal	10
PCA	12
Conclusión	13
Ejercicio 2. Gravedad en accidentes de tráfico	14
Preparación de datos	14
Discretización	16
Valores perdidos	18
Selección de características	19
Selección de instancias.....	21

EJERCICIO 1. FACTORES DE RIESGO PARA EL CÁNCER DE CUELLO UTERINO

ANÁLISIS DE DATOS

Empezamos analizando los atributos de los que disponemos en el documento de la práctica 2.

- Age - Edad (**numérico**).
- Number of sexual partnerts – Parejas sexuales (**numérico**).
- First sexual intercourse – primera relación sexual (**numérico**).
- Num of pregnancies – número de embarazos (**numérico**).
- Smokes - Fumadora (**booleano**).
- Smokes (years) – años fumando (**numérico**).
- Smokes (packs/year) – paquetes de cigarrillos al año (**numérico**).
- Hormonal Contraceptives – hormonas anticonceptivas (**booleano**).
- Hormonal Contraceptives (years) – años tomando hormonas anticonceptivas (**numérico**).
- IUD - usa DIU (**booleano**).
- IUD (years) - años usando DIU (**numérico**).
- STDs - enfermedad de transmisión sexual (**booleano**).
- STDs (number) - Número de enfermedades sexuales contraídas (**numérico**).
- STDs:[n] - Variedad de enfermedad sexual (**booleano**)....
- STDs: Time since first diagnosis - Tiempo desde el primer diagnóstico (**numérico**).
- STDs: Time since last diagnosis - Tiempo desde el último diagnóstico (**numérico**).
- Dx: Cancer – Diagnóstico de cáncer (**booleano**).
- Dx: HPV – Diagnóstico de virus papiloma (**booleano**).
- Dx: Diagnóstico de alguno de los anteriores (**booleano**).
- Hinselman: se realizó colonoscopia (**booleano**).
- Schiller – se realizó prueba de Schiller (**booleano**).
- Citology – se realizó citología (**booleano**).
- Biopsy – se realizó biopsia (**booleano**).

Una vez cargado el set de datos en Knime podemos observar que todos los datos los toma como números (enteros y decimales) incluyendo las variables consideradas **booleanas** en el apartado anterior. Habría que transformar estas columnas a tipo categórico.

También se pueden observar en la Foto 1 los valores perdidos que poseen algunas variables marcadas con el carácter '?'. Esto habría que controlarlo llegando hasta incluso eliminar atributos con más de un cierto margen de valores perdidos.

Row ID	I	Age	D	Number...	D	First se...	D	Num of ...	D	Smokes	D	Smokes...	D	Smokes...	D	Hormon...	D	Hormon...	D	IUD	D	IUD (ye...	D	STDs	D	STDs (n...	D	STDs:c...	D	STDs:c...
Row372	21	5	13	3	1	1.267	0.513	1	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row373	23	2	19	2	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row374	21	5	14	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row375	22	1	16	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row376	41	3	19	2	0	0	0	1	4	1	5	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row377	20	2	14	4	1	3	3	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row378	18	2	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row379	21	2	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row380	18	3	15	1	1	2	0.003	1	0.58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row381	19	3	15	2	0	0	0	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Row382	25	5	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row383	19	2	15	3	0	0	0	1	0.58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row384	18	1	16	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row385	21	3	14	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row386	18	2	15	3	1	7	3.5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Foto 1. Muestra inicial de las variables.

PREPROCESADO

Empezamos pasando todas las variables de tipo decimales a entero, exceptuando las variables:

- Años de fumadora - Smokes (years).
- Paquetes de tabaco por año – Smokes (packs/year)
- Años tomando anticonceptivos – Hormonal Contraceptives (years)

Estas variables son las únicas que tienen decimales porque incluyen el porcentaje de los días en esos decimales.

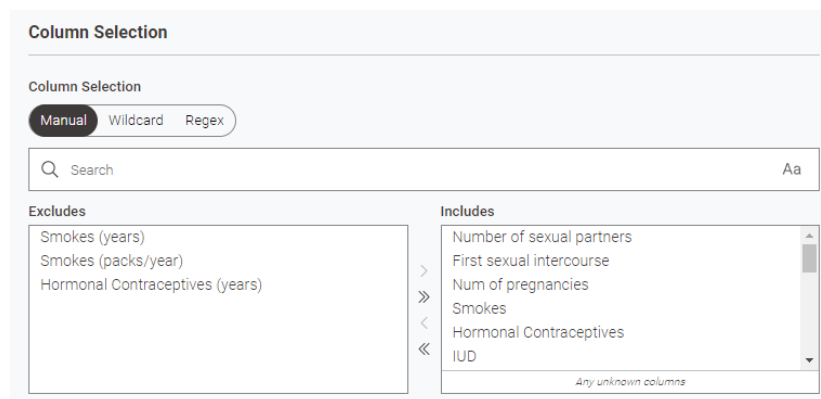


Foto 2. Decimal a entero.

Y transformamos las variables categóricas identificadas en el apartado anterior.

