UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

DEPARTAMENTO DE SISTEMAS DE INFORMACIÓN, CONTROL Y EVALUACIÓN DE RECURSOS INFORMÁTICOS

CARRERA LICENCIATURA EN INGENIERÍA DE SOFTWARE

GESTIÓN DE LA INFORMACIÓN

Proyecto Final

Estudio de análisis de datos

Integrantes:

Cortez, Brandool 3-742-294

Delgado, Oriel 8-970-187

Restrepo, Yulissa 8-961-1900

Profesora:

Ing. Carmen Ortega R.

SEMESTRE I, 2022

Contenido

INTRODUCCIÓN

La gestión de la información es un conjunto de procesos por los cuales se controla el proceso del ciclo de vida del software, esto va desde su obtención, hasta su disposición final. Esta información es útil para responder a necesidades, como por ejemplo la toma de decisiones. Con el paso de los años han surgido un conjunto de herramientas y técnicas que nos permiten a nosotros como usuarios gestionar fuentes de datos y organizar la información registrada, ya sea mediante técnicas de clasificación, asociación, aplicación de filtros, etc.

A continuación aplicaremos los conceptos, métodos, técnicas y herramientas asignadas durante el curso de gestión de la información para estudiar un conjunto de datasets asignados. Cada uno de estos datasets manejan información totalmente diferente a la anterior, por lo que se dispondrá de unos análisis y estudios totalmente distinto para cada conjunto de datos.

Nos dispondremos a revisar cada uno de los conjuntos de datos e idear posibles propuestas para la resolución de cada uno de estos. Mediante la generación de modelos que respondan a algún escenario.

DESCRIPCÍON DE LOS DATASET

1. Conjunto de datos titanic.arff

Descripción del dataset

El dataset original del Titanic, contiene cada uno de los datos sobre los pasajeros a bordo del barco. Los datos contenidos no guardaban información con respecto a la tripulación, pero contiene las edades reales de la mitad de los pasajeros.

La fuente principal de los datos recabados fue obtenida principalmente la Enciclopedia Titánica.

Información de dataset:

a. **Número de instancias:** 2201
b. **Número de atributos:** 4

El dataset proporcionado los generaliza la mayor parte de los atributos, de tal manera que quedan a libre interpretación. Los datos contenidos están clasificados respecto a los siguientes atributos:

- **Clase:** Se refiere al tipo de pasaje adquirido por cada uno de los pasajeros que abordaron el navío. Este se encuentra distribuido en el dataset de la siguiente manera: primera clase (1), segunda clase (2), tercera clase (3). Pero al comprender el valor 0 en el conjunto de datos, para tratarlos de una mejor manera, se los asignaremos a la tripulación.
- **Edad:** Corresponde a la edad de los pasajeros a borde. Para propósito de este dataset, estos sido clasificado en los valores 0 y 1, representando en el mismo orden a los adultos e infantes.
- **Sexo:** Corresponde al género de la persona. Los datos contenidos en este son los valores 0 y 1. Para propósito de nuestro análisis asignaremos para masculinos el valor 0, y 1 para femeninos.
- **Sobrevivió?:** Nos indica si el pasajero sobrevivió o caso contrario no, representado con los valores 1 y 0, referidos al mismo orden.

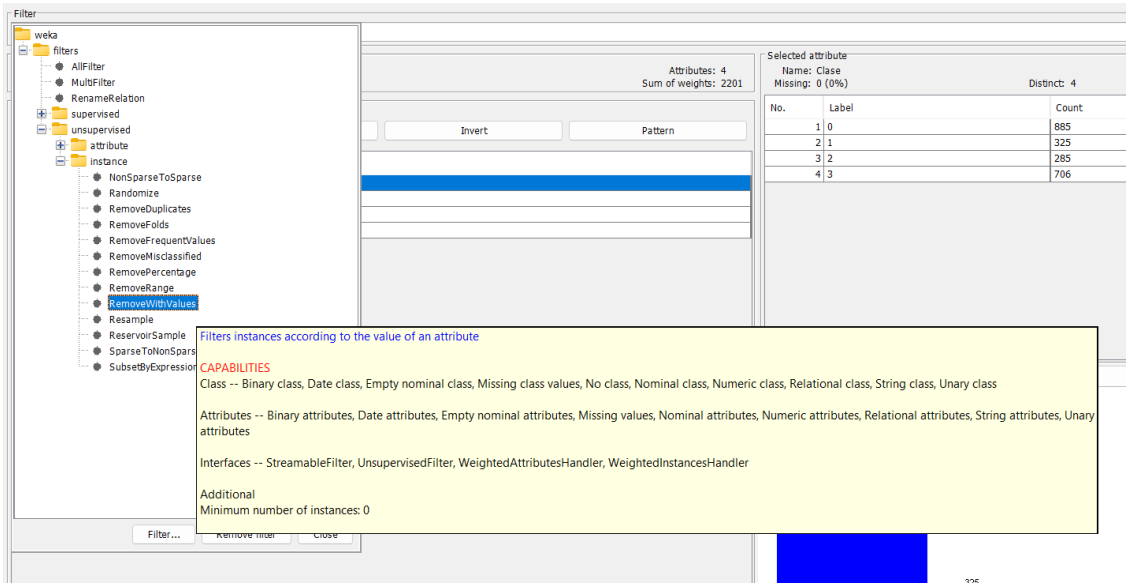
Preparación de los datos

Al realizar una evaluación previa a los datos de entrada podemos concluir que no posee data errónea o que simplemente los atributos no carecen de data para continuar con la evaluación. En general, el dataset está completo y los datos que este contiene corresponden a cada uno de los atributos. Por lo que podemos pasar a siguiente paso.

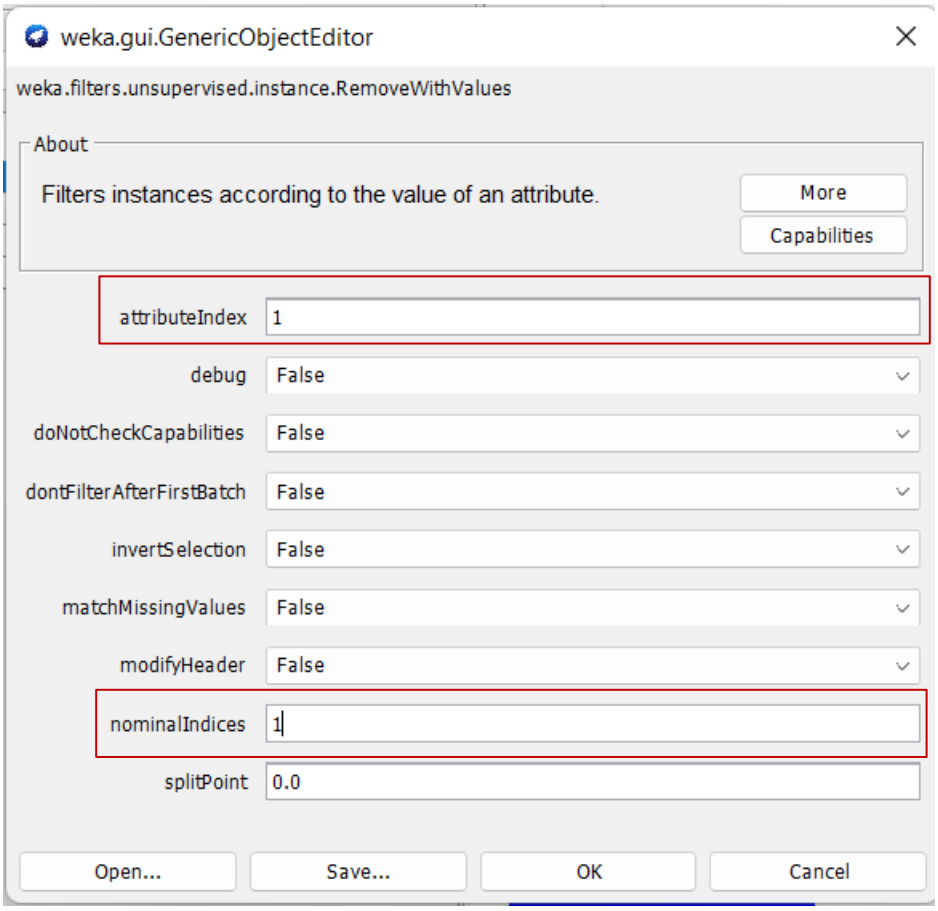
Aplicación de filtros

Para propósitos de este análisis nos interesa evaluar los datos de supervivencia sobre los pasajeros a bordo de la nave con relación a su estatus (clase), por lo que no tendremos en consideración a la tripulación a bordo.

Lo primero que haremos es aplicar un filtro no supervisado a instancias denominado **RemoveWithValue**.



Recordemos que no utilizaremos para la evaluación y análisis los datos referentes a la tripulación, este se encuentra referenciado en la primera fila del atributo Clase, por lo que, al momento de aplicar el filtro, debemos tenerlo en consideración.



En el apartado **attributeIndex** indicamos al atributo que vamos a modificar. El valor 1 hace referencia al atributo clase. Los mismo sucede con el apartado **nominalIndices**, allí indicamos la fila con respecto al atributo que queremos modificar, en nuestro caso sería la fila 1, donde se encuentra ubicada la clase 0 referente a la tripulación. Luego de realizar esta modificación la confirmamos y la aplicamos.

Al aplicarlo, lo primero que vamos a notar es que el número de instancias ha disminuido, y esto se debe a que ya no existe data referente a la clase 0 (tripulación).

Ilustración 1 dataset antes de la aplicación del filtro

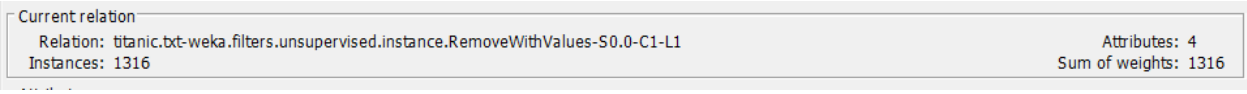


Ilustración 2 dataset luego de la aplicación del filtro

Luego de aplicar el filtrado podemos realizar nuestro análisis de una manera más directa y detallada hacia nuestros objetivos propuestos.

Modelado y análisis de datos

Lo primero que vamos a realizar va a ser una asociación entre nuestros atributos con el objetivo de obtener relaciones entre ellos, por lo que utilizaremos el algoritmo de **Apriori** proporcionado por Weka, con este obtendremos un conjunto de reglas de asociación.

Para esto seleccionamos el algoritmo en la pestaña **Associate** y lo aplicamos.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (197 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 12

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1

Best rules found:

