ANALITICA DE DATOS CON APLICACIÓN EN UN CASO PRÁCTICO, MEDIANTE EL USO DE UNA HERRAMIENTA LIBRE.

CARLOS MARIO LUGO CABRERA
JHOHANN LÓPEZ HERRERA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PEREIRA
2018

ANALITICA DE DATOS CON APLICACIÓN EN UN CASO PRÁCTICO, MEDIANTE EL USO DE UNA HERRAMIENTA LIBRE

CARLOS MARIO LUGO CABRERA
JHOHANN LÓPEZ HERRERA

DIRECTOR
MSC. JUAN DE JESÚS VELOZA MORA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN

PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PEREIRA
2018

Nota de Aceptación

____

____

____

____

____

____

____

Firma del Jurado

Pereira, Abril de 2018

DEDICATORIA

A nuestras familias, que gracias al esfuerzo que hicieron por nosotros y sus ánimos para continuar adelante, fue posible culminar esta etapa en nuestras vidas.

AGRADECIMIENTOS

Para todas aquellas personas que de una u otra forma fueron parte de esta etapa en nuestras vidas y nos ayudaron en este proceso, y especialmente al Profesor Juan de Jesús Veloza Mora por haber sido nuestro director en este trabajo de grado.

CONTENIDO

1. ANEXOS

2. FIGURAS

3. INTRODUCCIÓN

En los últimos años una gran cantidad de datos se han acumulado en todas las organizaciones, y esta tendencia continúa a un ritmo acelerado. Esto ocurre por el alto uso de los sistemas computarizados .El crecimiento explosivo de las bases de datos, de Internet y el empleo de técnicas y herramientas que en forma automática y eficiente, generan información a partir de los datos almacenados, permiten descubrir patrones, relaciones y formular modelos.

Saber analizar la información disponible de forma sistemática y organizada proporciona a la organización y al gestor tranquilidad y seguridad, ya que sienten que su entorno de responsabilidad está bajo control.

Para un mejor aprovechamiento de la información existen técnicas y herramientas que determinan un camino al aprovechamiento de los datos de forma sistemática y eficiente, para ellos debemos hablar de minería de datos o big data que son conceptos que caben dentro de la analítica de datos.

Con respecto a lo anterior la analítica de datos trata de convertir los datos en información útil y relevante, al tiempo que se descubre la metodología más eficiente para realizar un análisis de calidad. Conocer y dominar una buena metodología de análisis de datos es el paso previo para poder realizar de forma rápida y eficaz estudios personalizados que se adaptan con precisión a las necesidades de cada momento y que facilitan la interpretación de la información para llegar a conclusiones de relevancia.

Las empresas en la actualidad deberían ser capaces de procesar datos en tiempo real, aquellas que ya lo están haciendo encuentran en la analítica de datos que su inversión ha empezado a rentabilizarse cuando por medio de este tipo de análisis empiezan a verse favorecida el área de TI comienza a liberar carga de trabajo, el servicio al cliente mejora considerablemente, entre otras. Por ende, aquellas empresas u organizaciones que quieran aplicar la analítica de datos deben entrar a hacer un análisis sobre la herramienta que quiere utilizar puesto que existen varias en el mercado y algunas son libres, estas herramientas pueden ser de tipo analítico con respecto a lo estadístico como lo es R y R estudio, con respecto al entorno visual como lo es RapidMiner u otra herramienta que combina tanto lo estadístico como lo visual que es WEKA.

4. DEFINICIÓN DEL PROBLEMA

El presente trabajo pretende responder y aportar información al avance en la tecnología con respecto a la analítica de datos, el almacenamiento de datos y su procesamiento. La disponibilidad de los datos es un importante activo para cualquier organización, en la medida en que puedan ser transformados en información de interés, utilizando técnicas y métodos de Data Mining.

Teniendo en cuenta que en la actualidad se están generando una cantidad considerable de datos, la mayor dificultad es la recopilación de los mismos dependiendo del tipo de datos que se quieran recopilar puede conllevar mucho trabajo u obligación de tecnología de elevado coste. Cada organización debe elegir de manera correcta los datos según la necesidad, la elección errónea podría ocasionar el siguiente y más concurrido problema que se encontrará en el pre procesamiento de datos, puesto que puede llevar demasiado tiempo, y este no asegura la obtención de un modelo válido de analítica.

5. JUSTIFICACIÓN

El motivo que nos impulsó a realizar este trabajo es la importancia que se le ha dado a la utilización de los datos en la actualidad, ya que son una pieza clave para el entorno empresarial, educativo, gubernamental, entre otros.

Ya que los datos van creciendo de forma exponencial, y saber usarlos es la prioridad en la actualidad (2018), además las herramientas informáticas son la mejor opción para trabajar con estos, por eso el principal motivo de este documento es brindarle al lector una visión de la analítica de datos en la actualidad (2018) y exponerle cuatro herramientas informáticas libres de analítica de datos con sus características.

6. OBJETIVOS

6.1 OBJETIVO GENERAL

Mostrar el estado del arte de la analítica de datos con el desarrollo de un caso práctico de aplicación teórica.

6.2 OBJETIVO ESPECÍFICOS

- Observar el marco teórico actual de la analítica de datos.
- Observar distintas técnicas para realizar analítica de datos.
- Realizar estudio sobre herramientas libres más utilizadas en la analítica de datos.
- Seleccionar una herramienta y un caso aplicado que muestre un modelo de analítica de datos.

7. DESARROLLO

7.1 MARCO TEORICO ACTUAL DE LA ANALITICA DE DATOS

La analítica de datos se ha vuelto una herramienta demasiado importante para empresas y organizaciones, para conocer su estado actual, el estado en el que se encontraban y para tomar las mejores decisiones correspondientes a sus objetivos en un futuro, si bien todo esto ya se hacía hace bastante tiempo en las empresas, pero la inclusión de las tecnologías de la información en la analítica de datos con técnicas como el big data o minería de datos ha facilitado el realizar estas tareas y se ha dado un paso más adelante para poder utilizar datos no estructurados, y realizar estos análisis de datos en tiempo real.
Algunos de los términos más importantes en la analítica de datos son:

Análisis de Datos: El Análisis de Datos (*Data Analysis, o DA*) es la ciencia que examina datos en bruto con el propósito de sacar conclusiones sobre la información. El análisis de datos es usado en varias industrias para permitir que las compañías y las organizaciones tomen mejores decisiones empresariales y también es usado en las ciencias para verificar o reprobar modelos o teorías existentes. El análisis de datos se distingue de la extracción de datos por su alcance, su propósito y su enfoque sobre el análisis. Los extractores de datos clasifican inmensos conjuntos de datos usando software sofisticado para identificar patrones no descubiertos y establecer relaciones escondidas. El análisis de datos se centra en la inferencia, el proceso de derivar una conclusión basándose solamente en lo que conoce el investigador. [9]

La ciencia generalmente se divide en **análisis exploratorio de datos** (EDA), donde se descubren nuevas características en los datos, y en **análisis confirmatorio de datos** (CDA), donde se prueba si las hipótesis existentes son verdaderas o falsas. **El análisis cuantitativo de datos** (QDA) es usado en las ciencias sociales para sacar conclusiones de datos no numéricos, como palabras, fotografías o videos. En la tecnología de la información, el término tiene un significado especial en el contexto de las auditorías informáticas, cuando se examinan los sistemas, operaciones y controles de los sistemas de la información de una organización.