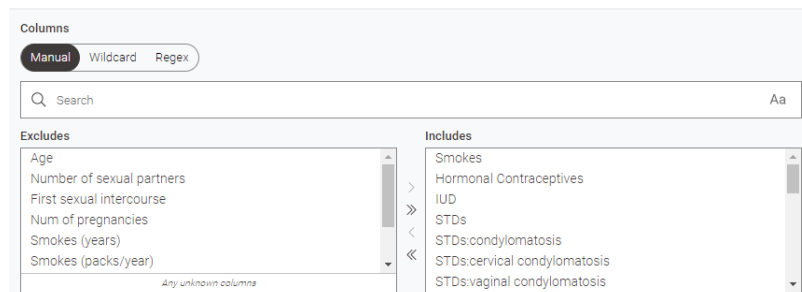


Foto 3. Transformando las variables numéricas a categóricas.

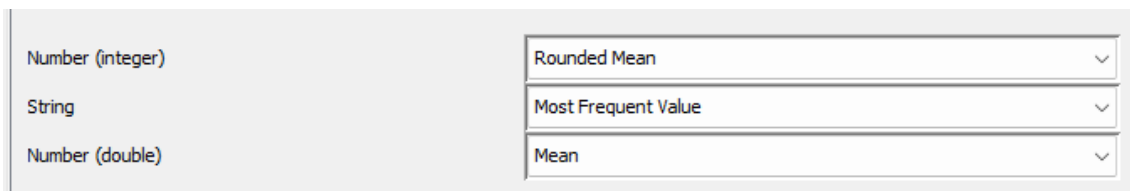
VALORES PERDIDOS

El siguiente paso es tratar con los valores perdidos. Las variables con mucha cantidad de estos valores no aportan nada y podemos prescindir de ellas.

Vamos a aplicar el nodo *Missing Value Column Filter* con un valor del 90%, es decir, que filtrará los atributos con máximo de 10% de valores perdidos.

El resultado es que hemos eliminado dos columnas, en concreto *STDs: Time since first diagnosis* y *STDs: Time since last diagnosis*. Estos atributos se pueden mantener y gestionar sus valores de distintas formas, pero al tener tantos valores perdidos resultaría que tendríamos bastantes de ellos repetidos y falsos.

Seguimos teniendo valores perdidos en nuestros datos, por lo que es necesario usar una función que les de valores. Hay diferentes técnicas, usar la media, varianza, el valor más frecuente; yo voy a usar el nodo *Missing Value* para imputar estos valores como se muestra en la Foto 4.



Number (integer)	Rounded Mean
String	Most Frequent Value
Number (double)	Mean

Foto 4. Imputando valores perdidos.

Otra opción sería eliminar las filas que contengan los valores perdidos, resultando en la eliminación de 190 filas (más del 20% de nuestros datos). Seguiré con lo comentado en el párrafo anterior.

VISUALIZACIÓN

Tras realizar una visualización sobre las variables numéricas podemos ver que algunas características se suelen agrupar en rangos.

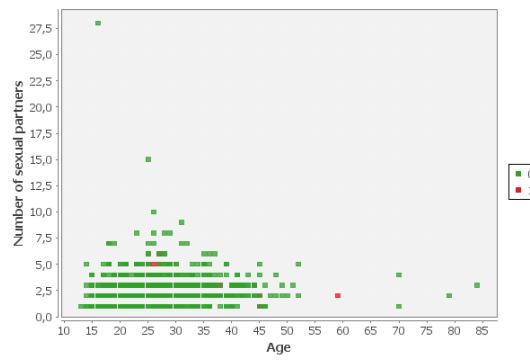


Foto 5: Agrupación de casos 1.

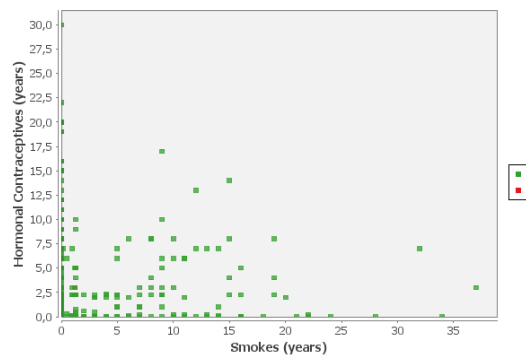


Foto 6: Agrupación de casos 2.

Por lo que sería interesante crear “grupos” en estos atributos para poder verlos más claramente. Esto podemos realizarlo usando distintos nodos, los cuales vamos a explicar a continuación.

CAIM BINNER

Vamos a aplicar el nodo *CAIM-Binner* para categorizar los datos numéricos como se ve en la Foto 7. Discretizamos sobre la variable **Dx:Cancer**.