 1. Clase=1 325 ==> Edad=1 319    <conf:(0.98)> lift:(1.07) lev:(0.02) [20] conv:(3.85)
 2. Clase=1 Sobrevivió?=1 203 ==> Edad=1 197    <conf:(0.97)> lift:(1.06) lev:(0.01) [10] conv:(2.4)
 3. Sexo=1 Sobrevivió?=0 694 ==> Edad=1 659    <conf:(0.95)> lift:(1.04) lev:(0.02) [22] conv:(1.6)
 4. Sobrevivió?=0 817 ==> Edad=1 765    <conf:(0.94)> lift:(1.02) lev:(0.01) [15] conv:(1.28)
 5. Sexo=1 869 ==> Edad=1 805    <conf:(0.93)> lift:(1.01) lev:(0.01) [7] conv:(1.11)
 6. Clase=3 Sexo=1 Sobrevivió?=0 422 ==> Edad=1 387    <conf:(0.92)> lift:(1) lev:(-0) [0] conv:(0.97)
 7. Clase=2 285 ==> Edad=1 261    <conf:(0.92)> lift:(1) lev:(-0) [0] conv:(0.94)
 8. Sexo=0 Sobrevivió?=1 324 ==> Edad=1 296    <conf:(0.91)> lift:(1) lev:(-0) [-1] conv:(0.93)
 9. Clase=3 Sexo=1 510 ==> Edad=1 462    <conf:(0.91)> lift:(0.99) lev:(-0) [-5] conv:(0.86)
10. Clase=3 Sobrevivió?=0 528 ==> Edad=1 476    <conf:(0.9)> lift:(0.98) lev:(-0.01) [-8] conv:(0.83)
```

Como podemos observar el algoritmo nos generó un conjunto de reglas con respecto a los atributos que este considero convenientes para la asociación. Cabe destacar que todas las reglas generadas no necesariamente nos resultarían útiles.

1. La **regla 1** nos permite saber que de la primera clase 319 personas eran adultas, por lo que podemos considerar que solo había 6 niños que corresponden a esta clase.
2. La **regla 2** sustenta de los 203 superviviente de primera clase, 197 eran adultos. Por lo que los 6 niños de primera clase los podemos considerar como supervivientes.
3. La **regla 3** define que de las 694 víctimas femeninas, 659 eran adultas. Por lo tanto, también podemos destacar que no sobrevivieron 35 niñas.

4. La **regla 4** señala que no sobrevivieron 765 personas adultas de 817. Con respecto a los infantes no sobrevivieron 52 niños.
5. La **regla 5** señala que 869 mujeres, 805 eran mujeres adultas. Por lo que podemos deducir que 64 eran infantes femeninas.
6. La **regla 6** nos señala que de las mujeres adultas de tercera clase 387 fueron víctimas del accidente. Entonces 35 eran infantes femeninas.
7. La **regla 7** sustenta que de 285 pasajeros de segunda clase 261 eran adultos. Por lo tanto, 24 infantes formaban parte de la tripulación de segunda clase.
8. La **regla 8** define que de 324 superviviente masculinos 296 eran adultos. Sobrevivieron 28 infantes masculinos.
9. La **regla 9** nos señala que de las 510 mujeres de tercera clase, 462 eran adultas. Por lo tanto, habían 48 infantes femeninas en tercera clase.
10. La **regla 10** nos señala que de los 528 pasajeros muertos de tercera clase, 476 eran adultos. Por lo que 52 de los no sobrevivientes de tercera clase eran infantes.

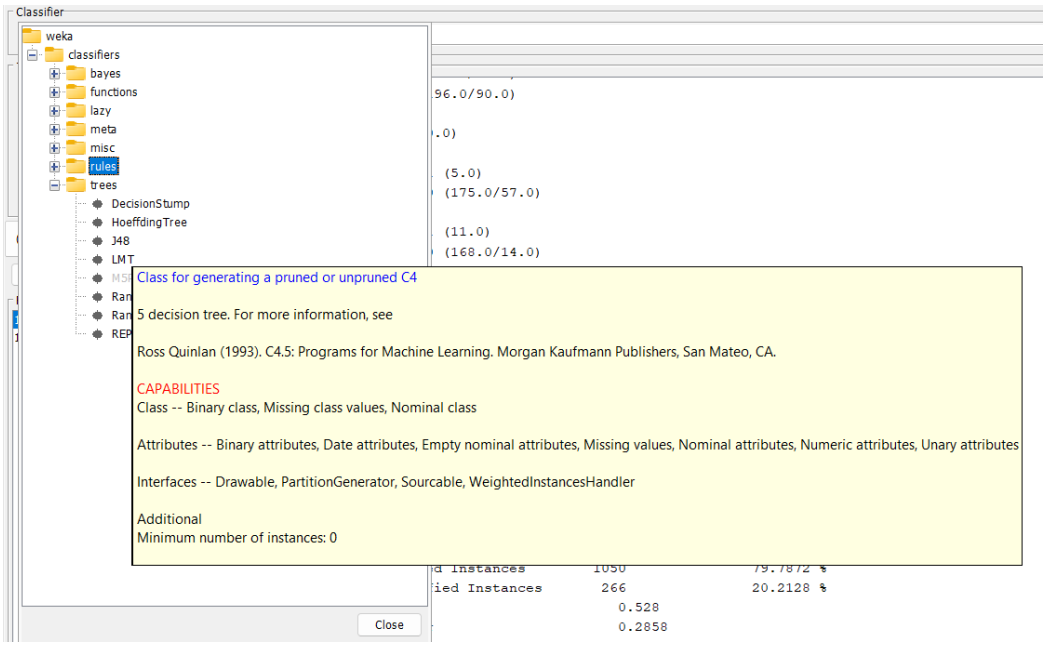
En base a estas métricas podemos deducir lo siguiente:

a. De las 1316 personas a bordo del navío:
    - Personas de primera clase 24,7% (325/1316)
    - Personas de segunda clase 21,7%(285/1316)
    - Personas de tercera clase 53,6% (706/1316)
b. 817 pasajeros no sobrevivieron (62,1% de los pasajeros)

En general, el algoritmo Apriori se encarga de brindarnos cierta información resultante de la evaluación del set de datos pero, esta información no nos resulta ser del todo completa ya que no toma como tal en consideración ciertos valores debido a la descompensación que este produce. Esto lo podemos notar en el caso de edades de los infantes, ya que si volvemos a visualizarlas, ninguna nos brinda información con respecto a las edades de infantes, solo de adultos. Aunque con las reglas generadas ya podemos completar ciertas de estas incógnitas, por lo que podemos afirmar que nos ha resultado útil para evaluar datos.

Ahora, con el objetivo de predecir si la clase afecta en el resultado de si sobrevive o no la persona utilizaremos un árbol de decisión proporcionado por la herramienta Weka.

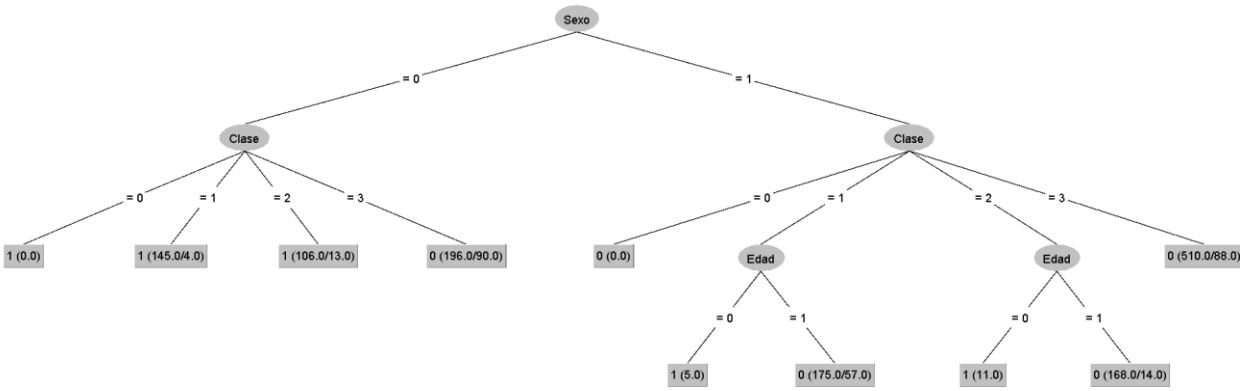
Para esto nos dirigiremos al apartado de clasificación y seleccionaremos el árbol J48, ya que este nos proporciona el porcentaje de error más bajo.



Vamos a cargar los mismos datos de entrenamiento, los seleccionamos, aplicamos e iniciamos. Luego el algoritmo nos generará los resultados:

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances        1050               79.7872 %
Incorrectly Classified Instances       266               20.2128 %
Kappa statistic                          0.528
Mean absolute error                      0.2858
Root mean squared error                  0.378
Relative absolute error                 60.6975 %
Root relative squared error             77.9122 %
Total Number of Instances             1316

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.499    0.763      0.979    0.857      0.579    0.833     0.842     0
                 0.501    0.021    0.936      0.501    0.653      0.579    0.833     0.771     1
Weighted Avg.    0.798    0.318    0.828      0.798    0.780      0.579    0.833     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 800  17 |   a = 0
 249 250 |   b = 1
```

El algoritmo pudo clasificar correctamente el 79.8% de los datos contra un 20.2% de instancias que no pudo clasificar correctamente.



Con esta evaluación podemos deducir lo siguiente:

- Las posibilidades de supervivencia de un pasajero de tercera clase masculino (ya sea adulto o infante) son bajas con respecto a las demás clases.
- Las posibilidades de supervivencia de un pasajero segunda clase son altas, con un 8,2% de posibilidades.
- Las posibilidades de supervivencia de un pasajero de primera clase son bastante alta, con un 36,2% de posibilidades.
- Las posibilidades de supervivencia de un pasajero femenino de tercera clase (independientemente de la edad) son relativamente bajas.
- Las posibilidades de supervivencia de un pasajero femenino de segunda clase dependerán de su edad; si es infante tiene más posibilidades de supervivencia por sobre el adulta.
- Ocurre lo mismo dentro de la primera clase, dependerá de la edad del pasajero.

Al realizar evaluaciones utilizando este algoritmo podemos dictaminar, que no es lo más conveniente el uso de árboles de decisiones ya que los datos que están suministrados en el dataset no pueden ayudar a determinar con seguridad este escenario.

Visualización de los resultados
Usaremos la herramienta Tableau para la visualización de los datos:

	A	B	C	D
1	Clase	Edad	Sexo	Sobrevivió?
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	1	1
17	1	1	1	1
18	1	1	1	1
19	1	1	1	1
20	1	1	1	1
21	1	1	1	1
22	1	1	1	1
23	1	1	1	1
24	1	1	1	1
25	1	1	1	1
26	1	1	1	1
27	1	1	1	1
28	1	1	1	1

Ilustración 3 Contenido del archivos.csv para su posterior traspaso a Tableau

Ilustración 4 Gráfico que representa la cantidad de pasajeros por clases, y datos de supervivencia correspondiente a cada clase

El grafico nos muestra la cantidad de pasajeros con respecto a las clases. Y los colores azul y celeste nos indica la supervivencia de los pasajeros. Celeste indica el número de personas que no sobrevivieron al accidente y azul el número de personas que sí sobrevivieron.

Esto lo podemos verificar con el gráfico generado por la aplicación Weka, el cual se muestra a continuación:

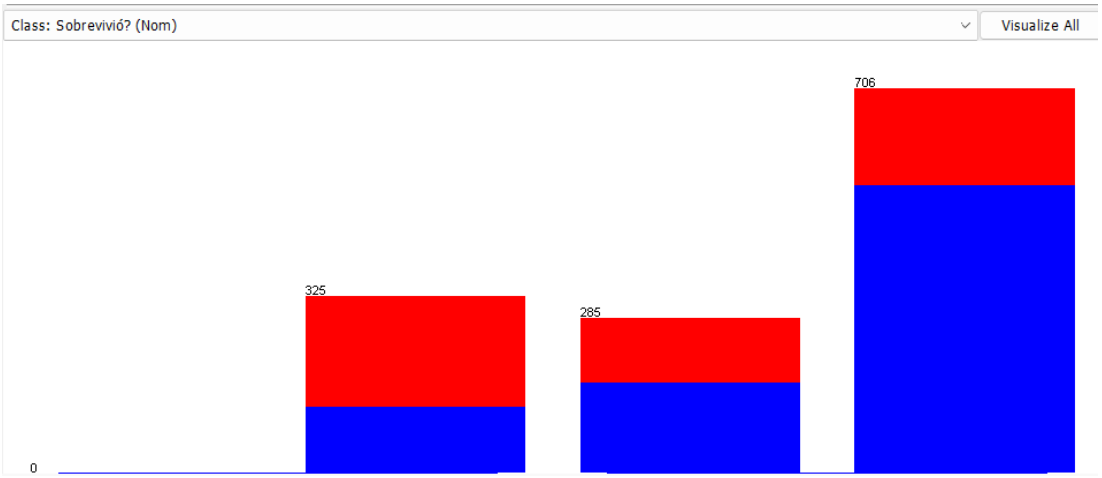


Ilustración 5 Gráfico de supervivencia representado en la Herramienta Weka

Con ayuda del gráfico generado en Tableau podemos simplificar la información y la podemos hacer sencilla a la vista.

Primera clase: El 62,5% de los pasajeros de primera clase sobrevivieron, por encima del 37,5% de los pasajeros que no siguieron la misma suerte.

Segunda clase: El 41,4% de los pasajero de primera clase sobrevivieron, por debajo 51,4% de los pasajeros que no sobrevivieron.

Tercera clase: El 25,2% de los pasajeros de primera clase sobrevivieron, por debajo del 74,8% de los pasajeros que no sobrevivieron.

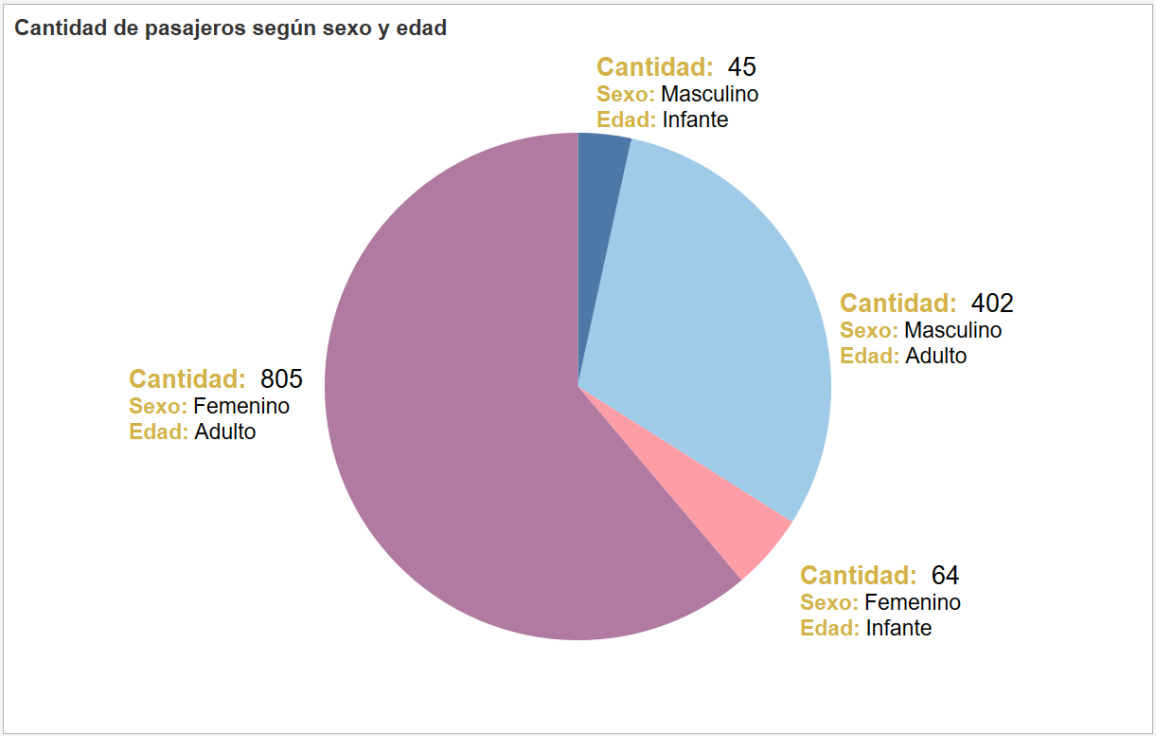


Ilustración 6 Gráfico que representa la cantidad de pasajeros según el género y la edad

El siguiente gráfico no muestra gran relevancia con respecto a nuestra conclusión, pero nos detalla un poco más la situación general con respecto a los pasajeros a bordo:

- El 92,6% de las mujeres a bordo eran adultas y el 7,4% eran infantes.
- El 90,0% de los varones a bordo eran adultos y el 10,0% eran infantes.