El análisis de datos se usa para determinar si el sistema existente protege los datos efectivamente, opera eficientemente y cumple con las metas de la organización.[9]
El término “análisis” ha sido usado por varios proveedores de software de inteligencia de negocios como una palabra de moda que describe varias funciones. El análisis de datos se usa para describirlo todo, desde el **procesamiento analítico en línea** (OLAP, por sus siglas en inglés) hasta el **análisis CRM** en centros de llamadas. Los bancos y las compañías de tarjetas de crédito, por ejemplo, analizan los retiros y los patrones de gasto para prevenir el fraude o robo de identidad.

El análisis de datos moderno normalmente usa tableros de información que se basan en flujos de datos en tiempo real. El llamado **análisis en tiempo real** implica análisis e informes dinámicos basados en los datos que introducidos en un sistema un minuto antes del tiempo actual de uso [9].

Análisis de Big Data: El análisis de 'big data’ es el proceso de examinar grandes cantidades de datos de una variedad de tipos (big data) para descubrir patrones ocultos, correlaciones desconocidas y otra información útil. Tal información puede proporcionar ventajas competitivas a través de organizaciones rivales y resultar en beneficios para el negocio, tales como el marketing más efectivo y mayores ingresos.

El análisis de big data puede hacerse con herramientas de software de uso común en el marco de disciplinas analíticas avanzadas, como el análisis predictivo y la minería de datos.[16] Sin embargo, las fuentes de datos no estructurados utilizados para el análisis de grandes datos tal vez no encajen en los almacenes de datos tradicionales. Además, los almacenes de datos tradicionales pueden no ser capaces de manejar las demandas de procesamiento de grandes datos.

Como resultado, una nueva clase de tecnología de datos grandes ha surgido y está siendo utilizado en muchos análisis de datos grandes. Las tecnologías relacionadas con el análisis de datos incluyen bases de datos grandes NoSQL, Hadoop y MapReduce. Estas tecnologías forman el núcleo de un marco de software de código abierto que soporta el procesamiento de grandes volúmenes de datos a través de sistemas en clúster.

Data Mining (minería de datos): Es el proceso de extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de factores ocultos, tendencias y correlaciones para permitir al usuario realizar predicciones que resuelven problemas del negocio proporcionando una ventaja competitiva. Las herramientas de *Data Mining* predicen las nuevas perspectivas y pronostican la situación futura de la empresa, esto ayuda a los mismos a tomar decisiones de negocios proactivamente [17].

SAS Institute define el concepto de *Data Mining* como el proceso de Seleccionar (*Selecting*), Explorar (*Exploring*), Modificar (*Modifying*), Modelizar (*Modeling*) y Valorar (*Assessment*) grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores [17].