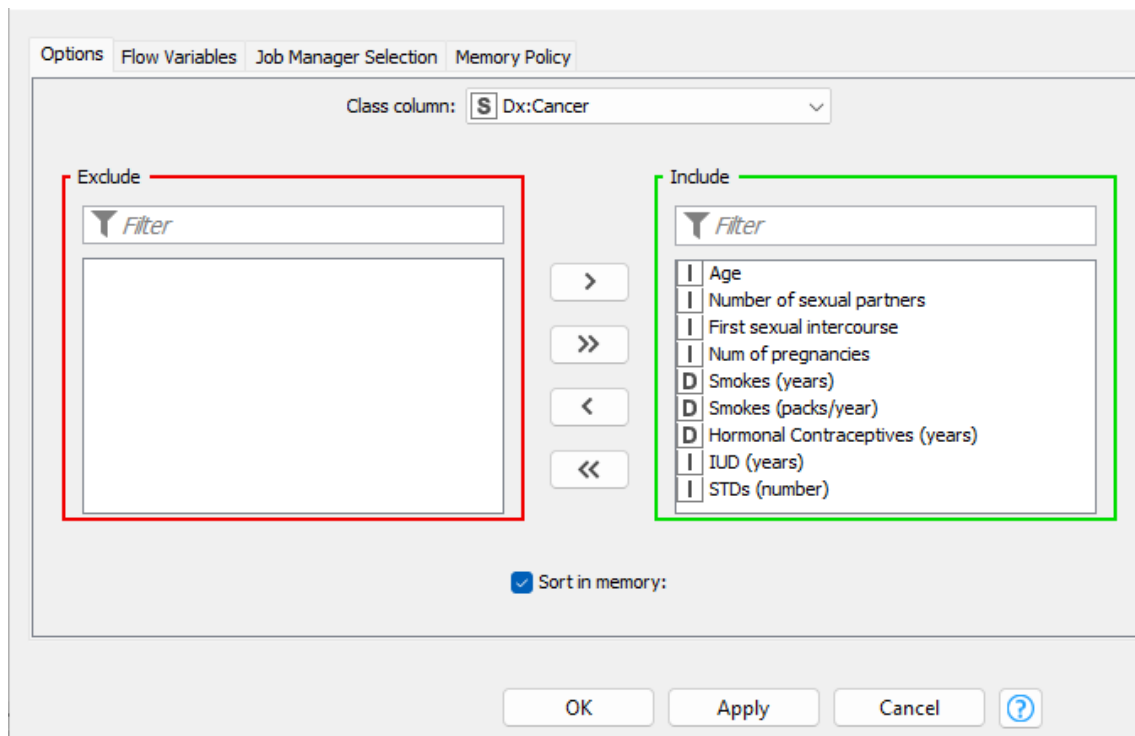


Foto 7. Categorización datos numéricos.

El problema de usar este nodo (como se puede ver en la Foto 8) es que al tener todos los valores tan agrupados, solo encuentra dos intervalos para casi todas los atributos del problema. Además no ofrece información sobre que rangos se mueven los valores de estos atributos. En mi caso queda descartado su uso.

Numeric

Nominal

Data Preview

Search:







Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
Age	<input type="checkbox"/>	0	2	Interval_1, Interval_0	
Number of sexual partners	<input type="checkbox"/>	0	2	Interval_0, Interval_1	
First sexual intercourse	<input type="checkbox"/>	0	2	Interval_1, Interval_0	
Num of pregnancies	<input type="checkbox"/>	0	2	Interval_0, Interval_1	
Smokes	<input type="checkbox"/>	0	2	0, 1	
Smokes (years)	<input type="checkbox"/>	0	2	Interval_0, Interval_1	

Foto 8. Problema con el nodo CAIM Binner

AUTO BINNER

Otro nodo que podemos usar es el *Auto Binner*. La configuración (ver Foto 9) nos deja crear un número fijo de grupos por atributo y además nombrar a este grupo. En mi caso, el nodo añadirá nueve columnas nuevas (reemplazando las columnas onjetivo) con cinco grupos en cada una que incluirán en el nombre los bordes de cada uno.

Auto Binner Settings | Number Format Settings | Flow Variables | Job Manager Selection | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

- ☐ Number of sexual partners
- ☐ First sexual intercourse
- ☐ Num of pregnancies
- ☒ Smokes (years)
- ☒ Smokes (packs/year)
- ☒ Hormonal Contraceptives (years)
- ☐ IUD (years)
- ☐ STDs (number)

☐ Enforce inclusion

> >> < <<

Binning Method

☒ Fixed number of bins

Number of bins: 5

Equal: width

☐ Sample quantiles

Quantiles (comma separated): 0.0, 0.25, 0.5, 0.75, 1.0

Bin Naming

☐ Numbered e.g.: Bin 1, Bin 2, Bin 3

☒ Borders e.g.: [-10,0], (0,10], (10,20]

☐ Midpoints e.g.: -5, 5, 15

☐ Force integer bounds

☒ Replace target column(s)

OK Apply Cancel ?

Foto 9. Configuración Auto Binner

Si filtramos por los casos positivos podemos sacar gráficos como los siguientes.