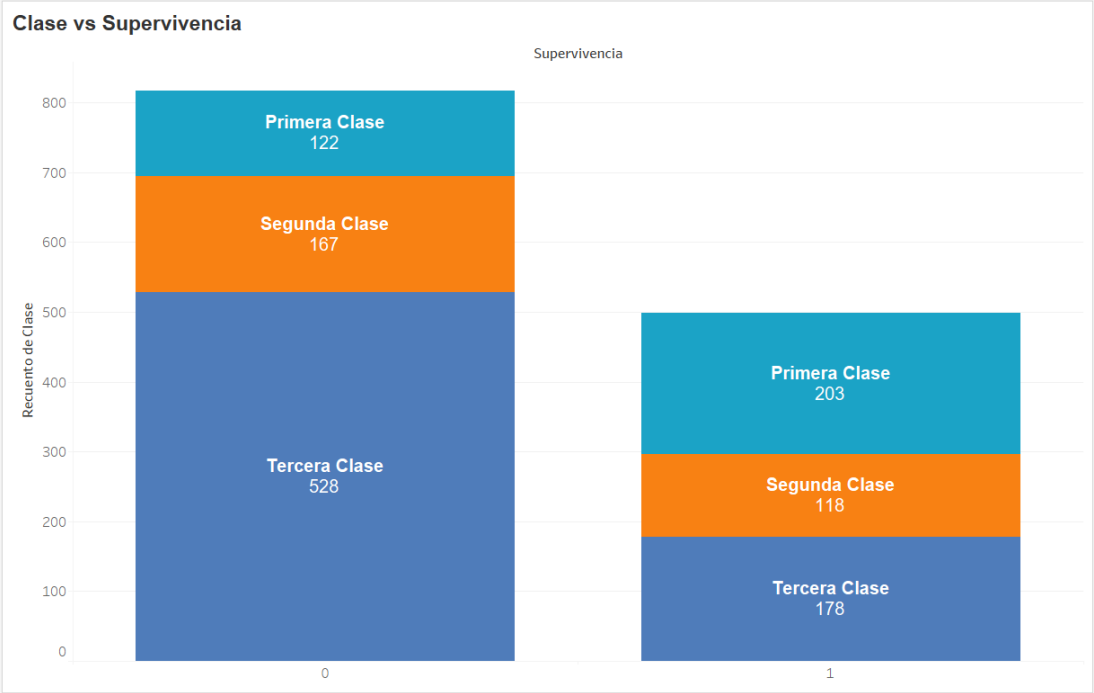


Ilustración 7 Gráfico de Clase vs Supervivencia

Con este gráfico basta para concluir nuestra duda sobre que la supervivencia de personas dependiendo de su clase. Al visualizar el gráfico se puede deducir que la mayor parte de los supervivientes con respecto a su cantidad total:

- Los pasajeros de primera clase encabezan con el 62,5% de supervivencia de la clase.
- Los pasajeros de segunda clase siguen con el 41,4% de supervivencia.
- Y los pasajeros de tercera clase 25,2% de supervivencia.

De manera general:

- 40,7% corresponden a los pasajeros de primera clase; 23,7% a los pasajeros de segunda clase; y 35,7% a los pasajeros de tercera clase.

De manera general podemos pensar que sobrevivieron muchos más pasajeros de tercera clase por sobre los de segunda, pero esto no tiene relevancia. Debido a que el que se encarga de definir es el porcentaje de supervivencia por sobre la cantidad de pasajeros correspondiente a la clase.

Por lo tanto la primera clase sería fue la más propensa a sobrevivir (aunque era de esperarse). Ver la película me sirvió de algo profe. No me fracase pls :3

2. Conjunto de datos Drug1n.arff

Descripción del dataset

Este dataset contiene información sobre un conjunto de pacientes, los cuales pudieron haber sufrido de la misma enfermedad. Los datos contenidos en el dataset nos servirán para determinar cuál es el medicamento más adecuado para los futuros pacientes.

Información de dataset:

a. Número de instancias: 200
b. Número de atributos: 7

Los atributos correspondientes para este conjunto de datos son:

a. **Age:** contiene la edad de cada uno de los pacientes (edades contenidas de 15 – 74).
b. **Sex:** genero del paciente, el cual puede ser masculino (M) o femenino (F).
c. **BP:** referido a la presión sanguínea. Los datos contenidos son declarados de la siguiente manera: HIGH (alta) o NORMAL (normal).
d. **Cholesterol:** indica los niveles de colesterol en sangre del paciente. Los datos son clasificados de la siguiente forma: HIGH (alta), LOW (baja) o NORMAL (normal).
e. **Na:** concentración de sodio en sangre.
f. **K:** concentración de potasio en sangre.
g. **Drug:** Medicamente prescrito al que respondió el paciente. Los medicamentos han sido nombrados de la siguiente manera: drugA, drugB, drugC, drugX o drugY.

Referencias: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=tutorial-drug-treatments-exploratory-graphsc50
http://dit.upm.es/~gfer/ssii/PracticaIntroWeka.html

Preparación de datos:

Se procede a realizar una evaluación de la data proveída mediante la opción de “edit” proporcionada por la herramienta Weka con el fin de verificar si había data errónea o faltante dentro del dataset. Una vez confirmado que la data se encuentra completa y correcta, procede a continuar con el análisis de este.

Aplicación de filtros

Para propósitos de analizar esta data, se estudió la relación que existe sobre el sodio y potasio en la sangre, los cuales están medidos por electrolitos. Se procedió a unificar ambos datos en una sola instancia mediante el filtro “Addexpression”. Esto se justifica al visualizar la relación entre ambos está separada de forma lineal a través de todos sus valores.

Modelado de datos y análisis de datos

En este dataset se busca analizar el mejor medicamento de entre los 5 medicamentos en total para una enfermedad X. Dada la cantidad de fármacos, se procede a realizar una primera visualización sin modificar ni agregar atributos con el fin de observar la predicción del software en este caso. Para esto nos dirigimos a la pestaña **Classify** y escogemos de los classifier, tree y luego J48. Se le deja con los valores predeterminados y se ejecuta el algoritmo.

A continuación, se observan los siguientes resultados:

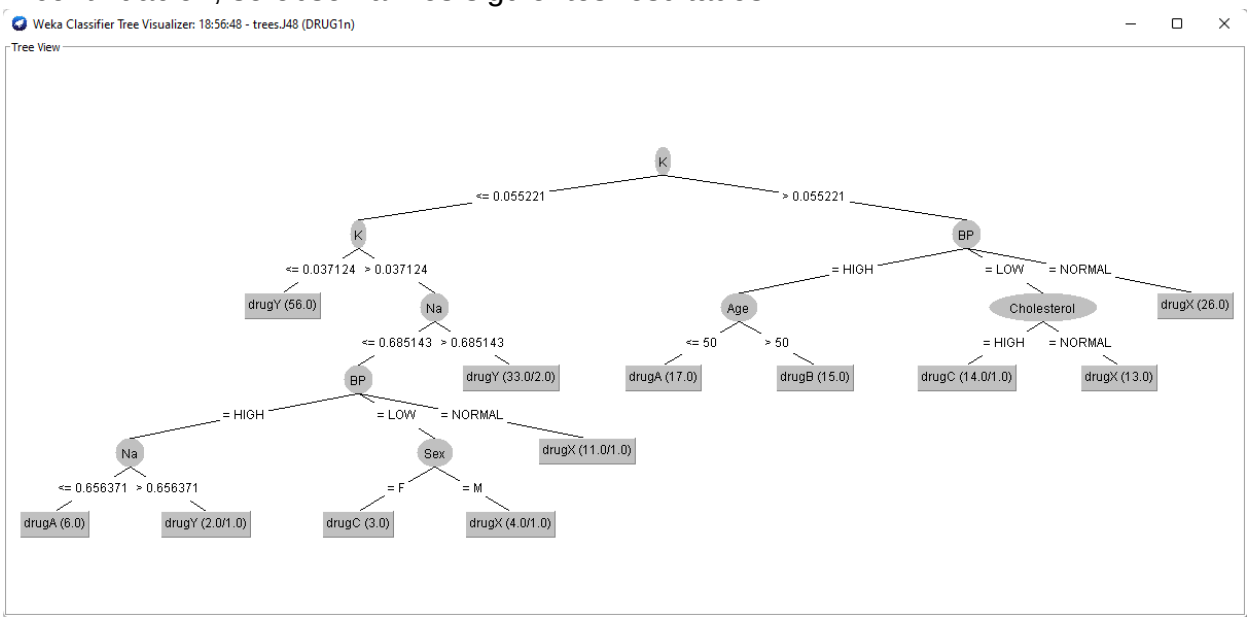


Ilustración 8. Árbol de decisión J48 generado sin modificaciones en el dataset.

Este árbol se generó entrenado con la data original y ha arrojado una precisión del 97% de un total de 200 instancias.

Dada la complejidad al visualizar este árbol, se procede a aplicar el filtro que añade el atributo que relaciona el sodio (Na) y el potasio (K) con el fin de obtener un árbol de decisión más comprensible a la vista. Como se mencionó en el apartado de aplicación de filtros, esto se justifica al observar el gráfico de dispersión generado en la sección Visualize el cual muestra lo siguiente:

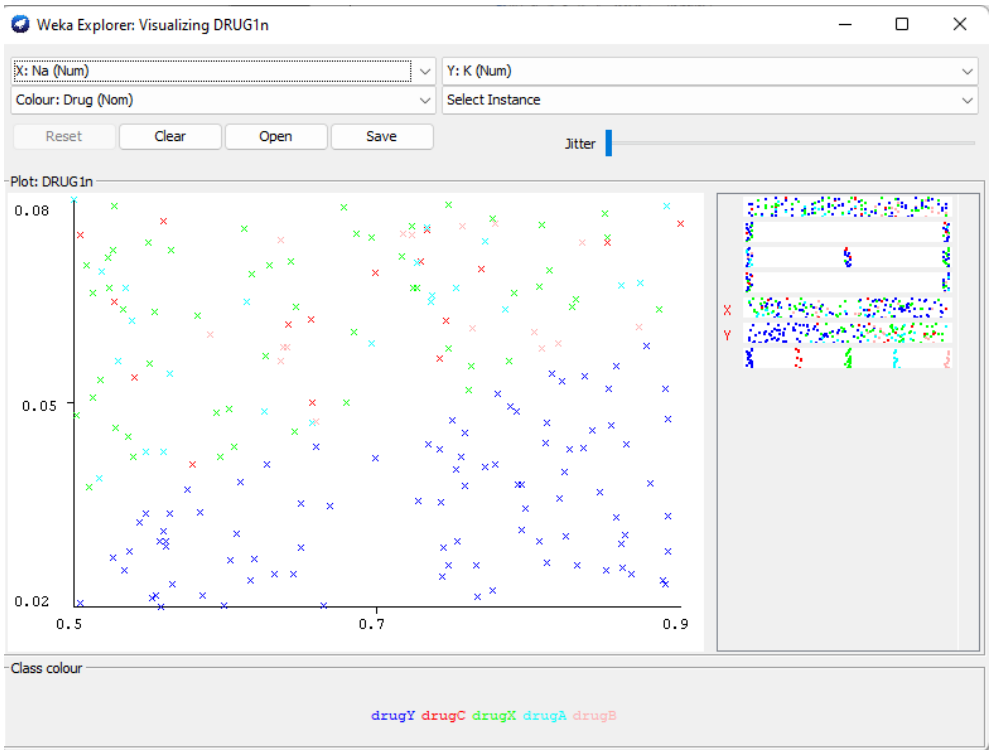


Ilustración 9. Gráfico de dispersión que muestra la relación Sodio-Potasio.

Al ver este gráfico podemos ver que existe una línea que separa al sodio del potasio de manera casi proporcional. Además, podemos deducir que el medicamento Y es más efectivo en casos donde el sodio se encuentra por debajo del rango normal y viceversa en el caso del medicamento X.

Habiendo visualizado lo dicho anteriormente, se procede a aplicar el filtro para agregar un nuevo atributo por medio de la opción **filter**, luego unsupervised, attribute, AddExpression:

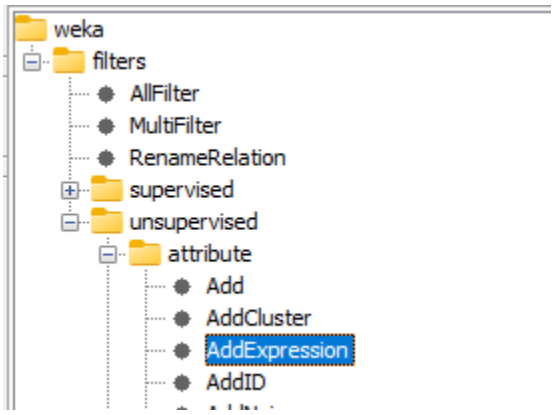


Ilustración 10. Aplicación de filtro para añadir un nuevo atributo en base a una expresión.

Posteriormente se modifican los valores del filtro de acuerdo con lo requerido:

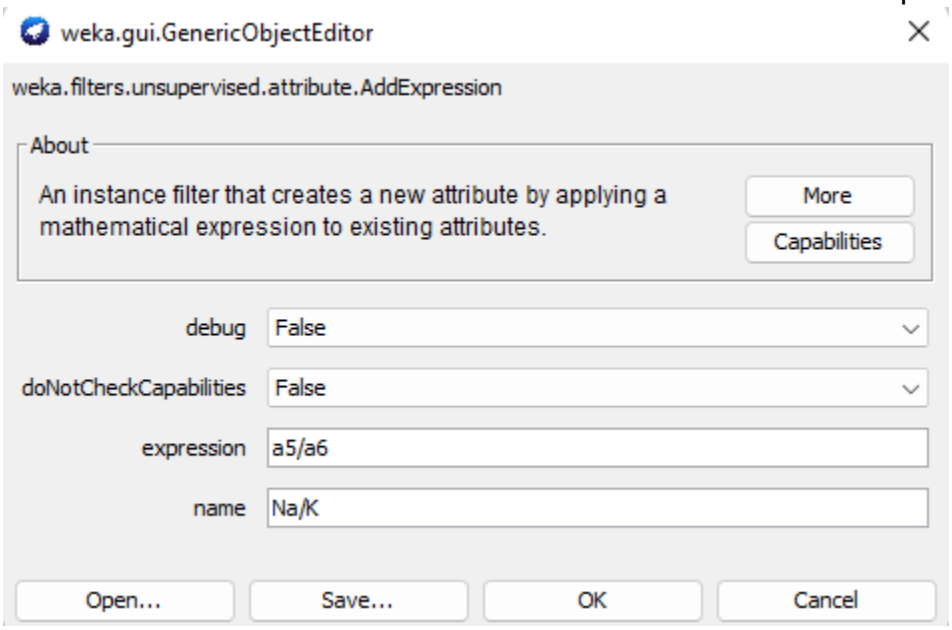


Ilustración 11. Pantalla de modificadores que ofrece el filtro con los parámetros a establecer

Se agrega al apartado de expresión a5/a6 correspondiente a la expresión Na/K el cual se coloca en el nombre del nuevo atributo. Por último, se aplica el filtro y se obtiene los siguientes resultados:

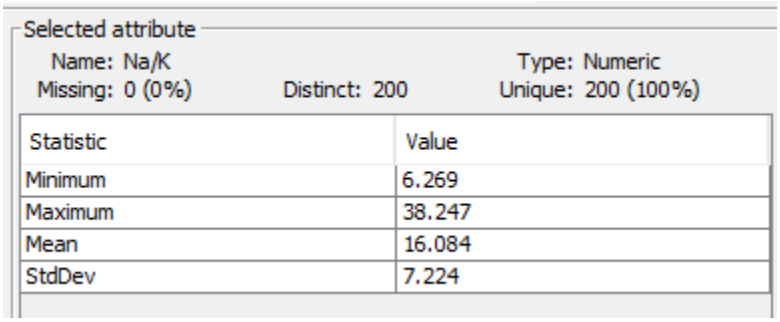


Ilustración 12. Datos estadísticos generados producto de la fusión entre los atributos sodio y potasio.

