Universidad de Talca
Facultad de Economía y Negocios
Auditoría e Ingeniería en Control de Gestión

Proyecto Final

Integrantes:

Rigoberto Pallamar Manríquez

Marcela Rojas López

Docente:

Jimmy Gutiérrez

Modulo:

Inteligencias de Negocio

Santiago, 25 de diciembre del 2017

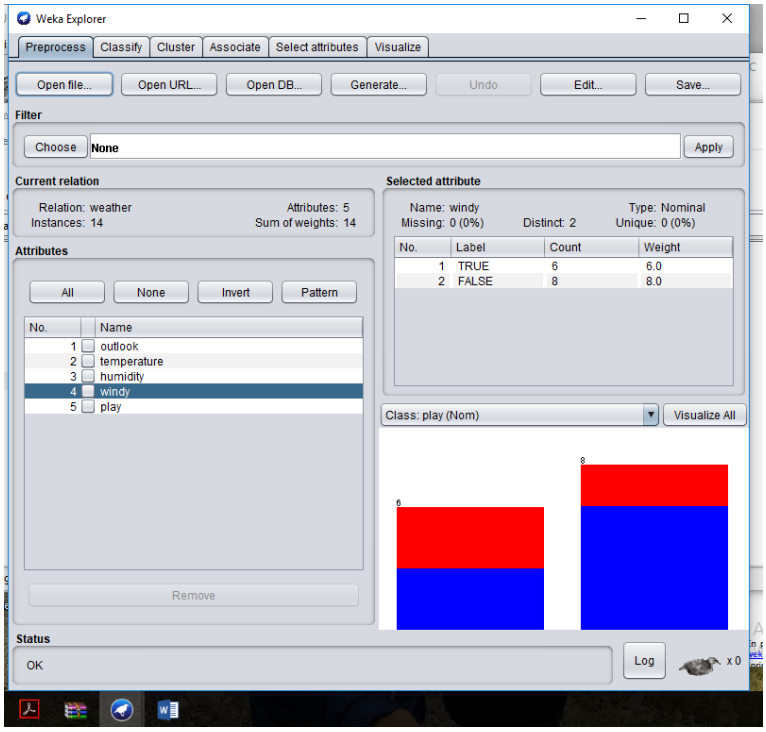
Introducción

Para comenzar definiremos algunos conceptos básicos del programa, donde Weka o "*Waikato Environment for Knowledge Analysis"* (en español entorno para análisis del conocimiento de la Universidad de Waikato) es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL.

Como se puede ver en la parte inferior de la Figura 1, Weka define 5 áreas de trabajo:

- Simple CLI: Entorno consola para invocar directamente con java a los paquetes de weka.
- Explorer: Entorno visual que ofrece una interfaz gráfica para el uso de los paquetes.
- Experimenter: Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala.
- KnowledgeFlow: Permite generar proyectos de minería de datos mediante la generación de flujos de información.
- Workbench

Si bien existen varias áreas de trabajo en el programa nosotros utilizaremos Explorer, ya que permite el acceso a la mayoría de las funcionalidades integradas en Weka de una manera sencilla.



Como se puede observar en la Figura 2 existen 6 sub-entornos de ejecución:

- Preprocess: Incluye las herramientas y filtros para cargar y manipular los datos.
- Classification: Acceso a las técnicas de clasificación y regresión.
- Cluster: Integra varios métodos de agrupamiento.
- Associate: Incluye unas pocas técnicas de reglas de asociación.
- Select Attributes: Permite aplicar diversas técnicas para la reducción del número de atributos.
- Visualize: En este apartado podemos estudiar el comportamiento de los datos mediante técnicas de visualización.