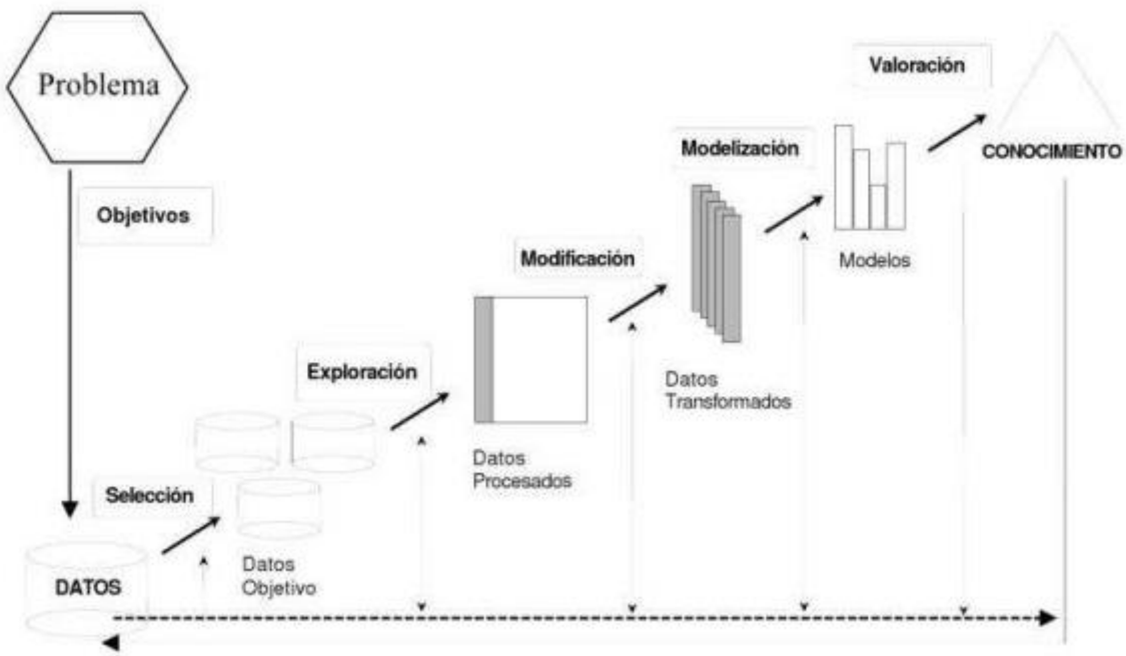

Figura 1.Data mining

7.2 TECNICAS PARA REALIZAR ANALITICA DE DATOS

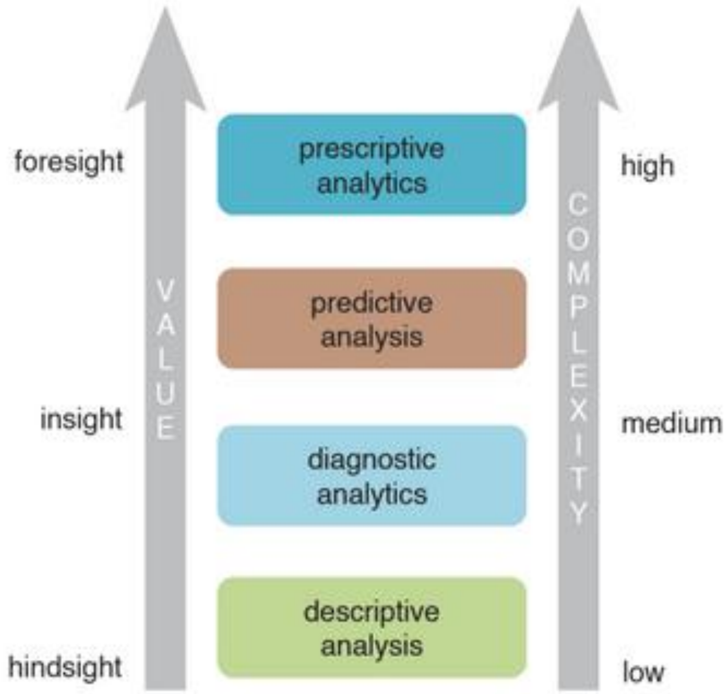

Figura 2.Orden de los tipos de analítica [6]

7.2.1 ANÁLISIS DESCRIPTIVO:

Alrededor del 80% de la analítica es de naturaleza descriptiva. En términos de valor, la analítica descriptiva proporciona un valor mínimo y requiere un conjunto relativamente básico de habilidades. [6]

La analítica descriptiva examina los datos y analiza los acontecimientos pasados para saber cómo abordar el futuro. La analítica descriptiva examina el rendimiento pasado y entiende ese rendimiento al extraer datos históricos para buscar las razones detrás del éxito o el fracaso del pasado, ya sea a través de medidas (estimadores), gráficas o tablas en donde se pueda apreciar claramente el comportamiento y las tendencias de la información recopilada. Casi todos los informes de gestión, tales como ventas, marketing, operaciones y finanzas, utilizan este tipo de análisis post-mortem.

Lo cual nos permite afirmar que la razón principal de este análisis es tratar de responder a la pregunta ¿Qué está pasando?

7.2.2 DIAGNÓSTICO DE DATOS:

Es la determinación de los datos necesarios y los métodos útiles para recolectar algún tipo de información dentro de la empresa. La recolección y el análisis de datos es una de las actividades más complejas del desarrollo organizacional. Incluye técnicas y métodos para describir el sistema organizacional y las relaciones entre sus elementos o subsistemas, así como los modos de identificar problemas y temas importantes.

La analítica diagnóstica proporciona más valor que la analítica descriptiva, y que requiere un conjunto de habilidades más avanzadas [6]. Cuando hacemos el diagnóstico del análisis de los datos recogidos se pasa a la interpretación y al diagnóstico para identificar preocupaciones y problemas y sus consecuencias, establecer prioridades, metas y objetivos. En el diagnóstico se verifican las estrategias alternativas y los planes para implementarlas.

Las características del diagnóstico de datos que se deben incluir son:

1. Funciones de la organización, como son: Metas, estrategias, estructura, políticas, procedimientos, etc.
2. Factores sociales, que incluye: Cultura, proceso de interacción, atributos y características individuales.
3. Tecnología: Herramientas, equipo y maquinaria, diseño de trabajo, conocimientos y sistemas técnicos.
4. Ambiente físico, que incluye: Configuración del espacio, diseño de interiores, diseño arquitectónico.

7.2.3 ANÁLISIS PREDICTIVO:

Es una parte de la analítica avanzada que se usa para hacer predicciones sobre sucesos futuros desconocidos. Utiliza diversas técnicas de la minería de datos para reunir toda la información tecnológica, la gestión y el proceso de construcción empresarial para elaborar predicciones de cara al futuro.

Los datos históricos y transaccionales de la empresa se pueden utilizar para identificar riesgos y oportunidades futuras [6]. Los modelos de análisis predictivo permiten evaluar los riesgos desde un determinado punto de vista. Cuando aplicamos la analítica predictiva al mundo de los negocios, las empresas pueden aprovechar las ventajas del Big Data en su propio beneficio [6].

La minería de datos y los textos analíticos, en conjunción con las estadísticas, permiten a los propietarios de un negocio construir su inteligencia predictiva, descubriendo tendencias y relaciones, tanto en el conjunto de datos estructurados como no estructurados.

Los datos estructurados que se pueden utilizar son por ejemplo la edad, el género, el estado civil, nivel de ingresos, etc. Los datos no estructurados pueden ser los contenidos en redes sociales u otros tipos de textos, incluso elementos que se pueden derivar de sus contenidos, como el sentimiento con el que pueden clasificarse [6].

La analítica predictiva permite a las organizaciones ser un poco más proactivas, tener la vista en el futuro, anticipando resultados y comportamientos, basándose así en datos y no en una serie de especulaciones. La analítica prescriptiva va un poco más allá y sugiere acciones que podemos poner en marcha, a raíz de las predicciones y sus implicaciones.

El proceso de analítica predictiva está compuesto por diferentes fases:

1. Define el proyecto

Señala los objetivos de negocio, las fuentes de datos que vas a usar, las decisiones, los resultados y el alcance que esperas obtener como resultado de tus esfuerzos.

2. Recoge los datos

La minería de datos para el análisis predictivo tiene el objetivo de recoger la información de diferentes plataformas para su análisis. Esto te permitirá tener una visión más clara de las interacciones con tus clientes.

3. Análisis de datos

Consiste en el proceso de inspeccionar, limpiar, transformar y clasificar los datos con el objetivo de descubrir información útil, que te permitirá llegar a conclusiones.

4. Estadísticas

El análisis de estadísticas te permite validar tus hipótesis y testearlas utilizando modelos estadísticos estándar.

5. Modelado

La modelación predictiva te da la oportunidad de crear, de forma automática, modelos predictivos específicos sobre el futuro. También hay opciones para elegir la mejor solución con una evolución multimodelo.

6. Puesta en marcha

La puesta en marcha de los modelos predictivos te permite desplegar los resultados analíticos de las decisiones de cada día, construyendo un proceso para obtener resultados e informes que nos permitan llegar a la automatización de decisiones.

Aplicaciones de la analítica predictiva

Veamos algunas aplicaciones prácticas que le podemos dar a las aplicaciones de analítica predictiva.