Una vez obtenida la nueva instancia, se procede a generar un nuevo árbol J48:

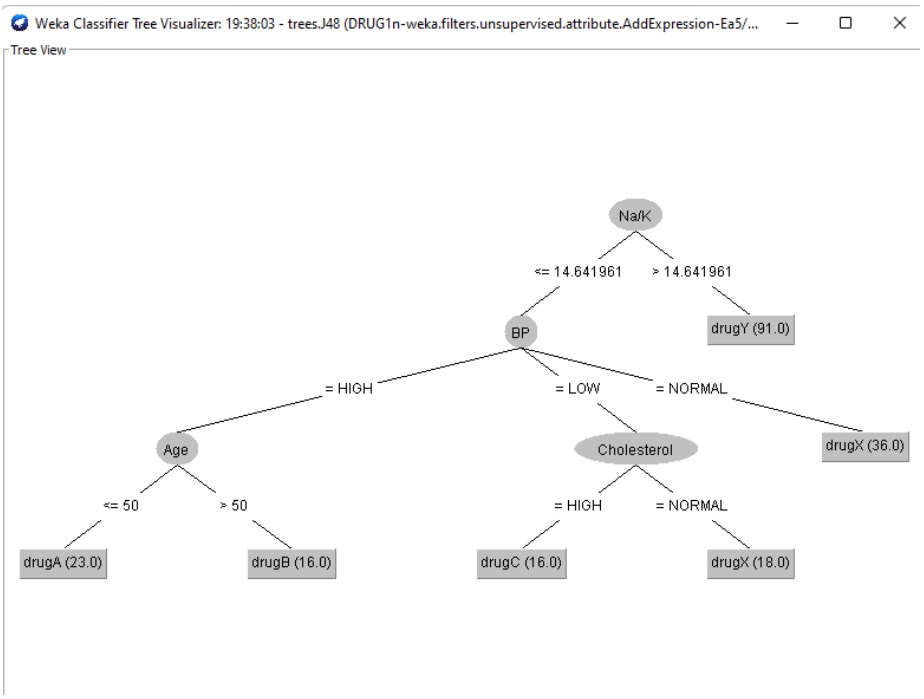


Ilustración 13. Árbol de decisión J48 generado con el nuevo atributo en cuenta.

Se puede observar un árbol más sencillo para propósitos de análisis en el que podemos ver claramente en base a que valores de los distintos atributos que se manejan, se puede iniciar el tratamiento para la enfermedad en estudio con el medicamento indicado para ese escenario. Cabe destacar que este árbol de decisión fue generado con una precisión del 100%.

Visualización de los resultados

Para propósitos de cumplir con lo dispuesto al inicio de este análisis, se procederá a analizar los atributos con respecto a los medicamentos.

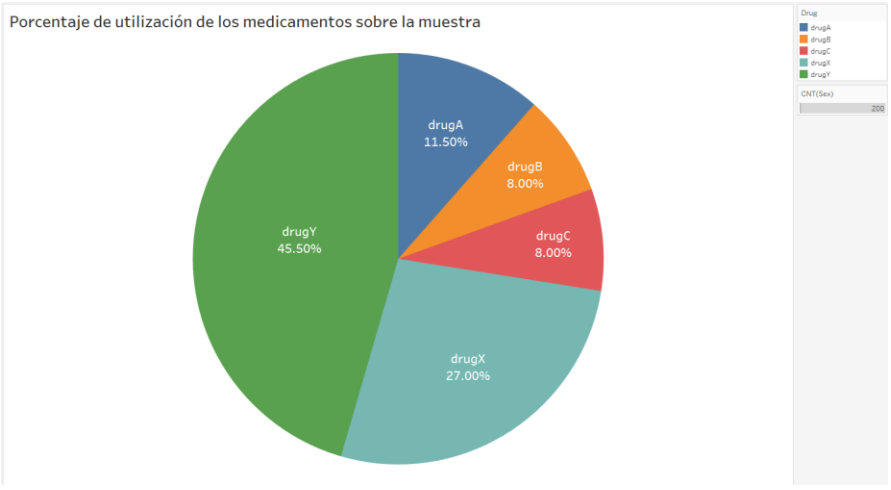


Ilustración 14. Gráfico de pastel sobre la utilzación de cada medicamento con respecto a la muestra.

En este gráfico de pastel podemos observar el porcentaje de utilización por medicamentos sobre la muestra. En este, podemos observar que el medicamento Y tuvo una mayor eficiencia en comparación sobre los demás medicamentos, casi doblando el porcentaje de utilización del medicamento X. Los demás medicamentos tuvieron un porcentaje menor, 11.5% para el medicamento A y 8% para los medicamentos B y C; estos han de haber tenido menor eficiencia sobre las personas en estudio o tuvieron su escenario especial en el que superaron la eficiencia de los demás medicamentos.

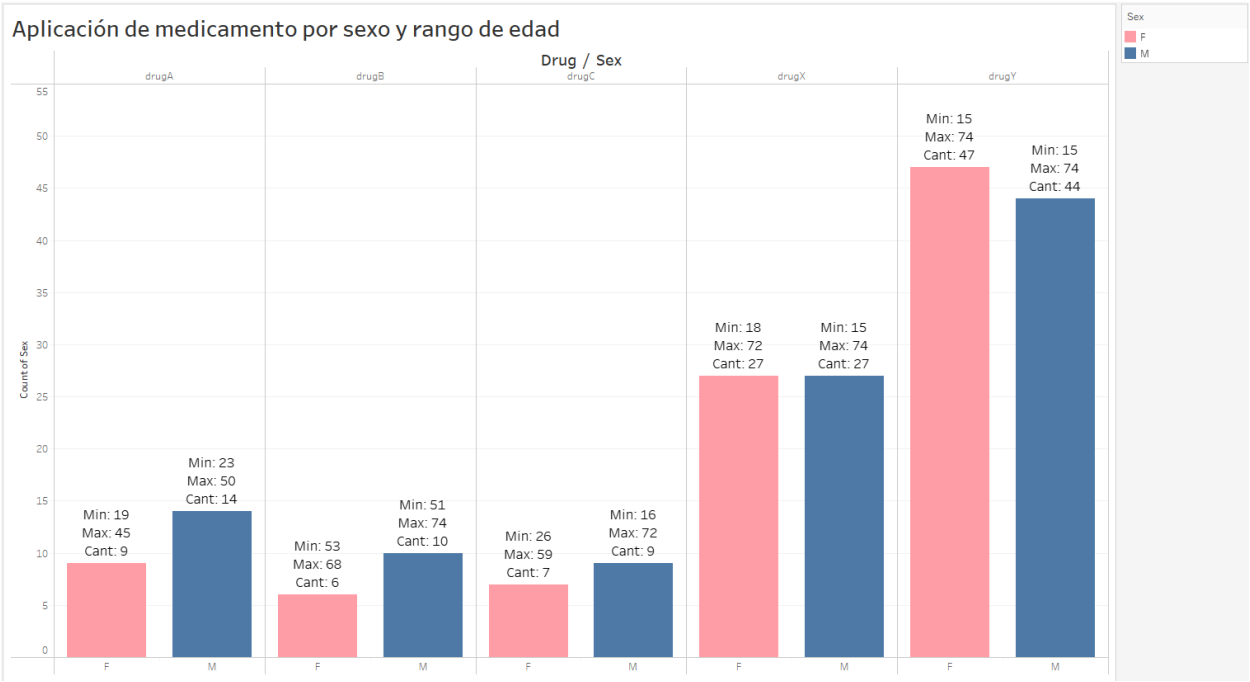


Ilustración 15. Gráfico de barras que muestra la cantidad de hombres como mujeres y su rango de edad por medicamento.

En este gráfico podemos visualizar el rango de edad y sexo en los que fueron aplicados cada medicamento. Podemos distinguir que:

a. El medicamento Y cubrió el 100% del rango de edad establecido en este estudio y se aplico en un total de 91 personas (47 femeninas y 44 masculinos).
b. El medicamento X fue aplicado en igual cantidad de personas (27 en ambos sexos) haciendo un total de 54 personas. Este solo cubrió el 100% del rango de edad para los masculinos; en femeninas solo fue aplicado en personas entre los 18 y 72 años.
c. El medicamento A fue el tercero más aplicado, 9 femeninas y 14 masculinos, haciendo un total de 23 personas. Este medicamento fue aplicado en jóvenes y adultos: mujeres entre 19 – 45 años y hombres entre 23 – 50 años.

d. El medicamento C también fue aplicado en personas entre los 16 y 72 años (26 – 59 años en mujeres y 16 – 72 años en hombres) haciendo un total de 16 personas que recibieron este medicamento.
e. Por último, el medicamento B, el cual tiene la peculiaridad de haber sido suministrado a personas de edad mayor, específicamente 51 – 74 años (53 - 68 años en mujeres y 51 – 74 años en hombres) haciendo un total de 16 personas.

Escenario médico de aplicación por medicamento

Cholesterol	BP	Drug drugA Median K	drugA Median Na	drugB Median K	drugB Median Na	drugC Median K	drugC Median Na	drugX Median K	drugX Median Na	drugY Median K	drugY Median Na
HIGH	HIGH	0.0599	0.5943	0.0584	0.6870					0.0291	0.7736
	LOW					0.0668	0.6768			0.0378	0.7927
	NORMAL							0.0644	0.6351	0.0354	0.7559
NORMAL	HIGH	0.0648	0.7244	0.0736	0.7575					0.0344	0.7421
	LOW							0.0649	0.7000	0.0281	0.7499
	NORMAL							0.0667	0.5918	0.0399	0.7785

Ilustración 16. Tabla de información que muestra los escenarios en los que se suministraron cada medicamento.

Finalmente, analizamos el escenario médico en el que se aplicó cada medicamento. Por propósitos de interpretación sobre el sodio y el potasio, se procedió a estudiar sus niveles en un rango de 0 a 1 para el sodio y 0.00 – 0.1 para el potasio y se utilizó la media de cada uno aplicado a los escenarios de los 5 medicamentos. Esto debido al hecho de que estos no están en las unidades establecidas para la medición de electrolitos en la sangre el cual es mEq/l (miliequivalentes/litro). Como resultado, se obtuvo lo siguiente:

a. El medicamento A fue aplicado en personas que tuvieron niveles de colesterol tanto normal como alto, presión sanguínea alta y la media de los valores de sodio y potasio fueron 0.59 y 0.58 para aquellos con presión sanguínea alta y

para quienes tuvieron la presión sanguínea baja, 0.064 de potasio y 0.72 de sodio.

b. Para el medicamento B se presentaron los mismos escenarios con respecto al nivel de colesterol y presión sanguínea. Para el escenario de colesterol y presión sanguínea alta se presentaron niveles de sodio en 0.68 y potasio en 0.05 y para su variación con nivel de colesterol normal, los niveles de sodio y potasio fueron 0.75 y 0.073 respectivamente.
c. El medicamento C solo fue utilizado en un escenario, el cual fue colesterol alto y presión sanguínea baja y cuyos valores de sodio y potasio fueron de 0.67 y 0.066.
d. El medicamento X presentó un total de tres escenarios:
    a. Escenario 1:
        i. Colesterol alto.
        ii. Presión sanguínea baja.
        iii. Sodio: 0.63
        iv. Potasio: 0.064
    b. Escenario 2:
        i. Colesterol normal.
        ii. Presión sanguínea baja.
        iii. Sodio: 0.7
        iv. Potasio: 0.064
    c. Escenario 3:
        i. Colesterol normal.
        ii. Presión sanguínea normal.
        iii. Sodio: 0.59
        iv. Potasio: 0.066
e. El medicamento Y presentó aplicación en todos los escenarios posibles y cuyos valores fueron los siguientes:
    a. Escenario 1:
        i. Colesterol alto.
        ii. Presión sanguínea alta.
        iii. Sodio: 0.77
        iv. Potasio: 0.029
    b. Escenario 2:
        i. Colesterol alto.
        ii. Presión sanguínea baja.
        iii. Sodio: 0.79
        iv. Potasio: 0.037
    c. Escenario 3:
        i. Colesterol alto.
        ii. Presión sanguínea normal.
        iii. Sodio: 0.75
        iv. Potasio: 0.035

d. Escenario 4:
    i. Colesterol normal.
    ii. Presión sanguínea alta.
    iii. Sodio: 0.74
    iv. Potasio: 0.035
e. Escenario 5:
    i. Colesterol normal.
    ii. Presión sanguínea baja.
    iii. Sodio: 0.74
    iv. Potasio: 0.028
f. Escenario 3:
    i. Colesterol normal.
    ii. Presión sanguínea normal.
    iii. Sodio: 0.77
    iv. Potasio: 0.039

Podemos concluir en este análisis que el medicamento Y tuvo mayor aplicación que el resto de los medicamentos. Esto habla de su eficiencia, pero no quiere decir que fue altamente probado en todos los escenarios posibles. Esto se demuestra al observar los valores de sodio y potasio, los cuales deben están dentro de un rango establecido por la medicina como valores normales.

Podemos observar en la tabla de escenario médico de aplicación por medicamento que el medicamento Y tuvo aplicación en un escenario completo con respecto al nivel de colesterol y presión sanguínea pero la porción de la muestra a la que le fue suministrada el medicamento Y mostró un desequilibrio pronunciado en los niveles de sodio y potasio, específicamente potasio bajo (0.02 – 0.03) y sodio alto (0.74 – 0.79).