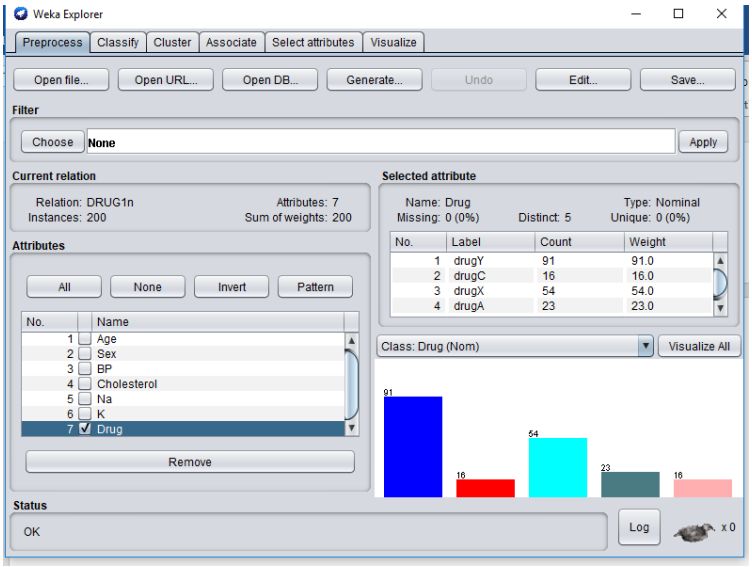
Desarrollo Base de Datos

La base de datos a utilizar es Drug1n, en este caso esta trata de predecir el tipo de fármaco que se debe administrar a un paciente afectado de rinitis alérgica según distintos parámetros/variables. Las variables que se recogen en los historiales clínicos de cada paciente son:

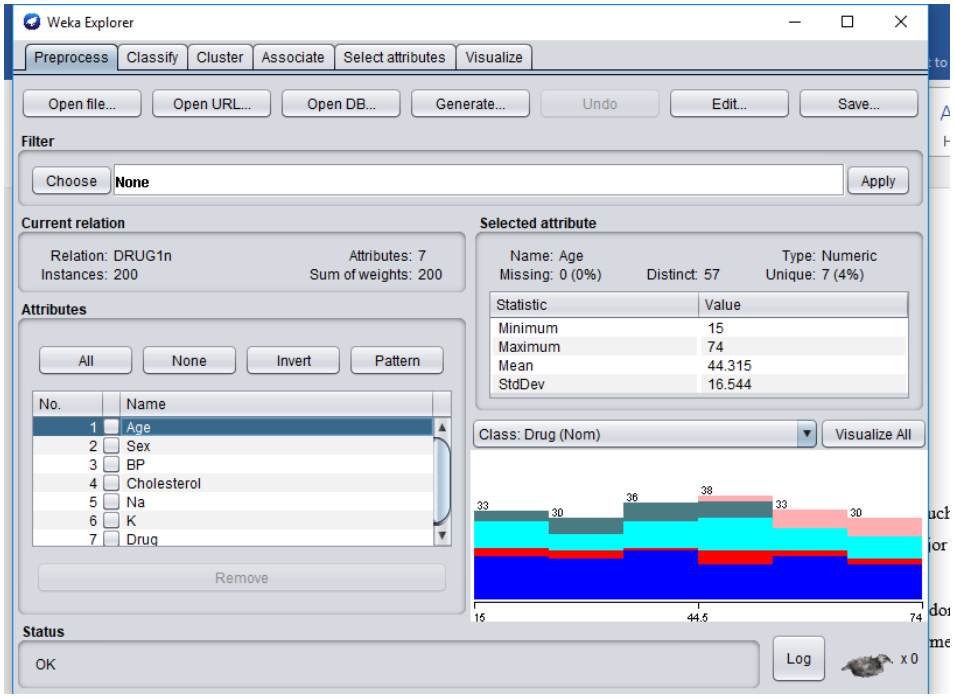
- Age: Edad
- Sex: Sexo
- BP (Blood Pressure): Tensión sanguínea.
- Cholesterol: nivel de colesterol.
- Na: Nivel de sodio en la sangre.
- K: Nivel de potasio en la sangre.

Hay cinco fármacos posibles drogas a administras entre ellas:

- DrugA
- DrugB
- DrugC
- DrugX
- DrugY

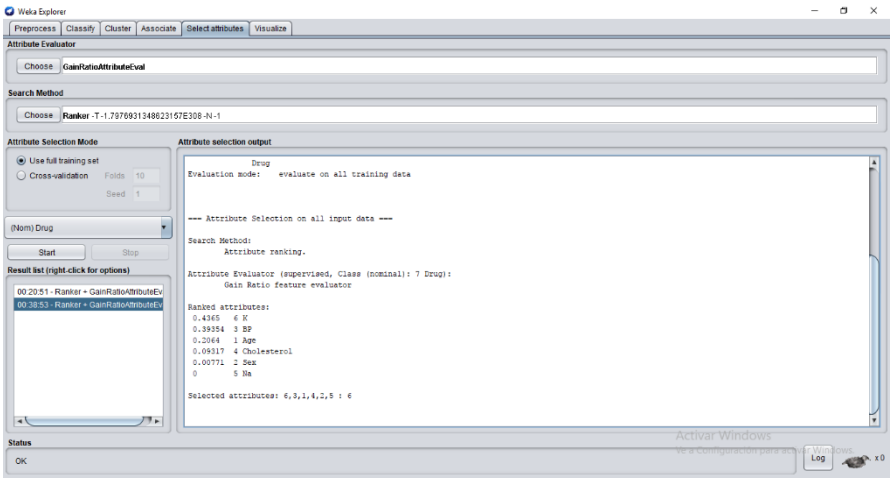


Se han recogido los datos del medicamento idóneo para muchos pacientes en cuatro hospitales. Se pretende, para nuevos pacientes, determinar el mejor medicamento a probar. Para comenzar cargaremos la base de datos en el programa Weka donde se verán los atributos que posee esta, en este caso son las variables nombradas anteriormente. La base de dato cuenta con 7 atributos y 200 instancias las que analizaremos más adelante.



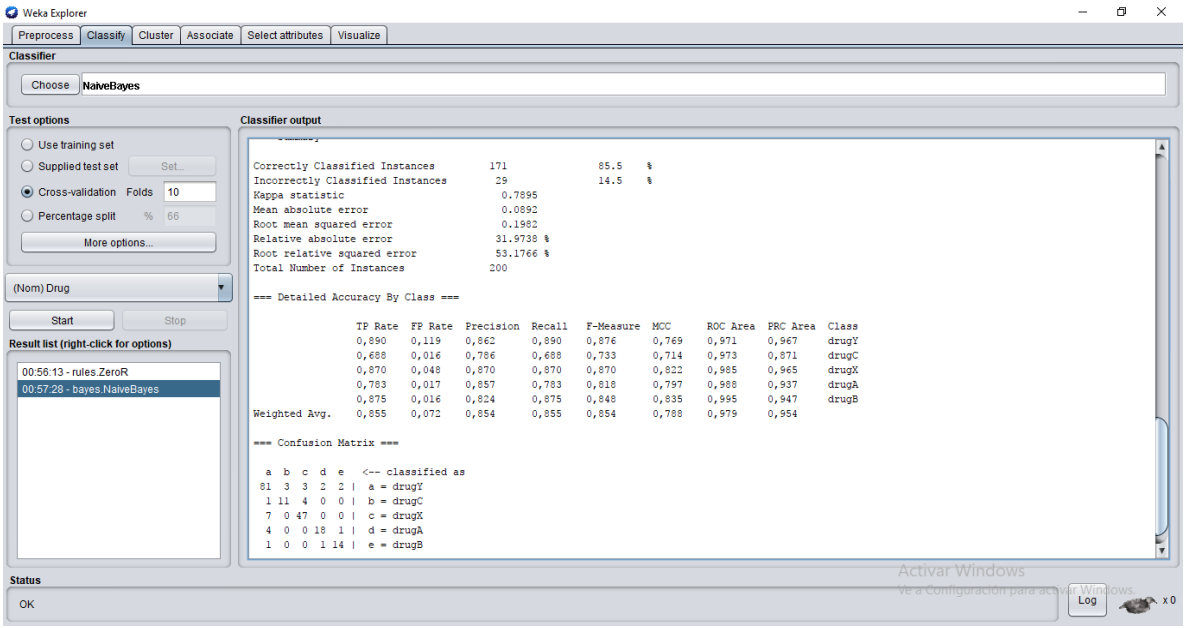
El objetivo es saber que tipo de droga recetar a nuevos pacientes según características recolectadas anteriormente.

Para esto comenzaremos con el ranking de los atributos tomando como el principal o de clase a Drug, lo que nos muestra el siguiente orden:



Según el ranking los atributos mas importantes es K (el nivel de potasio en la sangre) y el menos importante es Sex (Sexo) y Na (Nivel de sodio en la sangre). Por esto mismo queremos saber cuál es la importancia de cada atributo para encontrar la respuesta a nuestra pregunta.