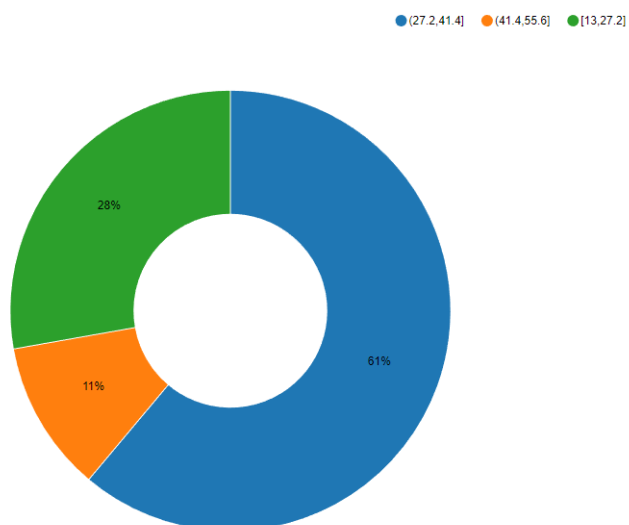


Foto 10. Gráfico de casos positivos x Edad

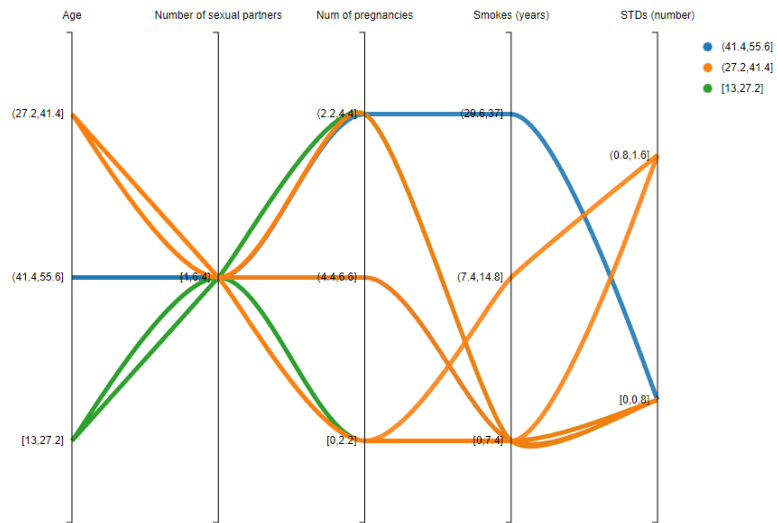


Foto 11. Gráfico paralelo estudiando todas las variables numéricas para los casos positivos.

CORRELACIÓN LINEAL

Para estudiar la correlación lineal, podemos usar el nodo *Linear Correlation*.

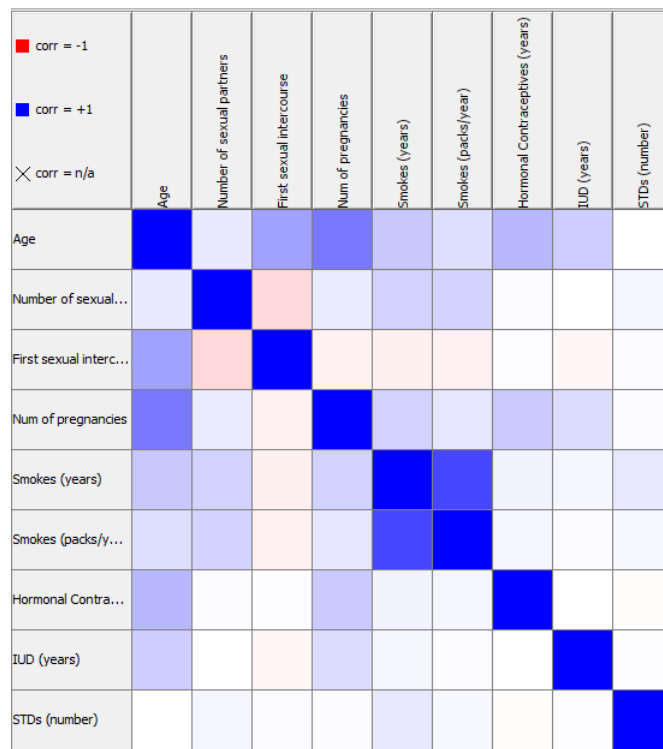


Foto 12. Correlación lineal de las variables numéricas.

En la foto 12, se observan ciertas correlaciones positivas empezando por:

1. Las características de Smoke (years) y Smoke (packs/year): A más años fumando más paquetes/año.
2. La edad y número de embarazos.

Y una negativa: First sexual intercourse y number of partners. A más edad, menos parejas.

Podemos usar otro nodo que se encargue de eliminar ciertas características que superen el grado de correlación que introduzcamos. *Correlation Filter* elimina variables que superen los límites definidos.

The screenshot shows the 'Correlation Filter' node configuration. On the left, under 'Columns from Model', a list of variables is shown with checkboxes: Age, Number of sexual p..., First sexual intercou..., Num of pregnancies, Smokes (years), Smokes (packs/year), Hormonal Contracep..., and IUD (years). The 'Smokes (years)' and 'Smokes (packs/year)' variables are highlighted. On the right, the 'Correlation Threshold' is set to 0,8. Below the threshold, the statistics are: Include columns: 9, Exclude columns: 0, and Total columns: 9. A 'Calculate' button is at the bottom right.