Analítica de la gestión de las relaciones con los clientes (CRM).

Las aplicaciones de análisis predictivo se utilizan para conseguir tanto los objetivos de CRM como en las campañas de marketing, ventas y atención al cliente. La gestión de las relaciones con nuestros clientes se puede aplicar a lo largo del ciclo de vida de los clientes, empezando por el proceso de adquisición, crecimiento de la relación, retención y reconquista.

7. Sanidad

Las aplicaciones de análisis predictivo también se pueden utilizar en sanidad para determinar los pacientes que están en riesgo de desarrollar algunas enfermedades como asma, diabetes y otras patologías.

8. Analítica de recopilación de datos

Las aplicaciones de análisis predictivo además se pueden emplear para la optimización de la asignación de recursos de datos, identificando bien a las agencias de recolección, las estrategias de contacto y las acciones legales para incrementar la recuperación de la información y reducir los costes de la recogida de datos.

9. Cross-sell o venta cruzada

Por otro lado, las aplicaciones de análisis predictivo también ayudan a analizar el gasto, los usos y comportamientos de los clientes. El objetivo es conseguir mejorar la venta de productos adicionales.

10. Detección de fraude

También sirven para detectar transacciones fraudulentas (tanto online como offline), robos de identidad y reclamaciones de seguros falsas.

11. Gestión del riesgo

Las aplicaciones de analítica predictiva también pueden usarse para predecir la mejor cartera para maximizar el retorno en el modelo de valoración de precios de los activos financieros (también conocido como CAPM o Capital Assets Pricing Model).

7.2.4 ANÁLISIS PRESCRIPTIVO:

El análisis prescriptivo sintetiza automáticamente grandes datos, ciencias matemáticas, reglas de negocio y machine learning para hacer predicciones y luego sugiere opciones de decisión para aprovechar las predicciones.

La analítica prescriptiva va más allá de predecir los resultados futuros al sugerir también acciones para beneficiarse de las predicciones y mostrar al tomador de decisiones las implicaciones de cada opción de decisión. La analítica prescriptiva no sólo anticipa lo que sucederá y cuándo ocurrirá, sino también por qué sucederá [6].

Además, la analítica prescriptiva puede sugerir opciones de decisión sobre cómo aprovechar una oportunidad futura o mitigar un riesgo futuro e ilustrar la implicación de cada opción de decisión. En la práctica, la analítica prescriptiva puede procesar continuamente y automáticamente nuevos datos para mejorar la precisión de la predicción y proporcionar mejores opciones de decisión.

El análisis prescriptivo combina sinérgicamente datos, reglas de negocio y modelos matemáticos. Las entradas de datos a la analítica prescriptiva pueden provenir de múltiples fuentes, internas (dentro de la organización) y externas (medios sociales, et al.).

Los datos también pueden estar estructurados, lo que incluye datos numéricos y categóricos, así como datos no estructurados, como texto, imágenes, audio y datos de video, incluyendo datos grandes. Las reglas de negocio definen el proceso empresarial e incluyen restricciones, preferencias, políticas, prácticas recomendadas y límites. Los modelos matemáticos son técnicas derivadas de ciencias matemáticas y disciplinas relacionadas incluyendo estadísticas aplicadas, aprendizaje de máquinas, investigación de operaciones y procesamiento del lenguaje natural [6].

7.3 HERRAMIENTAS LIBRES MÁS UTILIZADAS EN LA ANALITICA DE DATOS

7.3.1 HERRAMIENTA RAPIDMINER

RapidMiner es una de las herramientas de minería de datos de código abierto más ampliamente utilizada, fue desarrollada en 2001 por Ingo mierswa y Ralf Klinkenberg. Antes del 2006, era conocido como YALE (Yet another Learning Tool) [1].RapidMiner es una herramienta de minería de datos basados en XML que se utiliza para ejecutar los diferentes procesos de machine learning y minería de datos. Es una herramienta popular para implementar algoritmos de clasificación y agrupación. Una característica importante de RapidMiner es su capacidad para mostrar los resultados visualmente. "RapidMiner proporciona esquemas de aprendizaje y los modelos y algoritmos de WEKA y scripts de R que pueden ser utilizados a través de extensiones ". [2]

RapidMiner también proporciona un entorno integrado para el machine learning, análisis predictivo, minería de datos, minería de textos y análisis de negocios. Se utiliza para aplicaciones comerciales e industriales, así como para la investigación, la educación, la formación, la creación rápida de prototipos y desarrollo de aplicaciones y es compatible con todos los pasos del proceso de minería de datos. El mercado extensiones de Rapidminer proporciona una plataforma para los desarrolladores crear algoritmos de análisis de datos y publicarlos en la comunidad [2]. RapidMiner Studio proporciona 3 perspectivas para los usuarios trabajar con:

La primera perspectiva es inicio perspectiva:

Este es la perspectiva por defecto cuando se inicia RapidMiner por primera vez. RapidMiner proporciona información acerca de las nuevas versiones o versión actual que el usuario está ejecutando en esta perspectiva.

El segundo es el diseño perspectiva:

Es el principal perspectiva de RapidMiner estudio donde en todos los procesos se implementan y trabajan. La perspectiva de diseño se muestra en la figura 2 a continuación. Se muestra varios módulos de RapidMiner estudio.
La vista “operador” se utiliza para seleccionar diferentes operadores para llevar a cabo cualquier tarea de minería de datos.
La vista “Proceso” muestra la actual implementación de los algoritmos de minería de datos y muestra los resultados.
La vista “Repositorio” se compone de todos los repositorios de datos almacenados en RapidMiner estudio.
La vista “Problemas” es usada para mostrar cualquier error durante la ejecución del proceso y también ofrece sugerencias para eliminar esos errores.
La vista “Parámetros” se usa para establecer otros parámetros para un operador particular. Por ejemplo, el tamaño de las agrupaciones establecido para el K-means operador.