El medicamento X solo se aplicó en escenarios donde el nivel de colesterol fue tanto alto como bajo, pero los niveles de presión sanguínea fueron normales o bajos. Para estos escenarios los niveles de potasio y sodio fueron medianamente altos.

El medicamento C, A y B se pueden considerar como medicamentos para situaciones específicas. Para el medicamento C solo se mostró un escenario donde el nivel de colesterol fue bajo y los niveles de sodio y potasio fueron ligeramente altos. Los medicamentos A y B se aplicaron en una situación donde el colesterol y la presión sanguínea fueron altos. La diferencia entre estos es que para el medicamento B, con colesterol alto, su nivel de sodio mostró un valor ligeramente alto pero el nivel de potasio estaba rozando el límite de sus valores normales. En el caso de colesterol normal, ambos, tanto sodio como potasio mostraron niveles altos.

En cuanto al medicamento A, sus niveles de sodio y potasio con presión sanguínea alta fueron ligeramente altos y para el caso de colesterol normal fueron ligeramente altos para potasio y altos para el sodio.

Por lo tanto, definir un medicamento eficiente para todos los escenarios no es posible, dado el hecho que, como se ha explicado en el párrafo anterior, cada uno tiene sus escenarios específicos, siendo el medicamento Y el utilizado para un caso de un grave desequilibrio entre estos factores de estudios seleccionados dentro de este dataset.

3. Conjunto de datos Waveform-5000.arff

Descripción del dataset

Este dataset contiene información donde cada clase se genera a partir de una combinación de 2 de 3 ondas "base" y cada instancia se genera con ruido agregado (media 0, varianza 1) en cada atributo.

Información de dataset:

a. **Cantidad de instancia:** 5000
b. **Cantidad de Atributos:** 41

Desde el atributo x1 a x40 son ruidos y el atributo clase se clasifica en: 0,1,2

Preparación de datos

Primeramente visualizamos que la data no se encuentre errónea o con valores en atributos faltantes, además del tipo de datos que se manejara para obtener análisis claros en nuestros resultados utilizando weka.

Vemos que los datos no son consistentes para obtener un análisis concreto, por lo tanto aplicaremos filtros.

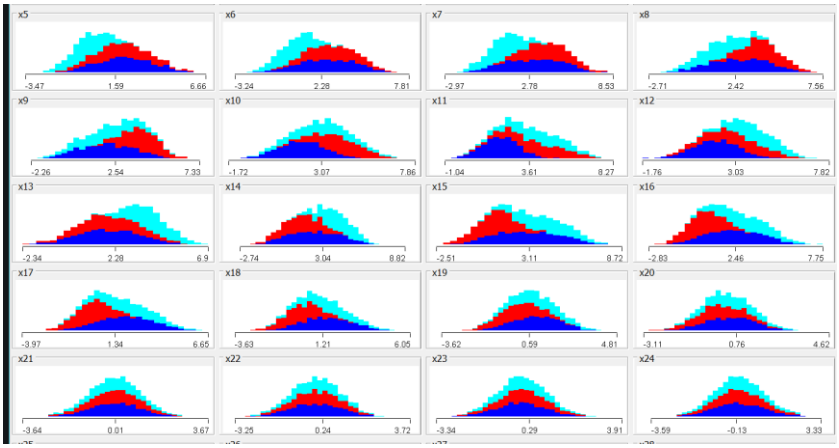


Ilustración 17: visualización del dataset

Aplicación de filtro

No es necesario aplicar filtro por ahora, para tener una clasificación mas precisa con la data original

Modelado y análisis de datos

Con este dataset preparado queremos conocer por clase (0,1,2) se clasifican los sonidos (instancias), por lo tanto estaremos usando el método de clasificación.

Usando la herramienta weka, iremos a la sección classify y escogemos el algoritmo de árbol J48

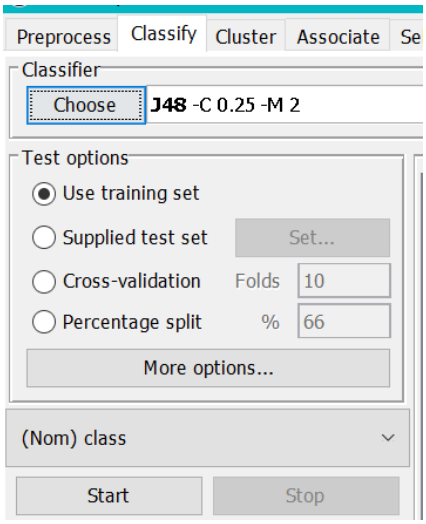


Ilustración 18: características del algoritmo J48

Ejecutamos con las características predeterminadas y analizamos los resultados.

Classifier output

```
Number of Leaves  :     247

Size of the tree :      493


Time taken to build model: 0.59 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances        4931               98.62   %
Incorrectly Classified Instances        69                1.38   %
Kappa statistic                          0.9793
Mean absolute error                      0.0167
Root mean squared error                  0.0915
Relative absolute error                  3.7682 %
Root relative squared error             19.4117 %
Total Number of Instances             5000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.006    0.988      0.982    0.985      0.977    0.998     0.996     0
                 0.984    0.006    0.987      0.984    0.986      0.979    0.999     0.997     1
                 0.993    0.008    0.984      0.993    0.988      0.982    0.999     0.996     2
Weighted Avg.    0.986    0.007    0.986      0.986    0.986      0.979    0.998     0.996
```

Ilustración 19: Resultado de algoritmo J48

Vemos que ha clasificado correctamente 98% de las instancias (4931) obtenidos en un numero de hojas de 247.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 1661   15   16 |    a = 0
   15 1627   11 |    b = 1
    6    6 1643 |    c = 2
```

Ilustración 20: matriz de confusion

podemos visualizar en la matriz de confusión, la diagonal principal las instancias bien clasificadas por clase.

Y alrededor los mal clasificado las cuales se puede deducir:

a. **Clase 0:** donde 15 instancias perteneciente a la clase1 y 6 instancias perteneciente a la clase2 se clasifico erróneamente.
b. **Clase 1:** donde 15 instancias perteneciente a la clase0 y 6 instancias perteneciente a la clase2 se clasifico erróneamente.
c. **Clase 2:** donde 16 instancias perteneciente a la clase0 y 11 instancias perteneciente a la clase1 se clasifico erróneamente.

Ilustración 21: visualizacion de resultado

Visualizamos por medio del margen de predicion(Y) dependiendo de las clases(X), donde se ve que las instancias mal clasificadas se encuentran debajo de el margen de 0.0025

Ahora, para obtener mejor interpretación en los resultados, aplicamos **discretizar** en los datos numéricos.

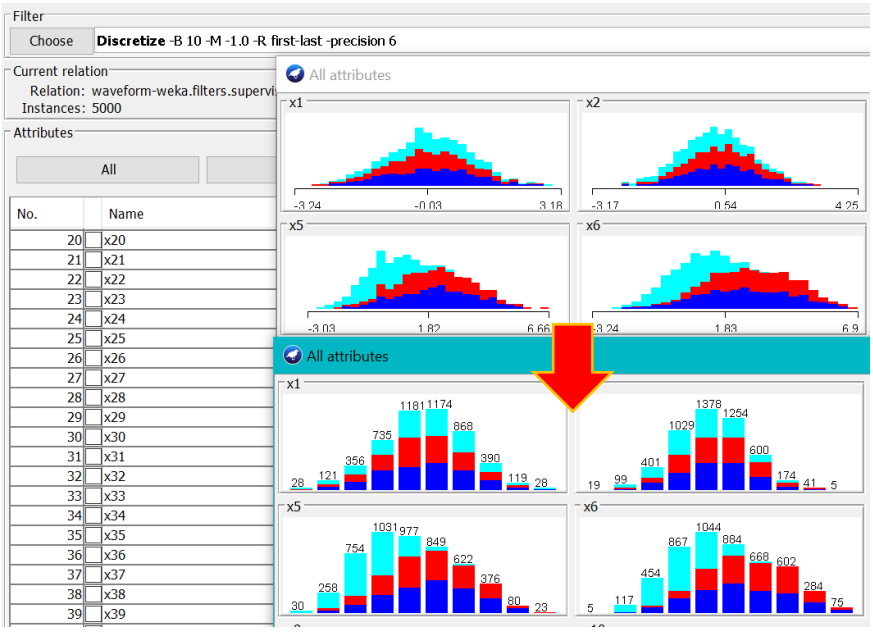


Ilustración 22: uso del filtro "discretizar" para visualizar mejor los datos.

Visualización de los resultados

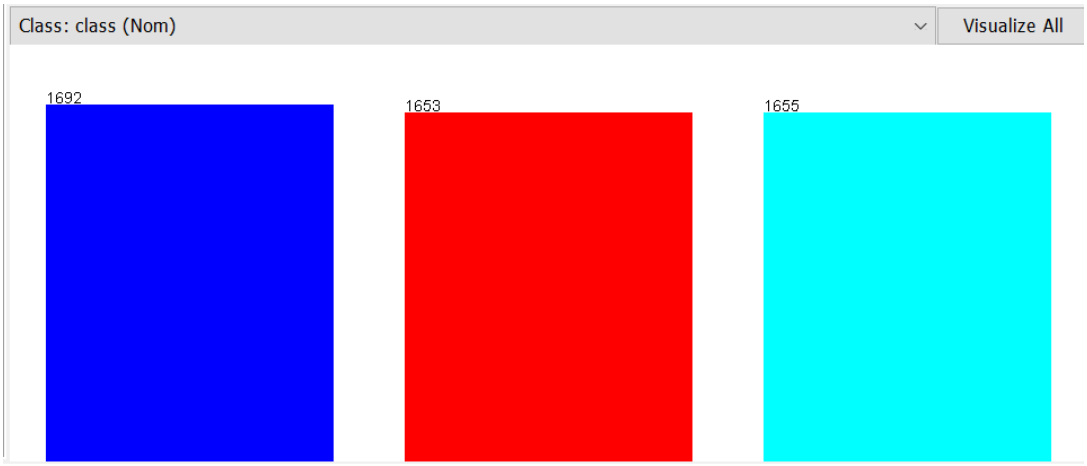


Ilustración 23: instancias por clases

En la **clase0** tenemos 1692 instancias, en la **clase1** tenemos 1653 instancias y en la **clase2** tenemos 1655 instancias clasificadas

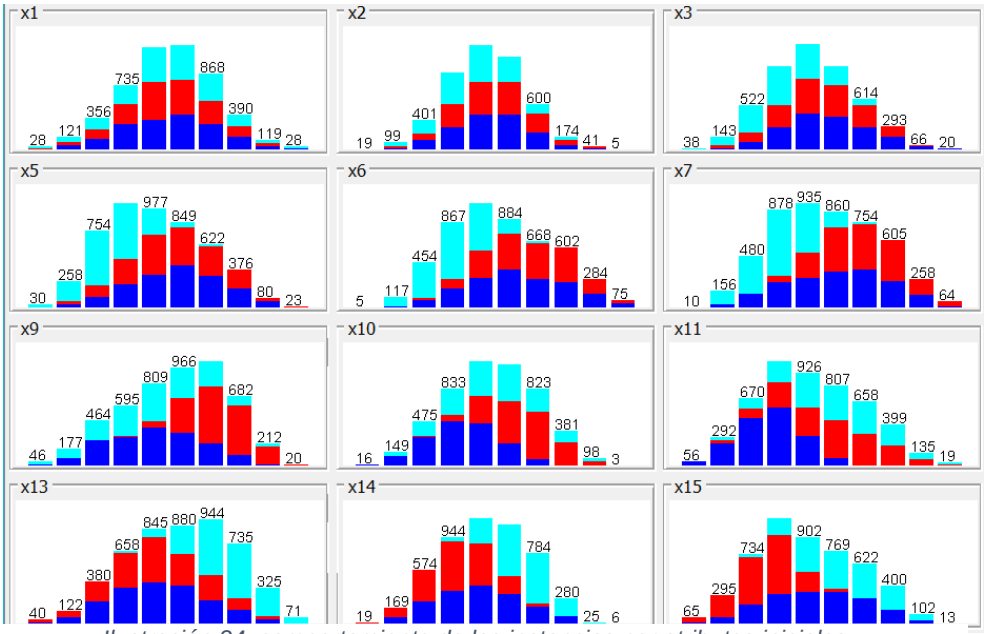


Ilustración 24: comportamiento de las instancias por atributos iniciales.

Podemos visualizar de manera mas clara que el comportamiento de las ondas desde los atributos x1 hasta x19, donde los sonidos de clase2 se encuentra muy a la derecha de la media, mientras que los sonidos de clase1 se encuentran al lado izquierdo de la media y los sonidos de clase0 mantienen su distribución estándar alrededor de la media.

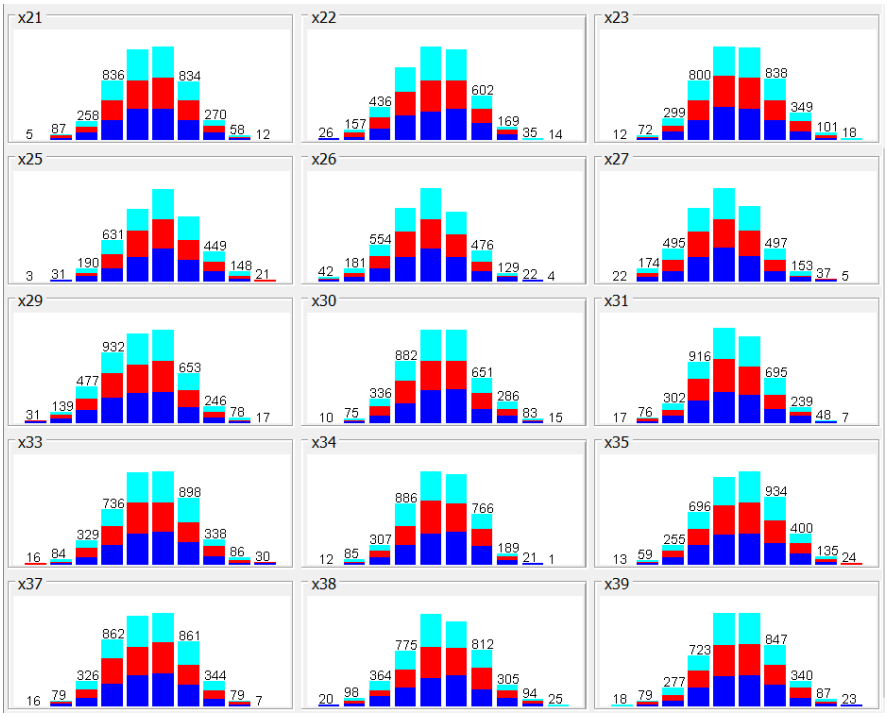


Ilustración 25: Comportamiento de las instancias por atributos finales.

Sin embargo, para las ondas de x21 hasta x40, tanto la clase2, clase1 y clase0 sus instancias están dentro de la distribución estándar.

4. Conjunto de Datos Vehicle.arff

Descripción del dataset

Vehicle.arff es un set de datos numéricos referente a siluetas de vehículos. Este dataset fue generado con el propósito de encontrar un método para distinguir objetos 3D en una imagen 2D mediante la aplicación de un conjunto de extractores de características de forma para siluetas 2D de un objeto. Para el experimento se utilizaron una serie de vehículos particulares de tipo “corgie”: un bus de dos pisos, una van Chevrolet, Saab 9000 y un Opel Manta 400. Se tomaron un set fotografías de los modelos seleccionados en base a un ángulo de visión fijo de 34,2 grados en orientación horizontal y posteriormente rotándolos para documentar distintos puntos del vehículo como de frente y de espalda (0-180 grados), perfiles en direcciones opuestas (90-270 grados), con el fin de obtener una imagen de 360 grados para cada uno. Además, fueron colocados sobre una superficie difusa y fueron pintados de negro mate para minimizar el nivel de reflejo.

a. Cantidad de instancias: 846
b. Cantidad de atributos: 19

Preparación de Datos

Primeramente se hace una revisión de los tipos de datos con los que trabajaremos, si este no tiene basura, vacíos o datos erróneo. Se evalua que la data no necesita ningún tipo de filtración por lo tanto, proseguimos al modelado para estudiar la información.

Modelado y análisis de datos

Para el estudio de esta datas utilizaremos cluster para hacer un estudio por agrupamiento y obtener un mejor análisis de los resultados. Utilizaremos el algoritmo de SimpleK-means

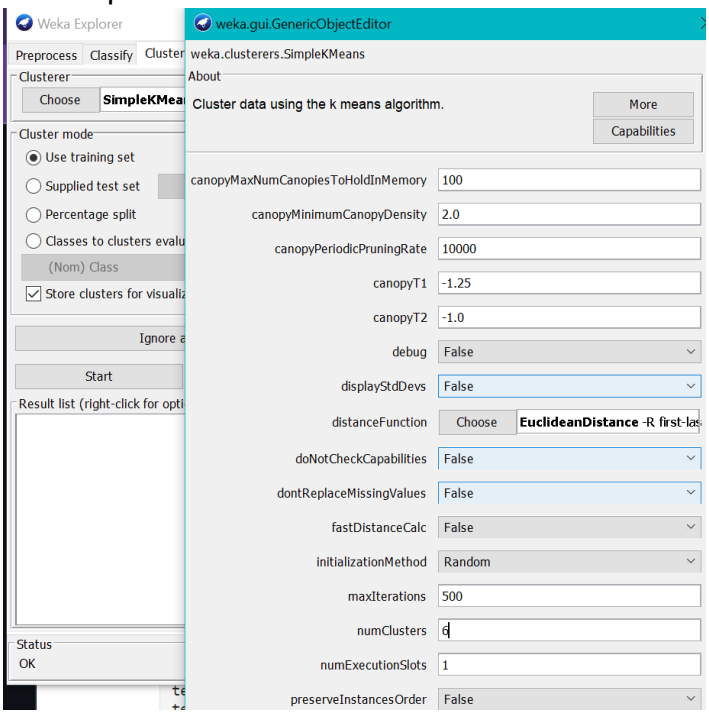


Ilustración 26: características predeterminadas del algoritmo SimpleKmeans

Donde el número de clústeres a utilizar será de 6, usando set de entrenamiento.

Clusterer output