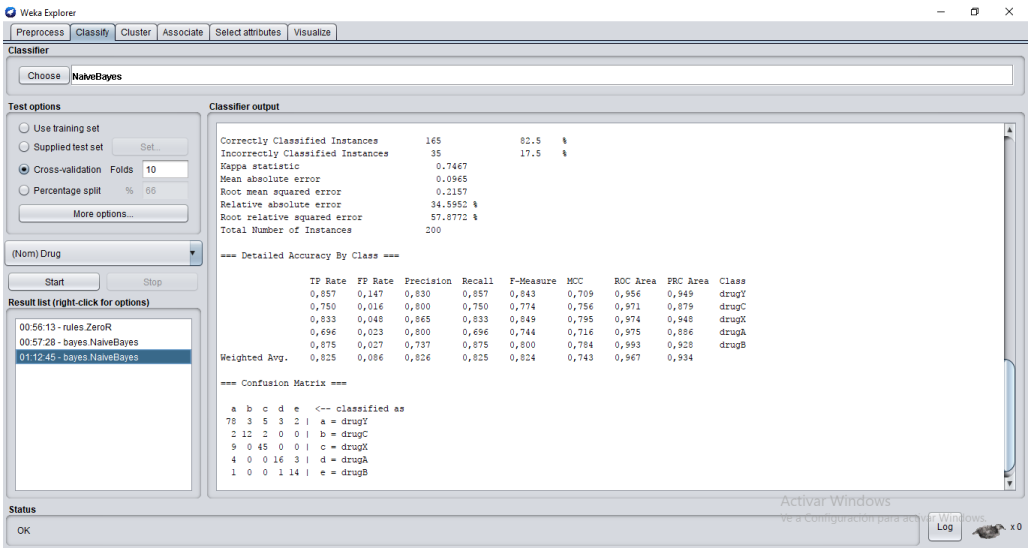
Por lo que comenzaremos con una clasificación la que incluye todos los atributos de la base de datos.



Con esta clasificación podemos ver que Weka al analizar los datos nos arroja una clasificación con 14,5% de instancias erróneas, lo que quiere decir que la información no está 100% claro, ahora veremos si al eliminar "Na" uno de los atributos que anteriormente quedo en último lugar esto mejora o empeora.

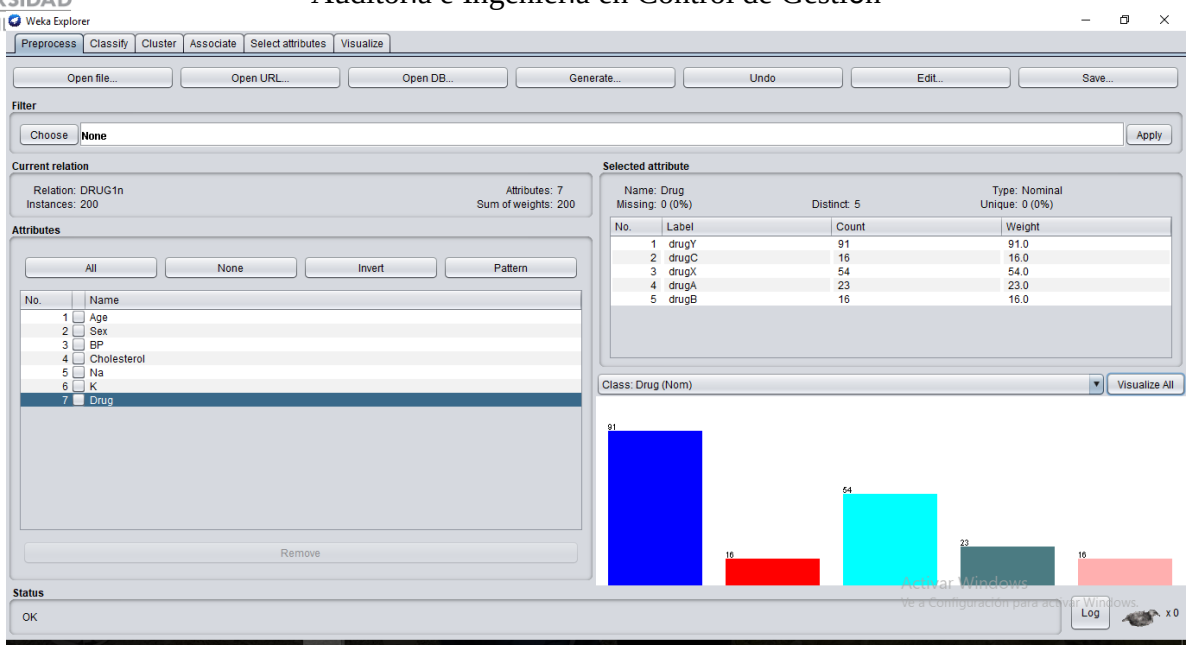
Al eliminar el atributo volvemos a realizar el ranking y nos damos cuenta de que este no varía en el orden, lo que era de esperar ya que la variable de clase sigue siendo Drug.