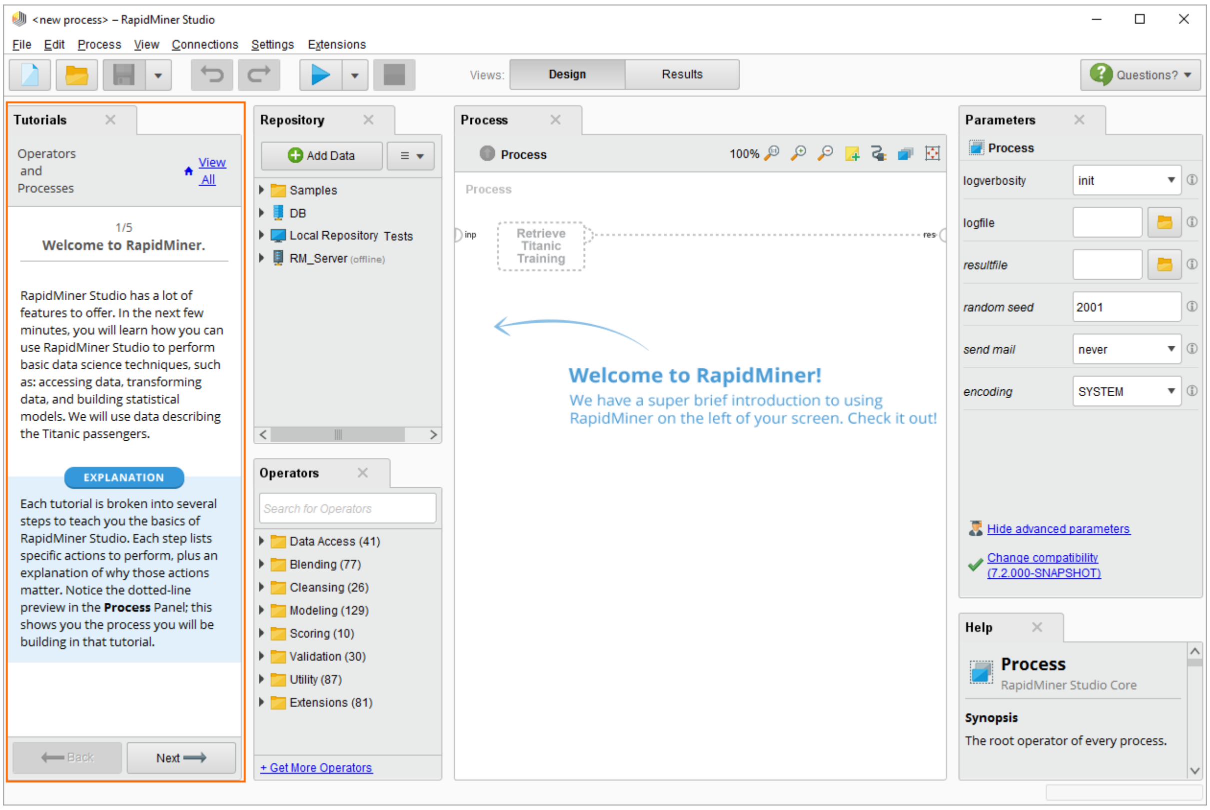

Figura 3.Visión general de diseño de perspectiva

La tercera perspectiva es perspectiva del Asistente:

Esta perspectiva se utiliza para aplicar algoritmos de minería de datos en el conjunto de datos del usuario. Es una característica muy útil proporcionada por RapidMiner estudio.

RapidMiner Studio proporciona una plataforma para las técnicas de minería de datos existentes. Una gran característica de RapidMiner es su metodología de arrastrar y soltar. Los usuarios pueden seleccionar los operadores y combinarlos mediante el uso de este enfoque [3]. Esto hace que sea más fácil de realizar las tareas deseadas para la aplicación de diversos algoritmos de minería de datos. Todo este procedimiento se completa usando combinaciones de diferentes procesos.

Otra característica importante de RapidMiner es que proporciona muy buenas visualizaciones de los datos como gráficos 3-D y matrices [3]. RapidMiner se centra mucho en proporcionar un mayor impacto visual de los datos, ya que podemos comprender imágenes visuales más rápidamente en comparación a texto sin formato. RapidMiner consta de más de 100 sistemas de aprendizaje clasificación y análisis de clustering. Otra característica importante de RapidMiner es su interfaz gráfica de usuario sensible e intuitiva.

7.3.2 HERRAMIENTA WEKA

La herramienta WEKA es una de las herramientas de código abierto de minería de datos más populares desarrollados en la Universidad de Waikato en Nueva Zelanda en 1992 [4]. Se trata de una herramienta basada en Java y se puede utilizar para implementar varios algoritmos de Machine Learning y minería de datos escritos en Java. La sencillez de uso de WEKA se ha convertido en un punto de referencia para el Machine Learning y la aplicación de minería de datos. WEKA apoya la lectura de archivos de varias bases de datos diferentes y También permite importar los datos a través de Internet, desde páginas web o desde un servidor de base de datos SQL situado a distancia, mediante la introducción de la dirección URL del recurso [4]. Entre todas las herramientas de minería de datos disponibles, WEKA es el más utilizado de todos debido a su gran rendimiento, WEKA puede ser fácilmente descargado y desplegado.

WEKA proporciona tanto, una GUI y CLI para realizar minería de datos y hace un buen trabajo de proporcionar apoyo a todas las tareas de minería de datos. WEKA soporta una variedad de formatos de datos como CSV, ARFF y binarios. WEKA se centra más en la representación textual de los datos en lugar de la visualización aunque proporciona soporte para mostrar algunos de visualización pero son muy genéricos [5].

También, WEKA no proporciona la representación visual de los resultados del procesamiento de una manera eficaz y entendible como RapidMiner, el desempeño de WEKA cuando el tamaño del conjunto de datos no es grande es muy preciso. Pero si el tamaño es grande, entonces WEKA tiende a experimentar algunos problemas de rendimiento. WEKA proporciona soporte para el filtrado de datos o atributos. En la página siguiente muestra las opciones disponibles para la interfaz de usuario.