Foto 13. Correlación lineal de 80%.

The screenshot shows the 'Correlation Filter' node configuration with the threshold set to 0,7. The 'Smokes (years)' variable is now highlighted. The statistics are: Include columns: 8, Exclude columns: 1, and Total columns: 9. The 'Calculate' button remains at the bottom right.

Foto 14. Correlación lineal del 70%.

The screenshot shows the 'Correlation Filter' node configuration with the threshold set to 0,5. The 'First sexual intercou...' variable is now highlighted. The statistics are: Include columns: 7, Exclude columns: 2, and Total columns: 9. The 'Calculate' button remains at the bottom right.

Foto 15. Correlación lineal del 50%.

Como podemos ver en las fotos 13, 14 y 15, a medida que se aumenta el valor de *correlation threshold*, van aumentando las columnas excluidas.

PCA

Por último, podemos aplicar el nodo *PCA* que funciona de forma similar al anterior, eliminando variables.

Foto 16. Configuración del nodo PCA.

La foto 16 muestra la configuración del nodo, donde podemos reducir las dimensiones al número que introduzcamos o, reducir las dimensiones sin perder mucha información. A este último según el porcentaje que digamos podemos reducir las variables numéricas (9) a:

- 5 variables, perdiendo un 5% de información.
- 4 variables, perdiendo el 10%.
- 2 variables, perdiendo el 30%.

CONCLUSIÓN

Para este caso, el mejor método a aplicar sería el *Correlation Filter* y el *Auto Binner*. El primero porque permite mantener el nombre de las variables y solo eliminarlas si superan un cierto rango, mientras que *PCA* hace sus propias variables y no sabemos que representan. *Auto Binner* funciona mejor que *CAIM* en este caso, porque los datos están muy agrupados en ciertas variables y *CAIM* no logra identificar muchos rangos.

Si queremos eliminar variables debemos tener en cuenta su correlación, y esta tiene que ser positiva y muy cercana al 1. De esta manera nos aseguramos de que no perdemos mucha información ya que esta representada por otra. Sin embargo, hay que tener cuidado de cuantas variables eliminamos porque se corre el riesgo de perder mucha información.

EJERCICIO 2. GRAVEDAD EN ACCIDENTES DE TRÁFICO

PREPARACIÓN DE DATOS

Antes de empezar hay que tratar los valores perdido tal y como lo pone en la descripción del dataset. Se ha usado alguna técnica de imputación de los valores perdidos para estas variables, y están recogidas en columnas extras (esto hay que tenerlo en cuenta para más tarde). Las variables son las siguientes:

Variable	Valor	Imputados
WEEKDAY	9	WKDY_I
HOUR	99	HOUR_I
INT_HWY	9	----
PROFILE	9	PROFIL_I
SUR_COND	9	SURCON_I
TRAF_CON	99	TRFCON_I
SPD_LIM	99	SPDLIM_H
LIGHT_CON	9	LGTCON_I
WEATHER	9	WEATHR_I
ALCOHOL	9	ALCHL_I

Tabla 1. Gestión de valores perdidos.

Para las pruebas iniciales usaremos el conjunto pequeño. He creado el siguiente metanodo donde todos los valores se ponen en blanco según se dice en la tabla anterior.

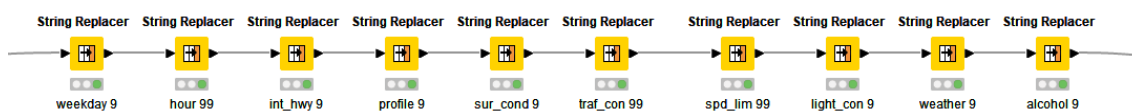


Foto 16. Estructura del metanodo.

El documento de la práctica nos pide crear la variable objetivo mezclando los valores de las variables *FATALITIES*, *INJURY_CRASH*, *PRPTYDMG_CRASH*. El método es de libre elección, por lo que optaré por lo siguiente:

- X tendrá será 0 cuando el accidente no tenga fallecidos
- X será 1 cuando haya fallecidos.

Antes se había usado otra fórmula (la cual esta mostrada en la foto 17), pero los resultados obtenidos no eran muy buenos y se descartó.

Expression

```

1 if($PRPTYDMG_CRASH$, 1,
2   if($INJURY_CRASH$, 2,
3     if($FATALITIES$, 3, 0)
4   )
5 )