Realizamos una clasificación nuevamente y esta nos arroja un aumento en las instancias erróneas de 14,5% a 17,5% por esto mismo decidimos utilizar todos los atributos de la base de datos, ya que cada uno aportara para darnos una respuesta concreta.

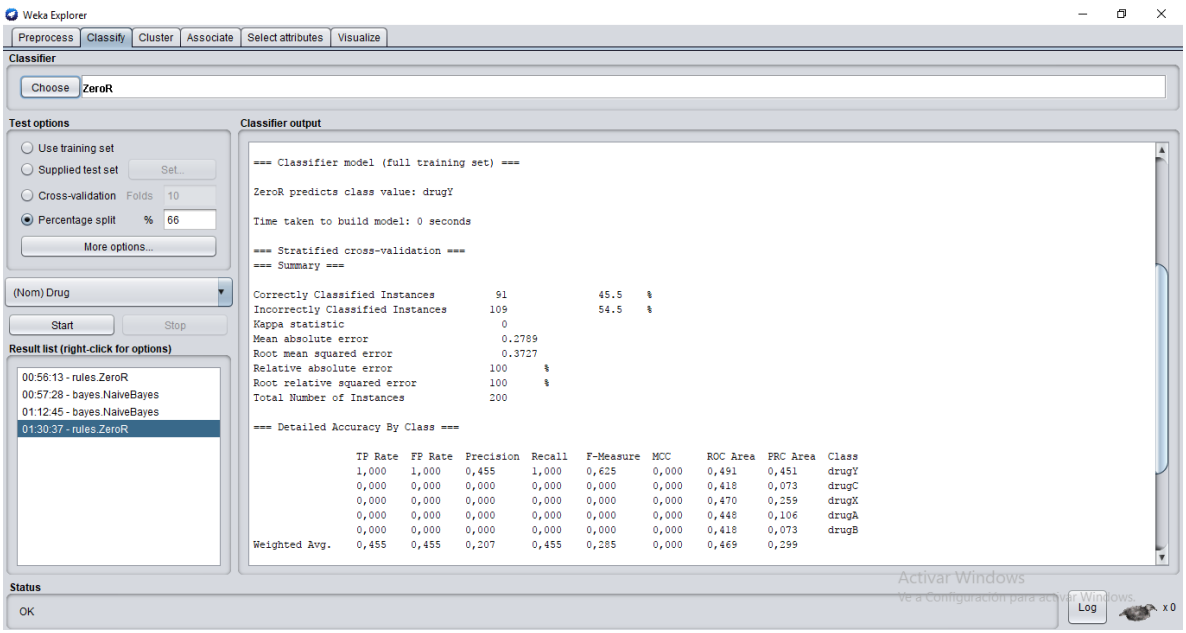


Ahora que ya sabemos que atributos considerar comenzaremos con desarrollar algunas preguntas que nos surgen al analizar la base de datos como por ejemplo es ver qué fármacos son más comunes en general, esto para ver si todos suelen ser igualmente efectivos.

Para esto seleccionamos el atributo drug, y de esta manera vemos a la distribución por clases. Podemos concluir que el fármaco más efectivo es el Y, que se administra con éxito en casi la mitad de los pacientes.



Weka tiene un método que permite generar este modelo tan simple, es decir asignar a todos los ejemplos la clase mayoritaria, recibe el nombre de *ZeroR en* la familia *rules*. Si vamos a la parte de Classify y ejecutamos este modelo evaluamos sobre todos los datos de entrenamiento, veremos que como era de esperar obtenemos un clasificador de precisión 45.5%.