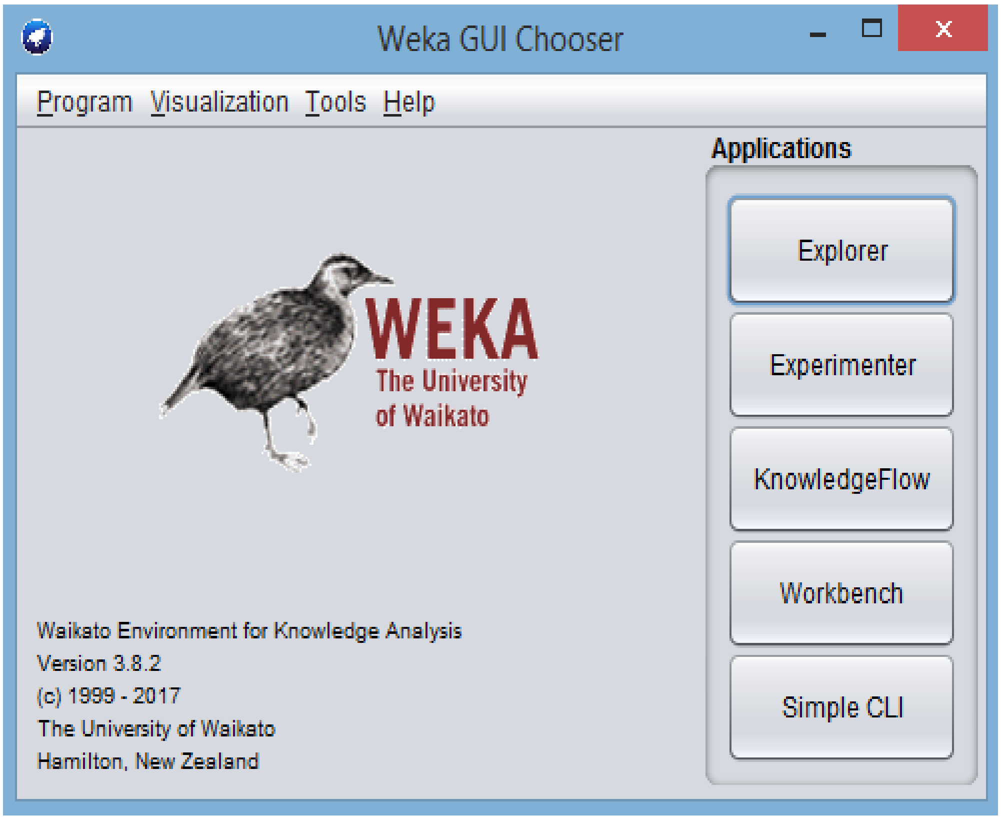

Figura 4.WEKA

1. El explorador:

Es la interfaz de usuario gráfica que se utiliza más comúnmente en Weka para implementar algoritmos de minería de datos. Es compatible con el análisis exploratorio de datos para realizar procesamiento previo, la selección de atributos, el aprendizaje y la visualización [4]. Esta interfaz se compone de diferentes pestañas para acceder a varios componentes para la minería de datos que se puede ver en la Figura 4 a continuación. Las diferentes pestañas son:

- **Pre procesamiento:** con esta ficha, podemos cargar los archivos de datos de entrada y realizar pre-procesamiento de estos datos utilizando filtros.

- **Clasificar:** Esta ficha se utiliza para implementar diferentes algoritmos de clasificación y regresión. Podemos hacer esto mediante la selección de un clasificador particular.

- **Asociar:** Esta ficha se utiliza para averiguar todas las reglas de asociación entre los diferentes atributos de los datos y que se pueden utilizar nuevas extracciones. Por ejemplo, la minería de reglas de asociación, etc.

- **Racimo:** Con esta ficha, podemos seleccionar un algoritmo de agrupamiento particular, para poner en práctica para nuestro conjunto de datos. algoritmos de agrupamiento como K-means se pueden implementar utilizando esta ficha.

- **Seleccione los atributos:** Esta ficha se utiliza para seleccionar los atributos particulares de la serie de datos útiles para implementar el algoritmo.

- **Visualizar:** Esta pestaña se utiliza para visualizar los datos siempre disponibles o apoyado por un algoritmo particular en forma de matriz gráfica de dispersión

.

2. El experimentador:

Esta interfaz de usuario proporciona un entorno experimental para probar y evaluar algoritmos de Machine Learning.

3. El flujo de conocimiento:

Flujo de conocimiento es básicamente una interfaz basada en un componente similar a Explorer. Esta interfaz se utiliza para las nuevas evaluaciones de proceso.

7.3.3 LA HERRAMIENTA TABLEU:

Tableau es una poderosa herramienta de visualización de datos utilizados en la inteligencia de negocio y análisis de datos. Tableau Software fue inventado por Chris Stolte, Christian Chabot y Pat Hanrahan en enero de 2003. La visualización proporcionada por Tableau ha mejorado por completo la posibilidad de obtener más conocimientos sobre los datos que estamos trabajando y se puede utilizar para proporcionar predicciones más precisas. "Las preguntas sobre el producto bases de datos relacionales, cubitos, bases de datos en la nube y hojas de cálculo y luego genera un número de tipos de gráficos que se pueden combinar en cuadros de mando que se pueden compartir con seguridad sobre una Red de computadoras o el Internet [7].

A diferencia de RapidMiner y WEKA, Tableau no implementa algoritmos de minería de datos proporciona visualizaciones de los datos. Para esto, Tableau proporciona integración con otro popular análisis estadístico herramienta R, para proporcionar apoyo a la minería de datos. “Tableau ofrece cinco productos principales, a saber
Tableau Desktop, Tableau Server, Tableau en línea, Tableau Reader y Tableau Public. Tableau Public y Tableau Reader son disponibles libremente, mientras que Tableau Server y Tableau de escritorio vienen con un período de prueba gratuita después que el usuario tiene que pagar” [7].

Tableau ha hecho posible explorar y presentar los datos de una manera mucho más simple y práctica. Trabajar en proyectos que usan Tableau consume menos tiempo y es fácil de manejar.

7.3.4 LA HERRAMIENTA DE PROGRAMACIÓN R:

R es una herramienta de análisis estadístico de código abierto basado en C y FORTRAN lenguaje de programación desarrollado por Ross Ihaka y Robert Caballero en la Universidad de Auckland, Nueva Zelanda. R fue lanzado en 1997 y está actualmente autorizado el uso de Licencia Pública General de GNU. Cuando hablamos de la herramienta R, hablamos de inclusión de símbolos y fórmulas matemáticas siempre que sea necesario. R utiliza una serie de diferentes paquetes para apoyar la minería de datos o estadísticas y proporciona una colección de herramientas bien integradas para análisis de datos [8].

R puede ser fácilmente compatible cuando deseamos integrar con otra herramienta, en este caso, la integración puede ser utilizada en combinación con Tableau para crear representaciones visuales de diversos algoritmos de minería de datos debido a las visualizaciones claras e interactivas que pueden ser creados usando Tableau.

Aunque R ofrece menos apoyo a los algoritmos de minería de datos en comparación a RapidMiner y WEKA, lo hace poner en práctica unos algoritmos de minería de datos. R utiliza una metodología impulsada con código, que implica el uso de un número de funciones construidas y comandos para realizar el análisis estadístico y minería de datos. Aparte de la ayuda para la extracción de datos, R tiene una gran colección de la biblioteca de estadística [8].