```

☒ Append Column: X
☐ Replace Column: PRPTYDMG_CRASH
☐ Convert to Int

Foto 17. Asignando valor a la variable objetivo

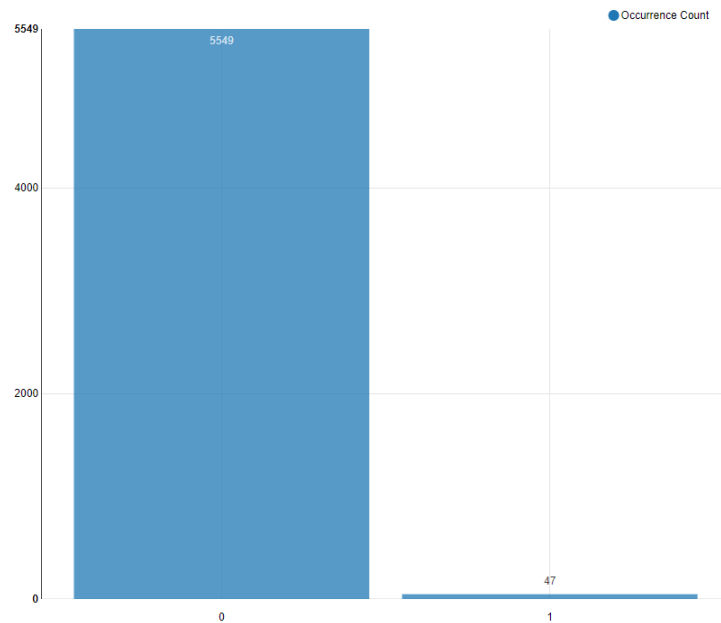


Foto 18. Distribución de la variable de clase

Podemos observar en la foto 18, que para ejemplos de la clase del valor 1 hay muy pocas muestras. Esto puede ser un problema a la hora de predecir, porque el algoritmo que se use estará mejor entrenado para reconocer las clases con valor 0, y podrá predecir mal si tiene valor 1.

DISCRETIZACIÓN

Parte A

Se puede apreciar la división de las características en la foto siguiente, también podemos observar que hay solo hay menos del 1% del caso positivo.

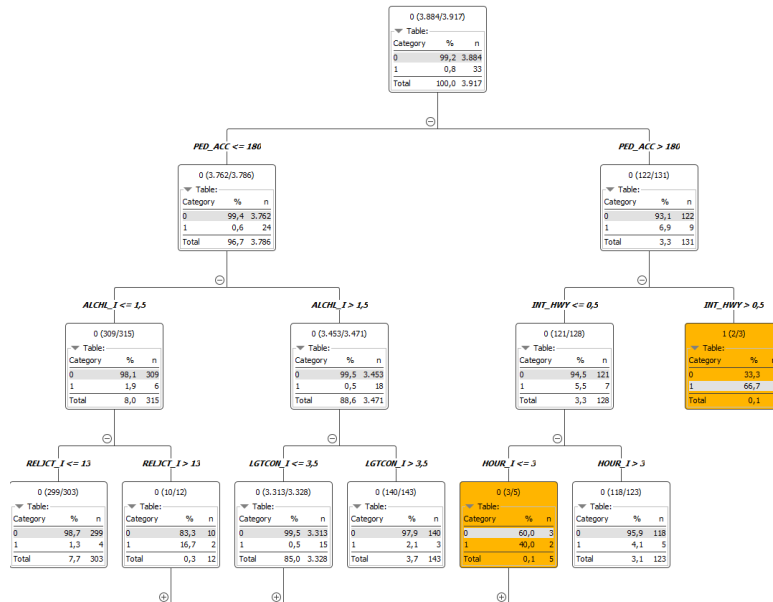


Foto 19. División de las características del árbol de decisión

La precisión de este modelo se puede observar en la siguiente foto. Se ha probado con el conjunto pequeño usando una partición del 70/30 de los datos, es decir, 70% para entrenar y el resto para validar, muestra estratificada con la semilla 12345.

Confusion Matrix...		
File Hilite		
X \ Predicti...	0	1
0	1661	4
1	14	0
Correct classified: 1.661 Wrong classified: 18		
Accuracy: 98,928% Error: 1,072%		
Cohen's kappa (κ):		

Foto 20. Precisión y error del modelo anterior.

Se puede observar hay una alta precisión, pero eso es debido a las pocas muestras del caso positivo. Si nos fijamos en la matriz de confusión de la foto 20, nunca acierta.

Si hacemos over-sampling (sobre muestreo) usando el nodo *SMOTE* sobre esta clase obtenemos los resultados de la foto siguiente.

X \ Predicti...	0	1
0	1632	33
1	2	1663

Correct classified: 3.295 Wrong classified: 35
 Accuracy: 98,949% Error: 1,051%
 Cohen's kappa (κ):

Foto 21. Resultados obtenidos tras hacer over-sampling

La precisión y el error mejoran un poco, tendremos que probar sobre el conjunto grande para comprobar si esta técnica mejora algo.

Over-sampling	Accuracy	Error
No	98.553%	1.447%
Si	98.713%	1.287%

Tabla 2. Resultados de usar over-sampling

Parte B y C

Se usarán a continuación los algoritmos CAIM Binner y Auto Binner probando la técnica de over-sampling (OS) con ellos.