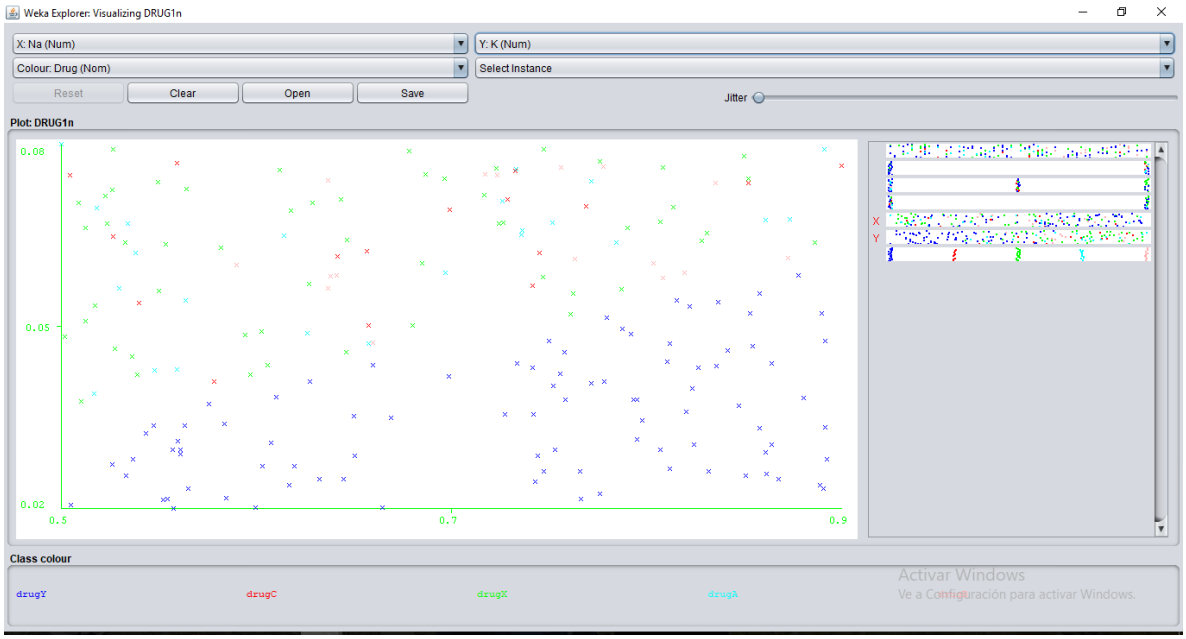
Este modelo nos da un error de más del 50%, es decir esta nos dice que el 54,5% de las veces el medicamento DRUGY no fue el adecuado según la muestra.

Es posible que otros modelos dieran mejor resultado, pero el asunto aquí es que posiblemente no hemos examinado suficientemente los datos de entrada.

Vamos a analizar, con más detenimiento, los atributos de entrada del problema. Es posible que se puedan establecer mejores modelos si combinamos algunos atributos. Podemos analizar pares de atributos utilizando diferentes gráficos.

Para comparar la relación entre atributos en Weka debemos acudir al entorno *Visualize*, donde podemos realizar gráficas entre pares de atributos y ver si tienen alguna relación con las clases.

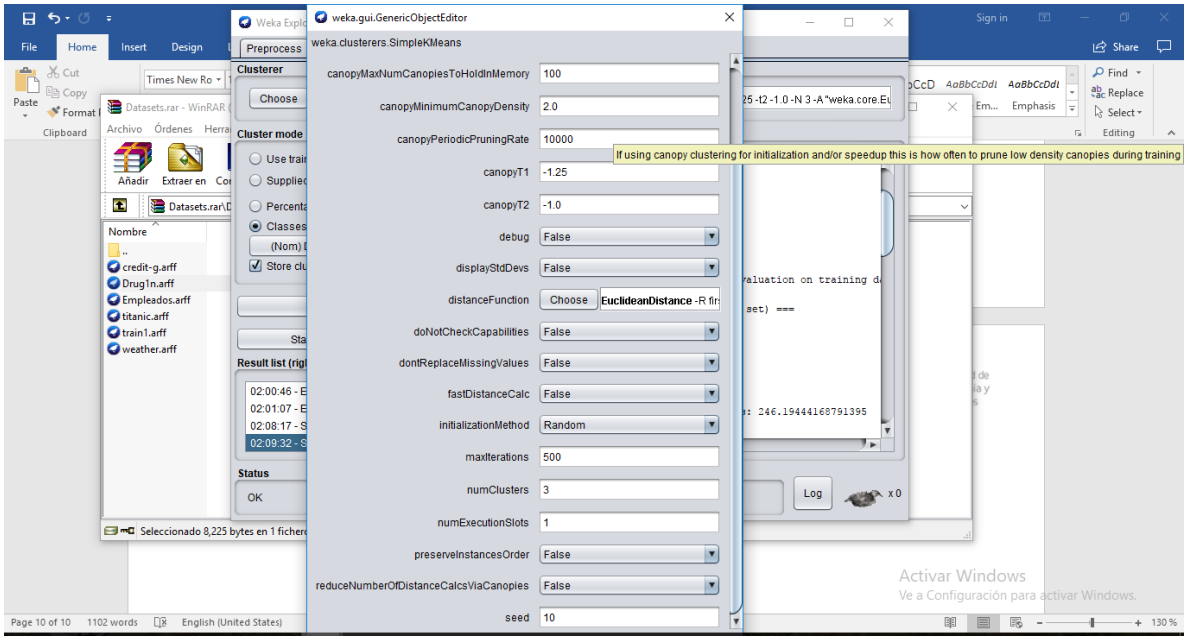
De entre todas las combinaciones posibles, destaca la que utiliza los parámetros de los niveles de sodio y potasio (K y Na), que eran según el ranking el atributo mas importante y el menos importante respectivamente.