R-proyecto puede decirse que es análoga a la programación en MATLAB debido a su funcionalidad similar. No hay función de pre procesamiento incorporado disponible en R.

7.4 ESTUDIO COMPARATIVO

Cada una de estas herramientas de minería de datos y los datos de visualización tiene sus propias características especiales y también inconvenientes. Las diferencias son las que los hacen únicos y popular en su propio camino. Después de estudiar las características de cada herramienta, he recopilado la siguiente lista de características comparables proporcionadas por cada una de estas herramientas de minería de datos.

1. Usabilidad:

Esta característica determina la facilidad de uso de cada herramienta. Esto describe que la interfaz de usuario es comparativamente más fácil de usar.

2. Velocidad:

La velocidad es un factor importante que distingue entre las diferentes herramientas de minería de datos. Esta característica ayuda a entender cómo la configuración del sistema impacta el trabajo de una herramienta de minería de datos en particular.

3. Visualización:

La visualización es la característica más importante de una herramienta de minería de datos. Esta característica comparativa distingue a cada herramienta de minería de datos en base a diferentes opciones de visualización proporcionadas.

4. algoritmos soportados:

Esta función clasifica las herramientas de minería de datos basados en la implementación del algoritmo con el apoyo de ellos y la elección de selección descriptor disponible.

5. Tamaño de conjunto de datos:

Datos pequeños o más grandes del soporte de conjuntos es otra de las características comparables entre diferentes herramientas de minería de datos.

6. Uso de memoria:

A medida que el uso de memoria afecta al rendimiento, el uso de memoria es otra característica importante para comparar herramientas de minería de datos.

7. Uso principal:

Cada herramienta de minería de datos tiene un uso particular que es una de las características comparables. Por ejemplo, tanto R y WEKA se pueden utilizar para implementar algoritmos de minería de datos, pero el uso principal de R está en el cálculo estadístico.

8. Tipo de interfaz soportado:

El tipo de interfaz que se proporciona para la implementación algoritmo es una de las características comparativas de este estudio.
Desde el contexto de este estudio comparativo, el uso de la interfaz gráfica de usuario (GUI) o la línea de comandos de la interfaz (CLI) que diferencia a cada herramienta.

	RapidMiner	**WEKA**	**Tableau**	**R**
Usabilidad	Fácil de usar	Más fácil de usar	Simple de usar	Complicado porque se requiere codificación
Velocidad	Requiere más memoria para operar	Funciona más rápido en cualquier máquina	Funciona rápido en cualquier máquina	Funciona rápido en cualquier máquina
Visualización	Varias opciones pero menos que Tableau	Menos opciones	Muchas opciones de visualización	Menos opciones en comparación con RapidMiner
Algoritmos soportados	Clasificación y Clustering	Clasificación y Clustering	No se utiliza para implementar algoritmos	Menos opciones en comparación con RapidMiner
Tamaño conjunto de datos	Soporta grandes y pequeños tamaños	Soporta solo pequeños tamaños	Es compatible con cualquier tamaño	Soporta grandes y pequeños tamaños

Uso de memoria	Requiere más memoria	Menos memoria por lo tanto funciona más rápido	Menos memoria	Más memoria
Uso principal	Minería de datos, Análisis predictivo	Machine Learning	Inteligencia de negocio	Estadística informática
Tipo de interfaz soportado	GUI	GUI / CLI	GUI	CLI

7.5 SELECCIÓN DE LA HERRAMIENTA Y CASO DE USO PRÁCTICO

Con la anterior comparación, se decidió utilizar la herramienta WEKA en el caso práctico que se realizará a continuación ya que es la que más se adapta a nuestras necesidades.

En este caso se trata de predecir el tipo de fármaco (drug) qué 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: 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.

Lo primero que vamos a hacer es cargar los datos en el área de trabajo. Para ello, hay que dar clic en el botón "Open file" del entorno "preprocess". Seleccionamos el fichero "Drug1n.arff" y si todo ha ido bien veremos la pantalla de la Figura 4. WEKA utiliza un formato específico de datos, el formato arff. Un fichero con este formato, no sólo contiene los datos desde donde vamos a efectuar el aprendizaje, además incluye meta-información sobre los propios datos, como por ejemplo el nombre y tipo de cada atributo, así como una descripción textual del origen de los datos. Podemos convertir ficheros en texto conteniendo un registro por línea y con los atributos separados con comas (formato csv) a ficheros arff mediante el uso de un filtro convertidor.