Algoritmo	Accuracy	Error
CAIM Binner	98.92%	1.07%
CAIM Binner (OS)	96.91%	3.10%
Auto Binner (5 bins)	98.68%	1.32%
Auto Binner (5 bins) (OS)	98.25%	1.75%
Auto Binner (3 bins)	99.11%	0.89%
Auto Binner (3 bins) (OS)	95.43%	4.57%
Auto Binner (2 bins)	99.17%	0.83%
Auto Binner (2 bins) (OS)	89.1%	10.9%

Tabla 3. Probando CAIM Binner y Auto Binner

Viendo los resultados de la tabla 3, obtenemos los mismos resultados aplicando CAIM o no en el conjunto sin over-sampling, y en ambos algoritmos empeoran tras hacerlo. Aplicando diferentes números de bins para Auto-Binner obtenemos mejores resultados que empeoran al hacer over-sampling. Esto es debido a que antes de aplicar over-sampling el algoritmo siempre predice una clase y nunca la otra, y al haber muy pocos ejemplos no empeoran la precisión, pero, en el momento que igualamos las clases los errores afectan más a esta medida empeorandola.

VALORES PERDIDOS

Vamos a ejecutar ahora el algoritmo sobre los datos con valores perdidos usando distintas técnicas. Usaremos el conjunto discretizado que obtuvimos en el punto anterior usando *Auto-Binner* y 5 bins.

Método usado	Accuracy	Error
Sin imputar	98.99%	1.01%
Sin imputar (OS)	97.25%	2.75%
Imputados	98.68%	1.32%
Imputados (OS)	98.22%	1.78%
Imputar media	98.99%	1.01%
Imputar media (OS)	97.64%	2.36%
Eliminar instancias	98.25%	1.75%
Eliminar instancias (OS)	97.16%	2.84%
Eliminar features	98.69%	1.31%
Eliminar features (OS)	96.79%	3.21%

Tabla 4. Distintos métodos de imputación de valores perdidos

No se ha conseguido mejorar al menos en el conjunto con over-sampling, el resultado obtenido de la variable con valores imputados que traía el problema. Aún así las medidas obtenidas son parecidas.

SELECCIÓN DE CARACTERÍSTICAS

Ahora vamos a utilizar el conjunto de datos completo, sobre el algoritmo obtenido en puntos anteriores, (*Auto-Binner* y con cinco *bins*).

Método usado	Accuracy %	Error %
Imputados	98.75	1.25
Imputados (OS)	96.65	3.35
Media	98.8	1.2
Media (OS)	95.07	4.92
Eliminando features	98.88	1.12
Eliminando features (OS)	92.73	7.27

Tabla 5. Comparación de medidas con el conjunto completo.

Vistos los resultados de la Tabla 5, hemos conseguido superar la precisión tanto usando la media para los valores perdidos, como eliminando características, a excepción de los conjuntos con sobre muestreo.

Por último, para esta sección, vamos a probar el metanodo de *Backward Feature Elimination* con el algoritmo C 4.5 para ver si obtenemos mejores resultados. Vamos a utilizar la versión que imputa la media como valor perdido.

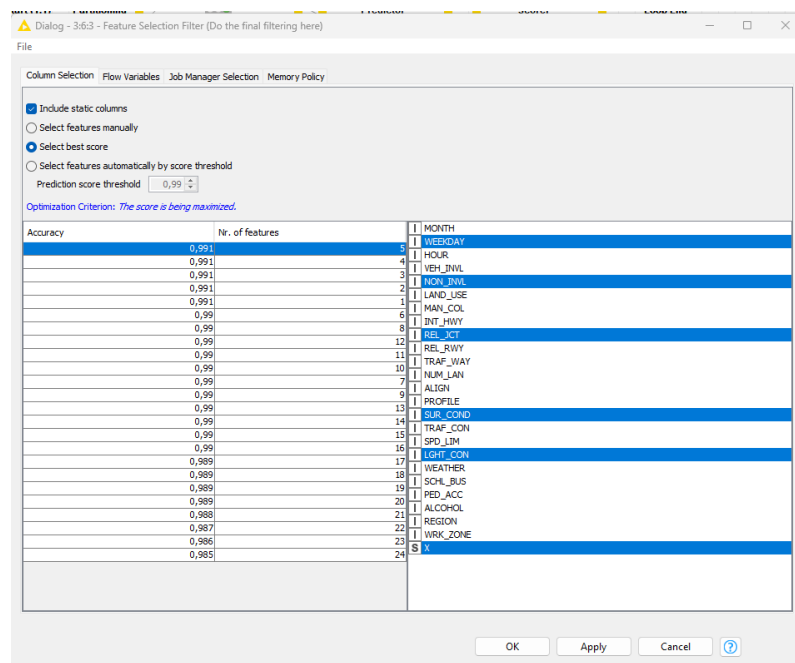


Foto 22. Resultado selección de características de forma recursiva.

Según los resultados vistos en la foto 22, podemos obtener una precisión del 99.1% con solo 5 variables.