En este gráfico sí que se ven algunas características muy significativas. Parece haber una clara separación lineal entre una relación K/Na alta y una relación K/Na baja. De hecho, para las concentraciones K/Na bajas, el fármaco Y es el más efectivo de una manera clara y parece mostrarse que por encima de un cierto cociente K/Na ese medicamento deja de ser efectivo y se debe recurrir a los otros cuatro.

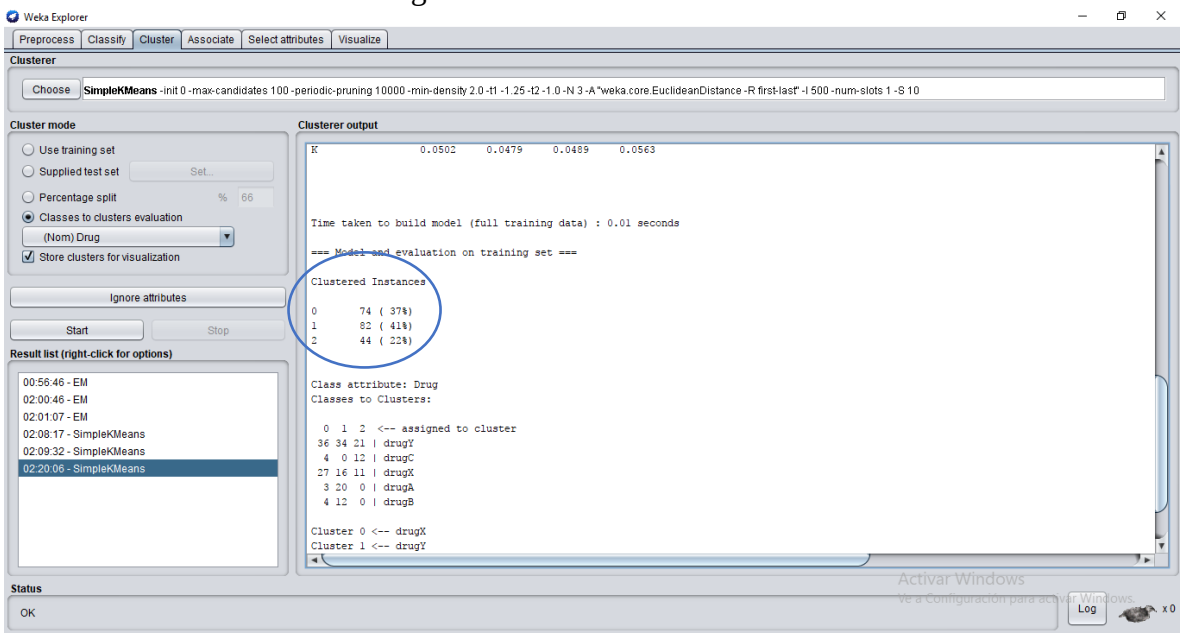
Podemos utilizar este conocimiento que acabamos de extraer para mejorar nuestros modelos. Hemos establecido que el medicamento a administrar depende en gran medida del cociente entre K/Na.

Por último, realizaremos un cluster donde seleccionaremos la opción SimpleKMeans, configuramos el número de cluster al cual le otorgamos un valor de 3 (supuesto solo para realizar prueba).