```
Number of iterations: 21
Within cluster sum of squared errors: 430.21217975069385

Initial starting points (random):

Cluster 0: 80,43,68,123,53,7,150,46,19,147,169,327,176,81,7,14,179,184,bus
Cluster 1: 86,45,73,152,63,6,149,44,19,145,170,335,176,71,6,1,189,196,bus
Cluster 2: 84,37,70,145,62,9,136,48,18,134,159,280,140,68,11,9,194,202,van
Cluster 3: 84,44,80,140,58,11,156,44,20,157,166,349,176,74,5,17,183,193,van
Cluster 4: 106,55,96,196,60,12,221,30,25,173,225,717,214,72,9,13,186,196,opel
Cluster 5: 80,45,71,128,56,7,151,45,19,147,171,337,176,79,3,16,181,187,bus

Missing values globally replaced with mean/mode

Final cluster centroids:
                                    Cluster#
Attribute                Full Data          0          1          2          3          4          5
                           (846.0)    (181.0)     (98.0)    (139.0)    (110.0)    (183.0)    (135.0)
=============================================================================================================
COMPACTNESS                93.6785    99.6022    95.2041    88.6835    91.2273     99.694    83.6148
CIRCULARITY                44.8617    47.2099    43.3469    37.8417    45.0182     50.4754   42.3037
DISTANCE CIRCULARITY       82.0887     94.232    77.8571    64.3525    81.2091     97.1803     67.4
RADIUS RATIO              168.9409   190.8508   185.9286   133.8633   159.1909   193.7486   137.6667
PR.AXIS ASPECT RATIO       61.6939    61.9834    67.4286    57.0432       64.7    61.235     60.1037
MAX.LENGTH ASPECT RATIO     8.5674     9.4254     6.2653     6.7122    11.8909     9.071      7.6074
SCATTER RATIO             168.8392   188.9669   169.3265   129.7266   152.2545   204.3224   147.1852
ELONGATEDNESS              40.9338    35.8011    38.7959    51.9137    44.3273    33.0874    45.9333
PR.AXIS RECTANGULARITY     20.5827    22.1271    20.4286    17.7482    19.3273    23.3224    18.8519
MAX.LENGTH RECTANGULARITY 147.9988   152.8122   142.5918   132.705    152.8909   160.1038   140.8222
SCALED VARIANCE_MAJOR     188.6253   206.0994   191.3469   151.7914   174.1636   220.8142   169.2963
SCALED VARIANCE_MINOR     439.9113   539.9503   444.9082   250.0719   341.7182   630.7213   318.9778
SCALED RADIUS OF GYRATION 174.7033   186.5028   166.9694   138.0216   172.2545   201.1803   168.3704
SKEWNESS ABOUT_MAJOR       72.4622    68.5304    70.898     71.5252    73.2273    71.5301    80.4741
SKEWNESS ABOUT_MINOR        6.3771     7.7514     5.1224     6.6619     6.4818     6.5464     4.837
KURTOSIS ABOUT_MAJOR       12.5993    15.6575    11.551     12.0216     9.9273    14.8634     8.963
KURTOSIS ABOUT_MINOR      188.9326   190.7459   195.7245   189.5396      188.2   188.9235   181.5556
HOLLOWS RATIO             195.6324   199.8619   199.3878   194.4317   197.6091   196.1585   186.1481
Class                          bus       saab        bus        van        van       opel        bus