Accuracy	Nr. of features	
0,988	22	D MONTH
0,988	21	D WEEKDAY
0,988	20	D HOUR
0,988	19	D VEH_INVL
0,988	18	D NON_INVL
0,988	17	D LAND_USE
0,988	16	D MAN_COL
0,987	15	D ENR_RWAY
0,987	14	D REL_JCT
0,987	13	D REL_RWAY
0,987	12	D TRAF_WAY
0,986	11	D NUM_LAN
0,986	10	D ALIGN
0,986	9	D PROFILE
0,986	8	D SUR_COND
0,985	7	D TRAF_CON
0,985	6	D SPD_LIM
0,984	5	D LIGHT_CON
0,983	4	D WEATHER
0,982	3	D SCHL_BUS
0,98	2	D PED_ACC
0,974	1	D ALCOHOL
0,922		D REGION
0,889		D WVR_ZONE
0,853		S X

Foto 23. Resultado de selección de características en el conjunto con over-sampling.

Probando con el conjunto con sobre muestreo, nos dice que tenemos una precisión del 98.8% seleccionando 18 características (ver foto 23).

Confusion Matrix - 3:...		
File Hilite		
X \ Predict...	0	1
0	16261	369
1	46	16584
Correct classified: 32.845 Wrong classified: 415		
Accuracy: 98,752% Error: 1,248%		
Cohen's kappa (κ): 0,975%		

Foto 24. Confirmación de los resultados seleccionando las características mostradas en la foto 23.

Como se ve en la foto 24, se han mejorado los resultados seleccionando esas 18 características.

SELECCIÓN DE INSTANCIAS

Podemos usar validación cruzada sobre los datos para ver si mejora el resultado. El metanodo *Cross Validation* implementa esta opción.

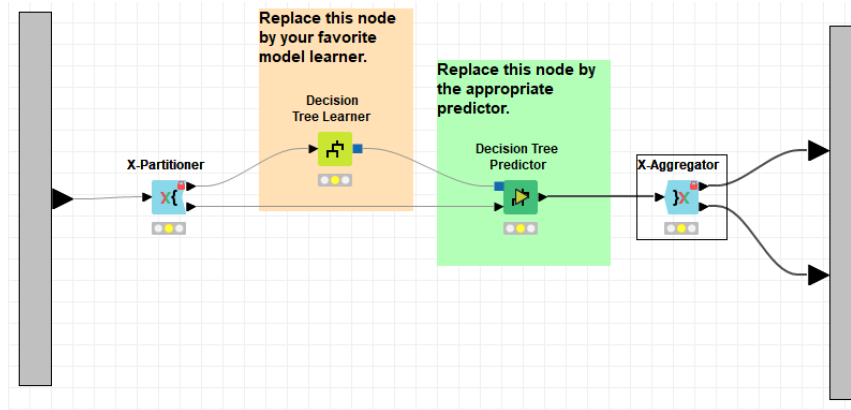


Foto 25. Metanodo Cross Validation.

En el nodo X-Partitioner definimos el número de validaciones, muestreo estratificado y una semilla fija. Lo que hacemos en la validación cruzada es dividir el conjunto en K lotes (el número de validaciones) y validar cada una por separado. Después nos quedamos con el mejor modelo (como se puede ver en la foto 26) que será aplicado al conjunto completo.

Error rates - 3:1:1 - X-Aggregator			
File Edit Hilite Navigation View			
Table "default" - Rows: 10 Spec - Columns: 3 Properties Flow Variables			
Row ID	D ▲ Error in %	I Size of Test Set	I Error Count
fold 8	0.992	11086	110
fold 9	1.064	11086	118
fold 7	1.073	11086	119
fold 3	1.109	11087	123
fold 4	1.127	11087	125
fold 1	1.136	11087	126
fold 6	1.146	11086	127
fold 0	1.2	11087	133
fold 5	1.254	11087	139
fold 2	1.353	11087	150

Foto 26. Prueba hecha con 10 validaciones ordenada por el error cometido.

En las siguientes tablas mostraremos las pruebas hechas con validación cruzada sobre los conjuntos obtenidos en el punto anterior (con la selección de características aplicada).

Sobre muestreo	Accuracy %	Error %
No	98.46	1.54
Si	98.75	1.25

Tabla 6. Resultado de los conjuntos sin aplicar validación cruzada.

Sobre muestreo	Lotes	Accuracy %	Error %
No	3	98.54	1.46
Si	3	98.55	1.45
No	5	98.56	1.44
Si	5	98.77	1.23
No	10	98.6	1.4
Si	10	98.85	1.15

Tabla 7. Resultado de los conjuntos aplicando validación cruzada.

Vistos los resultados de las tablas 6 y 7, podemos confirmar que usando una técnica de selección de instancias como la validación cruzada podemos mejorar la precisión de nuestro algoritmo. Se puede apreciar en la tabla 7, que a partir de 5 lotes se mejora el resultado de la tabla 6 tanto en el conjunto normal como en el que contiene sobre muestreo.

Por lo tanto, hemos aplicado como preprocesado:

1. Hemos usado *Auto Binner* para discretizar las características.
2. Hemos imputado los valores perdidos como la media.
3. Hemos usado el metanodo *Backward Feature Elimination* para hacer una selección de características y reducir la dimensionalidad.
4. Hemos realizado validación cruzada.

Para finalizar, con los pasos anteriores nos han servido para aumentar la precisión de nuestro algoritmo, las pruebas realizadas lo confirman.