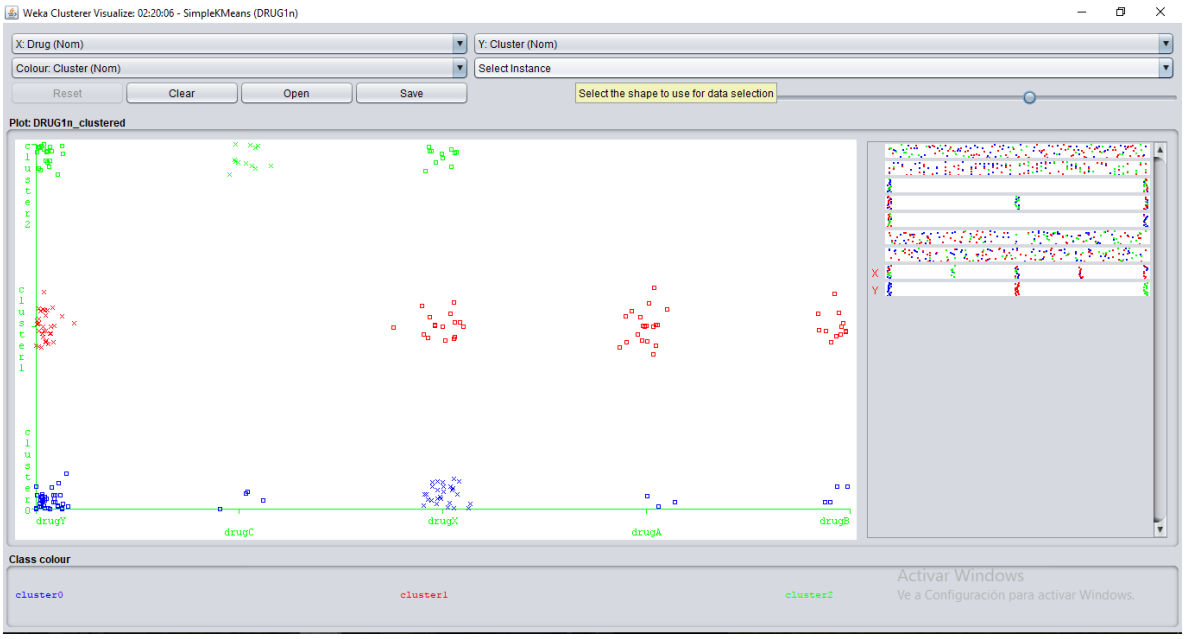


Luego seleccionaremos los atributos que no queremos utilizar en el proceso, que en este caso será Drug ya que no aporta ni una información útil para la separación del cluster.

Al realizar la prueba podemos ver que hemos conseguido 3 cluster con 74, 82 y 44 personas que usaron la droga.



A continuación, se visualiza la distribución obtenida de una manera más gráfica, donde comparamos el cluster versus el atributo Drug.



Para concluir según Cluster podemos deducir que para cada tipo de droga existen atributos a acompañar en este caso el de tres drogras mas utilizadas X, Y y C, lo que nos entrega un resumen de los atributos que se relacionaron a cada una de ellas. Por ejemplo, para la Droga X la edad es 39 años, Sexo Femenino, BP baja, Colesterol Normal, Na 0,69 y K de 0,47 y asi respectivamente con cada tipo de drogra.

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(200.0)	(74.0)	(82.0)	(44.0)
Age	44.315	39.2027	42.9512	55.4545
Sex	M	F	M	M
BP	HIGH	LOW	HIGH	LOW
Cholesterol	HIGH	NORMAL	HIGH	HIGH
Na	0.6971	0.6947	0.6705	0.7506
K	0.0502	0.0479	0.0489	0.0563

Conclusión

Con weka podemos concluir que con la prueba realizada no pudimos llegar al resultado esperado, el que dejara a una de las drogas como vencedora, pero si pudimos conocer un poco mas de los atributos de personas que utilizaron esta.

Por esto mismo podemos y según las pruebas realizadas podemos concluir que si bien existe datos concretos de que tipo de droga va utilizar cada persona según atributos estos no son 100% certeros, ya que si bien el Na era uno de los atributos menos importante según Weka al agruparse con K muestran concretamente la eficiencia de las drogas de las personas. En otras palabras, no nos podemos dejar llevar por el atributo que mas importa si no en como reaccionan al relacionarse con otros atributos.