Time taken to build model (full training data) : 0.2 seconds

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

Clustered Instances

0      181 ( 21%)
1       98 ( 12%)
2      139 ( 16%)
3      110 ( 13%)
4      183 ( 22%)
5      135 ( 16%)
```

Para que el algoritmo fuera entrenado por la data, realizo 21 interacciones y agrupo las instancias alrededor de los centroides de los clústeres (0,1,2,3,4,5) obteniendo asi :

- El mayor porcentaje en el cluster4 con 183 instancias.
- Con un 21% el cluster0 con una cantidad de 181 instancias.
- Luego con un 16% en los clusters 2 y 5 a diferencia de una instancia.
- Con un 13% el cluster3 con 110 instancias
- Y por ultimo,con un 12% el cluster1 con 98 instancias.

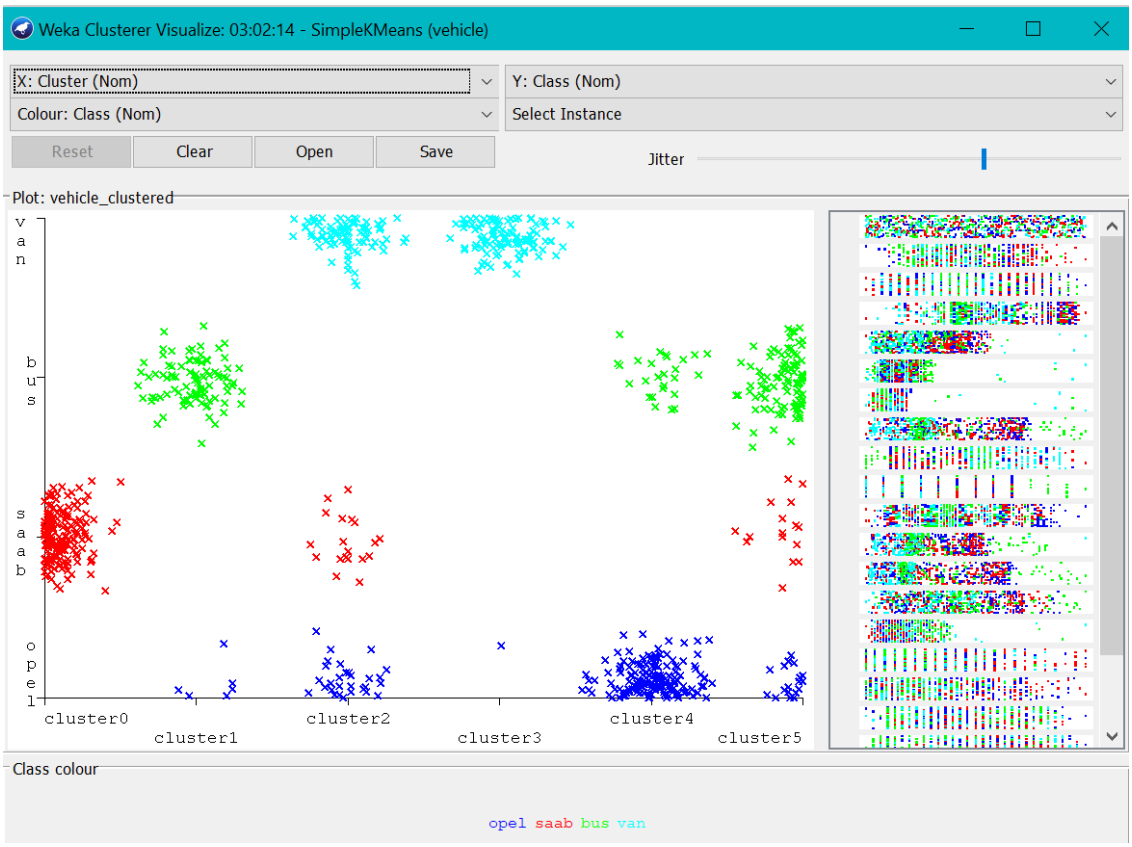


Ilustración 28: asignamiento de clusteres en visualizacion por clases

Para entender mejor el agrupamiento realizado por el algoritmo, podemos analizar por la relación de clases(Y) y clusters(X) que:

a. **El cluster0:** es donde solamente encontramos la mayoría de vehiculos categorizados *saab.*

b. **El cluster1:** es donde podemos ver vehículos categorizados *bus* y muy pocas instancias de vehiculos *opel.*

c. **El cluster2:** podemos ver una variedad entre vehículos *van* la cual puede ser el grupo con más instancia, luego vehiculos *saab* y *opel* en cantidades media.

d. **El cluster3:** Este tiene también en gran cantidad vehículos categorizados *van* y solamente podemos ver una instancia del vehículo *opel*.

e. **El cluster4:** podemos observar una gran cantidad de vehículos categorizados *opel*, además algunas instancias de vehículos *bus*.

f. **El cluster5:** este tiene una variedad, donde la mayoría son vehículo de la clase *bus*, seguido con una cantidad media los vehículos categorizado *saab* y por menor cantidad de instancias los vehículos *opel*.

También podemos visualizar el comportamiento de la data en sus otros atributos con relación a los clústeres como fue distribuida y clasificada por la clase.

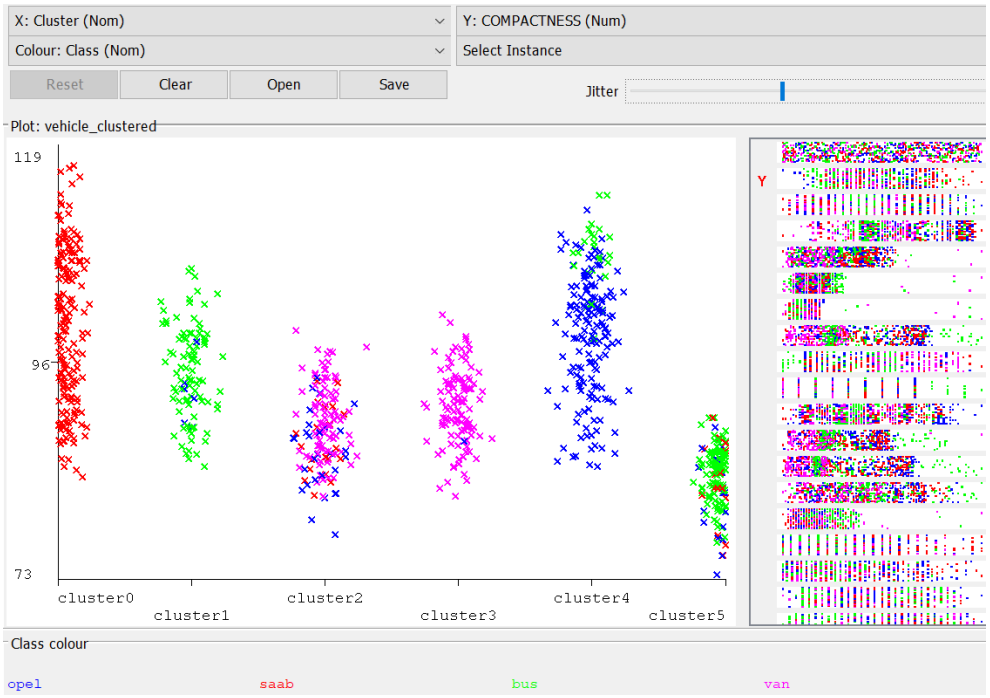


Ilustración 30: agrupación de clusteres por Compacidad.

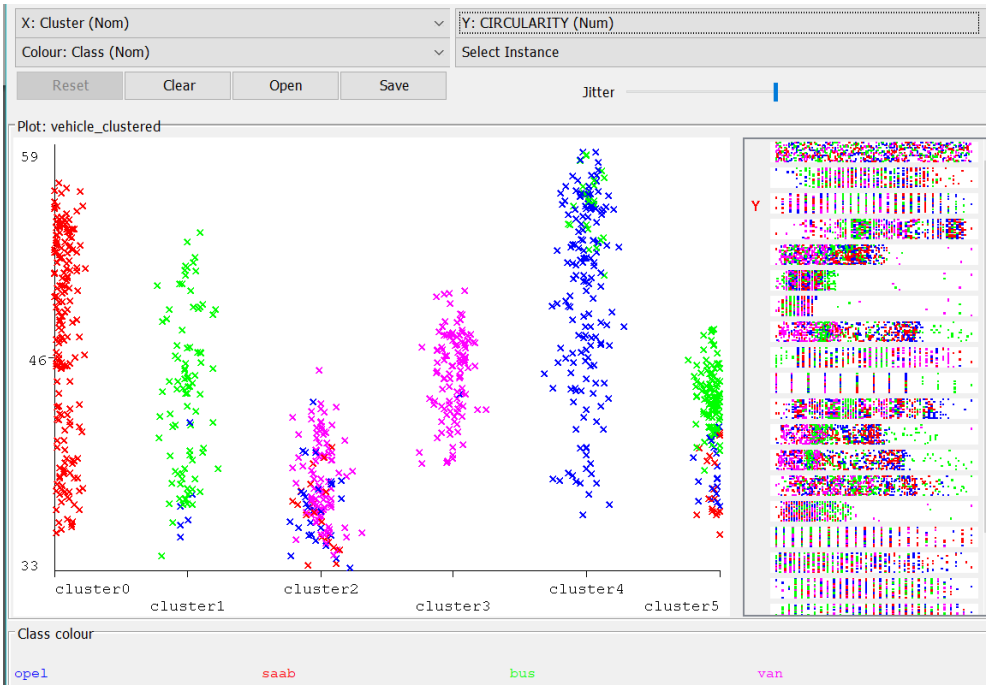


Ilustración 29: agrupación de clústeres por Circularidad

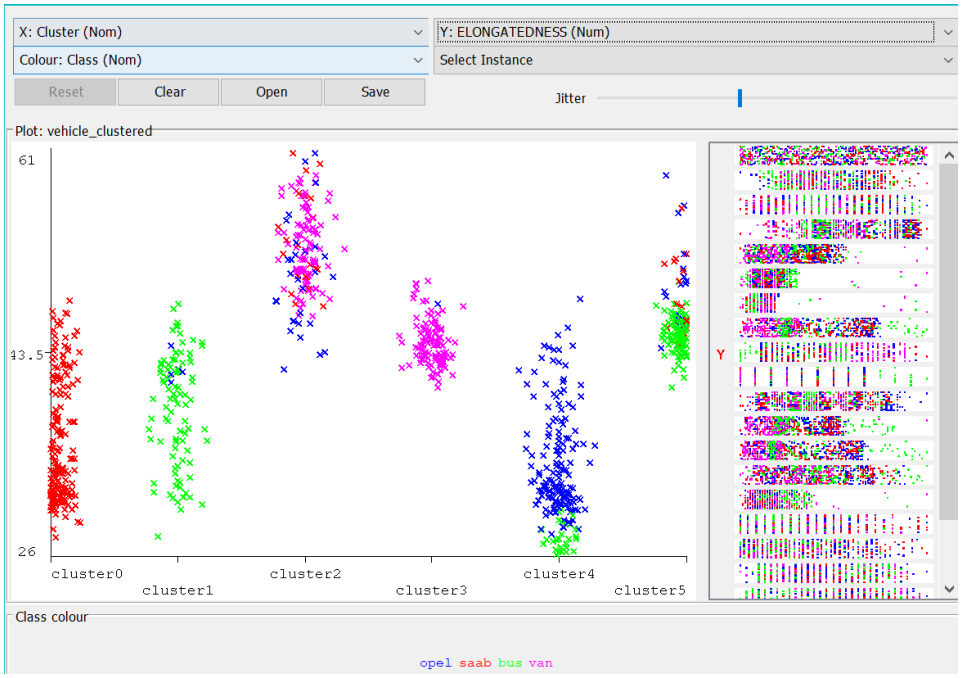


Ilustración 31: agrupación de Clusteres por Alargamiento.

Conjunto de datos glass.arff

Descripción del dataset

Este es un set de datos de sobre identificación de vidrio destinado a la clasificación. El estudio de estos datos fue motivado por investigación criminológica. Esto sustentado de manera tal que el vidrio que quede en la escena del crimen puede ser utilizado como prueba.

En general glass.arff contiene atributos relativos a varios tipos de vidrios

Información del dataset:

a. Cantidad de instancias: 214.
b. Cantidad de atributos: 10. RI (índice de refracción), Na (sodio), Mg (magnesio), Al (aluminio), Si (silicona), K (potasio), Ca (calcio), Ba (baria), Fe (hierro) y Type (referido al tipo de vidrio)

Preparación de datos

Al realizar una evaluación completa del dataset y verificar que este no contiene datos vacíos o data errónea en general podemos continuar. A su vez también comprobamos que no es necesario la aplicación de algún tipo de filtro, por lo que podemos pasar directamente al modelado y análisis de los datos.

Viewer

Relation: Glass

No.	1: RI Numeric	2: Na Numeric	3: Mg Numeric	4: Al Numeric	5: Si Numeric	6: K Numeric	7: Ca Numeric	8: Ba Numeric	9: Fe Numeric	10: **Type** Nominal
1	1.51793	12.79	3.5	1.12	73.03	0.64	8.77	0.0	0.0	build wi...
2	1.51643	12.16	3.52	1.35	72.89	0.57	8.53	0.0	0.0	vehic wi...
3	1.51793	13.21	3.48	1.41	72.64	0.59	8.43	0.0	0.0	build wi...
4	1.51299	14.4	1.74	1.54	74.55	0.0	7.59	0.0	0.0	tableware
5	1.53393	12.3	0.0	1.0	70.16	0.12	16.19	0.0	0.24	build wi...
6	1.51655	12.75	2.85	1.44	73.27	0.57	8.79	0.11	0.22	build wi...
7	1.51779	13.64	3.65	0.65	73.0	0.06	8.93	0.0	0.0	vehic wi...
8	1.51837	13.14	2.84	1.28	72.85	0.55	9.07	0.0	0.0	build wi...
9	1.51545	14.14	0.0	2.68	73.39	0.08	9.07	0.61	0.05	headlamps
10	1.51789	13.19	3.9	1.3	72.33	0.55	8.44	0.0	0.28	build wi...
11	1.51625	13.36	3.58	1.49	72.72	0.45	8.21	0.0	0.0	build wi...
12	1.51743	12.2	3.25	1.16	73.55	0.62	8.9	0.0	0.24	build wi...
13	1.52223	13.21	3.77	0.79	71.99	0.13	10.02	0.0	0.0	build wi...
14	1.52121	14.03	3.76	0.58	71.79	0.11	9.65	0.0	0.0	vehic wi...
15	1.51665	13.14	3.45	1.76	72.48	0.6	8.38	0.0	0.17	vehic wi...
16	1.51707	13.48	3.48	1.71	72.52	0.62	7.99	0.0	0.0	build wi...
17	1.51719	14.75	0.0	2.0	73.02	0.0	8.53	1.59	0.08	headlamps
18	1.51629	12.71	3.33	1.49	73.28	0.67	8.24	0.0	0.0	build wi...
19	1.51994	13.27	0.0	1.76	73.03	0.47	11.32	0.0	0.0	containers
20	1.51811	12.96	2.96	1.43	72.92	0.6	8.79	0.14	0.0	build wi...
21	1.52152	13.05	3.65	0.87	72.22	0.19	9.85	0.0	0.17	build wi...
22	1.52475	11.45	0.0	1.88	72.19	0.81	13.24	0.0	0.34	build wi...
23	1.51841	12.93	3.74	1.11	72.28	0.64	8.96	0.0	0.22	build wi...
24	1.51754	13.39	3.66	1.19	72.79	0.57	8.27	0.0	0.11	build wi...
25	1.52058	12.85	1.61	2.17	72.18	0.76	9.7	0.24	0.51	containers
26	1.51569	13.24	3.49	1.47	73.25	0.38	8.03	0.0	0.0	build wi...
27	1.5159	12.82	3.52	1.9	72.86	0.69	7.97	0.0	0.0	build wi...
28	1.51683	14.56	0.0	1.98	73.29	0.0	8.52	1.57	0.07	headlamps
29	1.51687	13.23	3.54	1.48	72.84	0.56	8.1	0.0	0.0	build wi...
30	1.5161	13.33	3.53	1.34	72.67	0.56	8.33	0.0	0.0	vehic wi...

Ilustración 32 Previa de los datos contenidos en el dataset glass.arff

Modelado y análisis de datos

Al ser este un dataset que se encarga de recopilar las característica de oxidación de los distintos tipos de vidrio, nos enfocaremos en realizar una clasificación para definir el patrón de composición de cada uno de los tipos de vidrios dependiendo de los elementos que a este lo conforman.

Para ello lo primero que haremos será aplicar un algoritmo de clasificación. En este caso utilizaremos el algoritmo más común (J48).

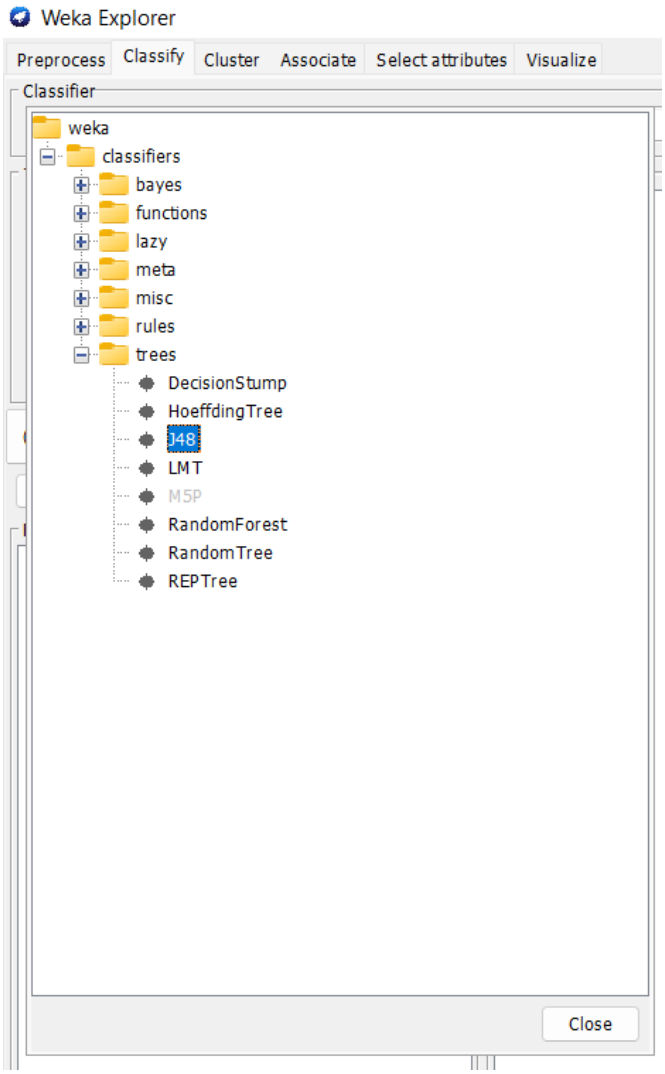


Ilustración 33. Selección del árbol J48 para realizar el modelado predictivo

Al ejecutar el árbol, este nos clasifica correctamente un 66.8% de las instancias. Con el objetivo de mejorar un poco más estos resultados realizaremos un pequeño ajuste a uno de los parámetros del árbol j48. Modificaremos el parámetro unpruned. Por defecto, el árbol lo mantiene desactivado, nosotros lo activaremos.

Activar el parámetro de poda hará que el árbol sea más fácil de entender, y también reduciremos el riesgo de sobreajustes de datos de entrenamiento. Por lo que el algoritmo será capaz de clasificar de una mejor manera los datos de entrenamiento.

Modificaremos el parámetro denominado “**minNumObj**”,con este parámetro modificado haremos que el algoritmo considere un número determinado de instancias para clasificar dentro de cada una de las hojas del árbol.

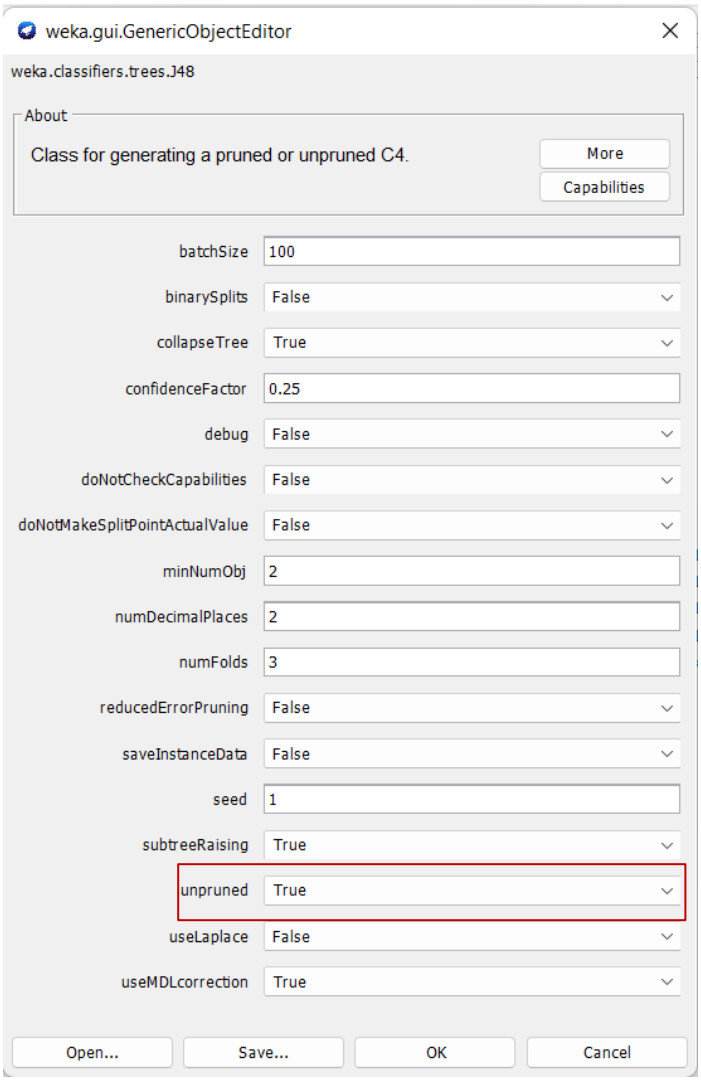


Ilustración 34. Modificación de parámetros al árbol J48

Luego de la aplicación e inicio de la evaluación surge una mejora de prácticamente un 1%, lo cual al no ser mucho, igualmente nos resulta muy útil.

El problema fundamental en este árbol radica en el análisis de esta:

```
Number of Leaves  :     30

Size of the tree :     59
```

Como se muestra en la imagen el número de hojas y el tamaño del árbol se vuelve un poco tedioso al momento de analizar, y esto se puede notar al momento de generar el árbol:

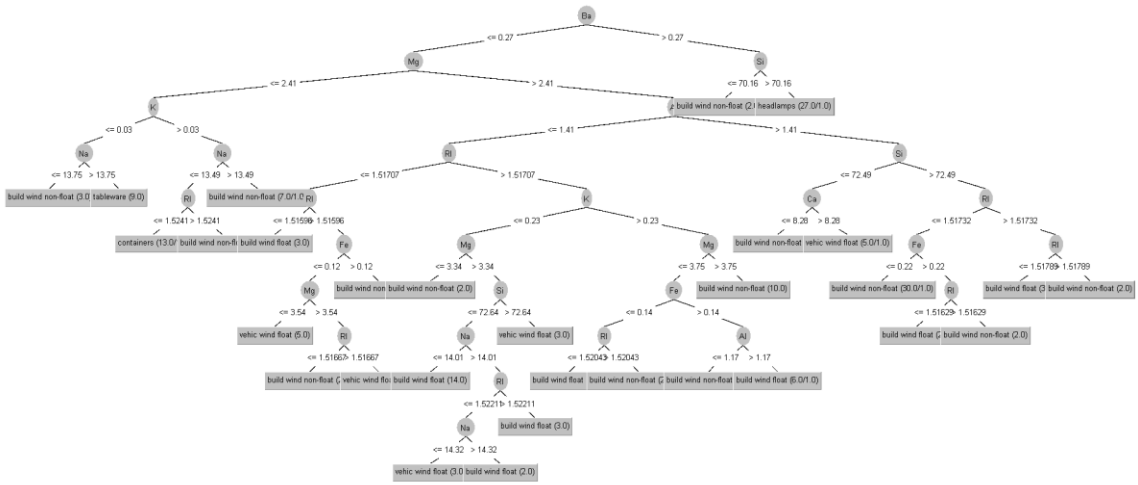


Ilustración 35. Árbol generado con el algoritmo J48

Como se puede ver en la imagen se vuelve un poco complejo realizar una evaluación a cada una da las ramas del árbol. Entonces con el objetivo de hacerlo un poco más eficiente y entendible aplicaremos otra modificación a los parámetros del árbol.

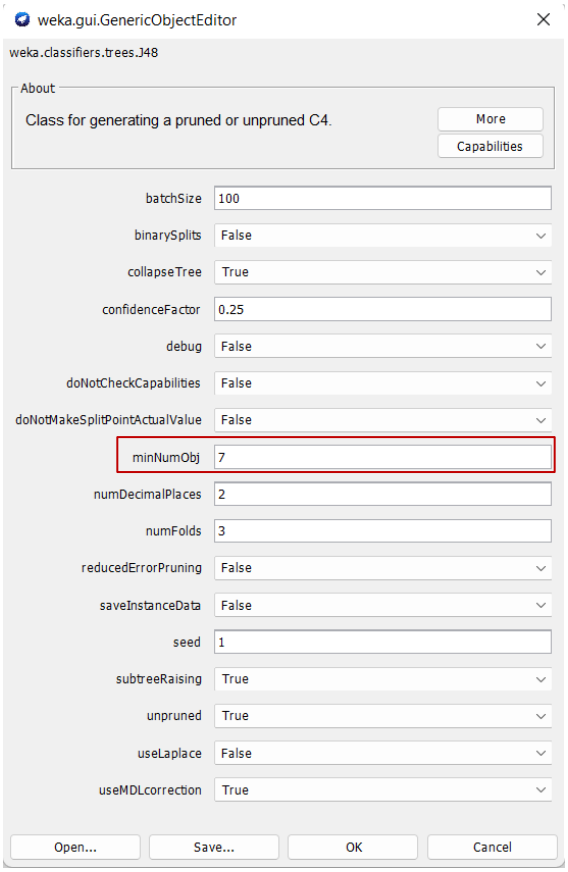


Ilustración 36 Ajustes a parámetros del árbol J48

Por defecto el árbol solo clasifica 2 instancias por hojas, lo que hace que surja la necesidad de el incremento de esta y por ende el aumento del tamaño de árbol. Ahora asignándole que este considere 7 instancias por hoja disminuiremos su tamaño considerablemente.

Al aplicar le filtro procedemos a evaluar que tanto afecto la aplicación de este parámetro al porcentaje de clasificación correcta de instancia. Para nuestra sorpresa, este no se ve prácticamente alterado con respecto a la clasificación anterior.

```
Number of Leaves  :     10

Size of the tree :      19


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               67.2897 %
Incorrectly Classified Instances        70               32.7103 %
Kappa statistic                          0.5589
Mean absolute error                      0.1152
Root mean squared error                  0.2755
Relative absolute error                 54.4035 %
Root relative squared error             84.898  %
Total Number of Instances              214

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

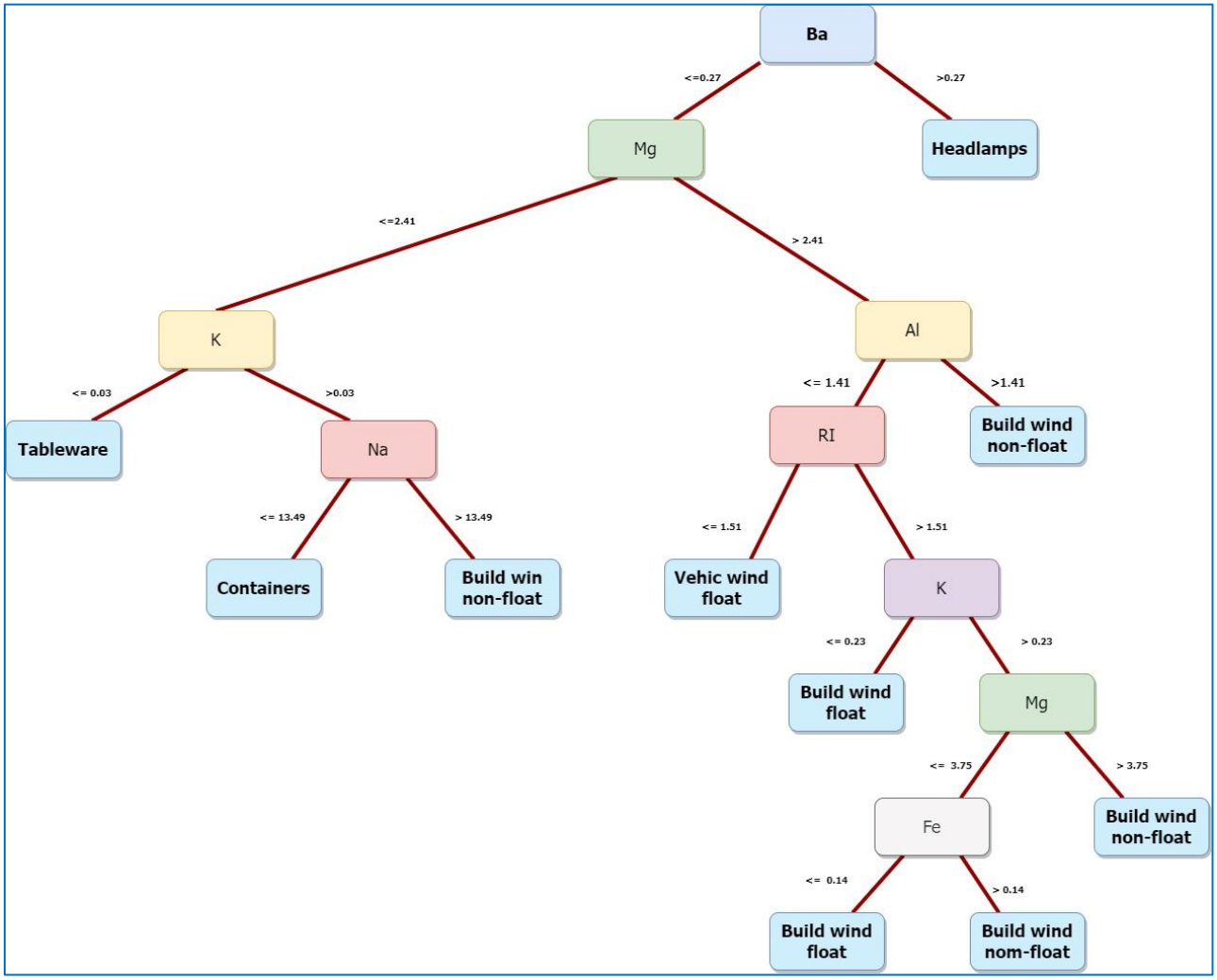
Ilustración 37 Informe de resultados del nuevo árbol J48

Como se puede notar el tamaño del árbol ha disminuido en consideración al anterior sin afectar considerablemente la clasificación correcta de la instancias.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 48 17  3  0  0  1  1 |  a = build wind float
 15 47  5  0  1  6  2 |  b = build wind non-float
  6  5  5  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  0  1  0  0  0  8  0 |  f = tableware
  1  1  0  0  0  1 26 |  g = headlamps
```

Al visualizar la matriz de confusión podemos denotar que la diagonal corresponde a la mayor parte de los datos clasificados correctamente.

Visualizacion de los resultados



De arriba hacia abajo podemos clasificar los tipos de vidrio en base a su composición de óxidos:

- Con una composición de 0.27 o menos de bario, 2.41 o menos de magnesio, y 0.03 o menos de potasio se puede deducir que se trata de un vidrio de tipo **tableware**.
- Con una composición de 0.27 o menos de bario, 2.41 o menos de magnesio, de más de 0.03 de potasio, y 13.49 o menos de sodio se puede deducir que se trata de un vidrio de tipo **containers**.
- Con una composición de 0.27 o menos de bario, 2.41 o menos de magnesio, de más de 0.03 de potasio, y más de 13.49 de sodio se puede deducir que se trata de un vidrio de tipo **build wind non-float**.
- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, 1.41 o menos de aluminio y 1.5 o menos de índice refractivo podemos señalar que estamos tratando con un vidrio de tipo **vehic wind float**.

- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, 1.41 o menos de aluminio, más de 1.5 de índice refractivo, y 0.23 o menos de potasio podemos estar hablando de un vidrio de tipo **build wind float**.
- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, 1.41 o menos de aluminio, más de 1.5 de índice refractivo, más de 0.23 de potasio, 3.75 o menos de magnesio, y 0.14 o menos de hierro hablamos de un vidrio de tipo **build wind float**.
- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, 1.41 o menos de aluminio, más de 1.5 de índice refractivo, más de 0.23 de potasio, 3.75 o menos de magnesio, y más de 0.14 de hierro podemos deducir que es un vidrio de tipo **build wind nom-float**.
- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, 1.41 o menos de aluminio, más de 1.5 de índice refractivo, más de 0.23 de potasio, y más de 3.75 de magnesio estaríamos hablando de un vidrio de tipo **build wind nom-float**.
- Con una composición de 0.27 o menos de bario, más de 2.41 de magnesio, y más de 1.41 de aluminio podemos considerar que es un vidrio de tipo **build wind nom-float.**
- Con una composición de más de 0,27 de bario consideramos que es un vidrio de tipo **headlamps**.

CONCLUSION

Gracias a todo los temas implementado sobre la manipulación de dataset e interpretarla a información, la cual fue un conocimiento adquirido durante el semestre, donde aprendimos a conocer y usar los filtros supervisados y no supervisado, y diferenciar la importancia de cada uno, como interpretar el uso indicado de los algoritmo de regresión y clasificación que son subcampos del aprendizaje automático supervisado cuyo objetivo es establecer un método para la relación entre un cierto número de características y una variable objetivo continua, permitiendo conocer la probabilidades. Además, que los algoritmos de agrupación (*clustering*) estos son procedimientos de una serie de vectores de acuerdo con un criterio que por lo general distancia o similitud. Todo esto fue eficiente y adquirido a las practica realizadas en la herramienta WEKA.

Además de saber la importancia de interpretar gráficos, que nos describen el comportamiento de la data, según la información que nos interesa extraer y poder tomar decisiones sobre estas. Todo esto fue posible con la herramienta tableau.

Con mucho esmero hemos podido concluir y demostrar que somos capaces de entender la información que se solicita, independientemente de los tipos de dataset, y como actuar si estos vienen erróneos para no estropear la información.