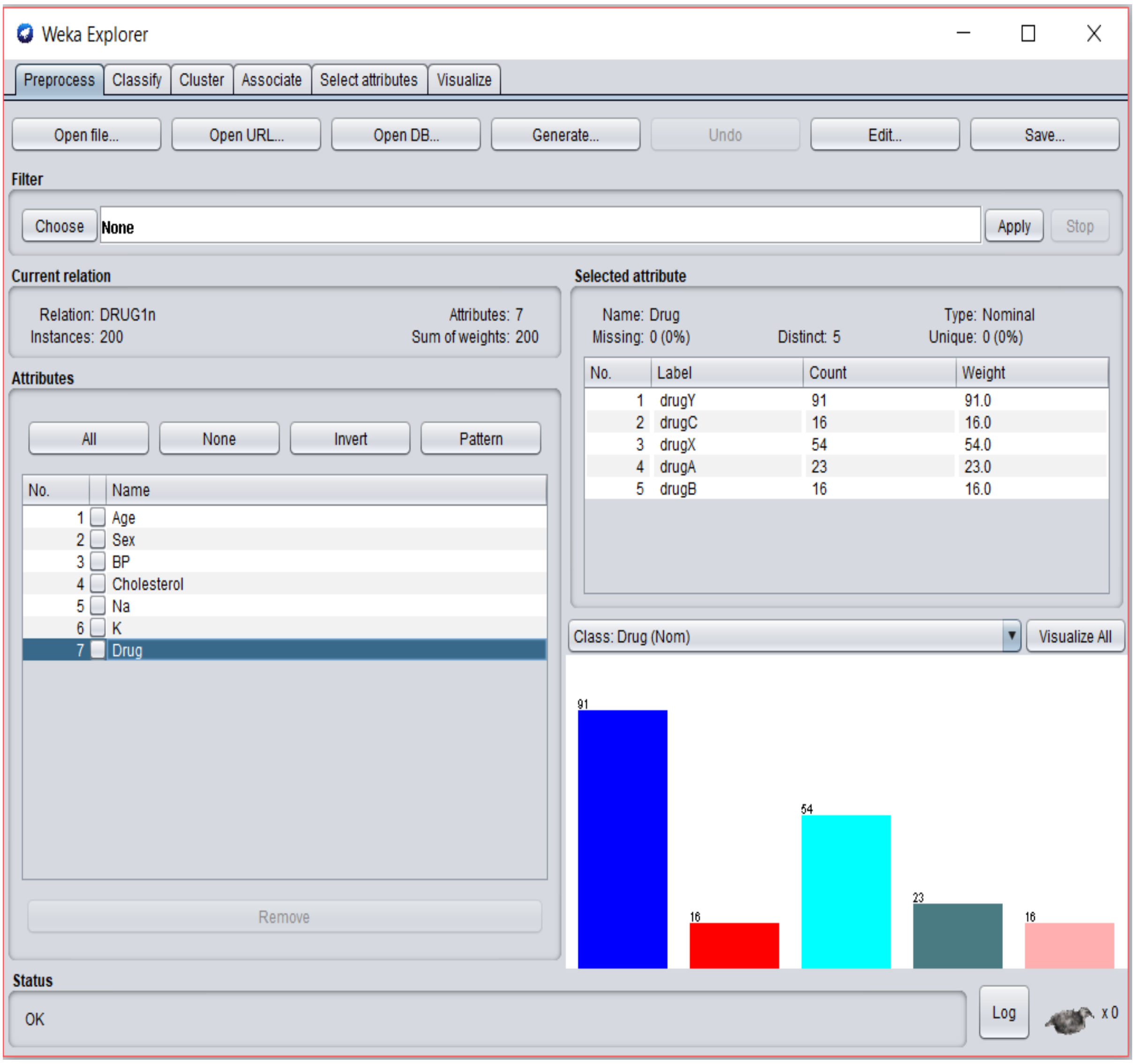

Figura 5.Cargando el dataset Drug1n

Luego vamos a la pestaña que dice classify y seleccionamos el método de árboles de decisión que en WEKA se llama j48 como se muestra en la Figura 5, en honor al algoritmo usado para generar un árbol de decisión desarrollado por Ross Quitan llamado C4.5, pero que su última versión fue la C4.8, y la j es debido a que WEKA está desarrollado en java de ahí su nombre j48.

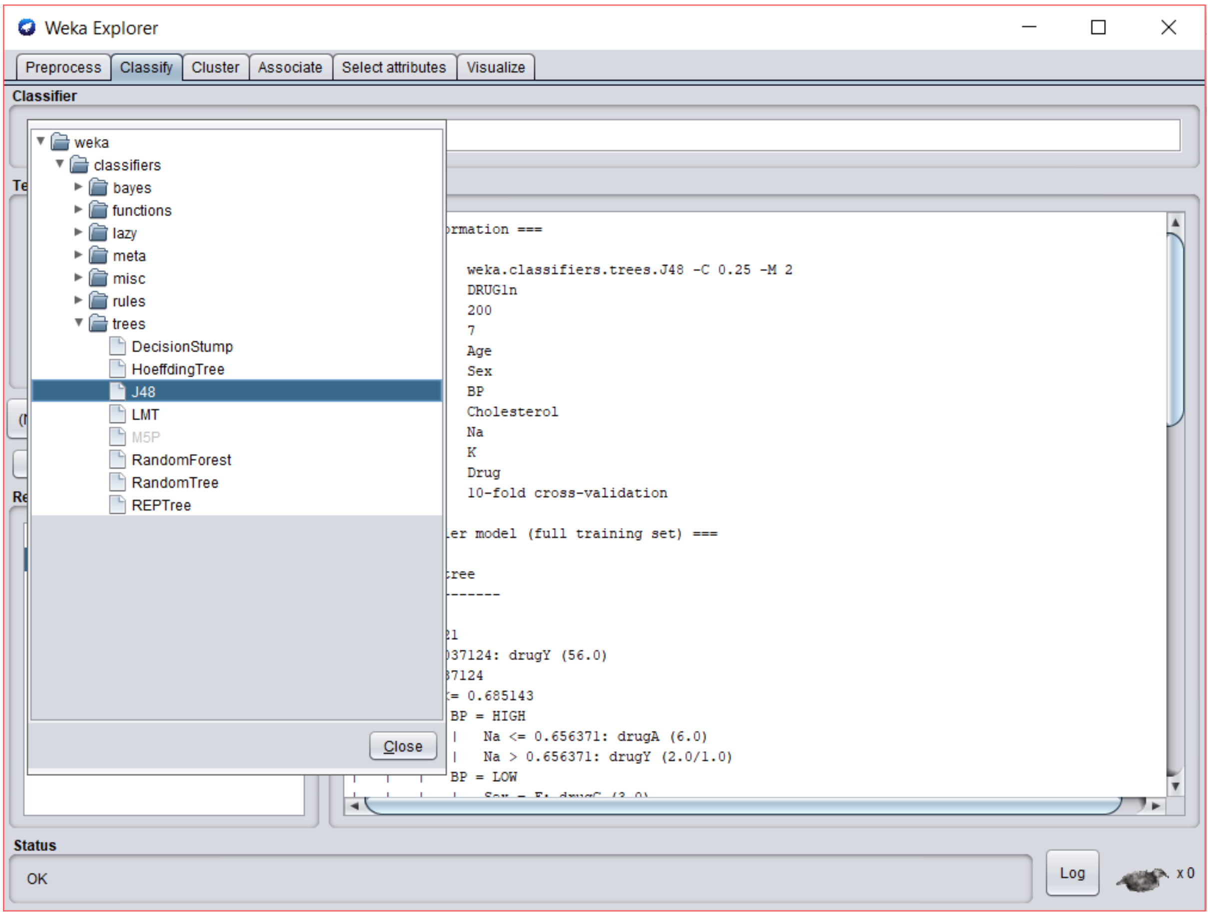

Figura 6. Seleccionando el método de árboles de decisión j48

Después de seleccionar el método de clasificación que en este caso fue el de árboles de decisión, podemos proceder a testear el árbol, tenemos 4 opciones para realizar su testeo “use training set” que sería testearlo con los mismo datos que elegimos para entrenarlo, “supplied test set” que sería testearlo con un dataset brindado por el usuario, “cross-validation” que sería realizar una validación cruzada de n pliegues y el usuario selecciona los pliegues que desea realizar y finalmente está “percentage split” que sería realizar en entrenamiento en un porcentaje de los datos dados anteriormente y con el otro porcentaje realizar el testeo, estas cuatro opciones se encontrarán en la parte superior izquierda como se muestra en la Figura 6.

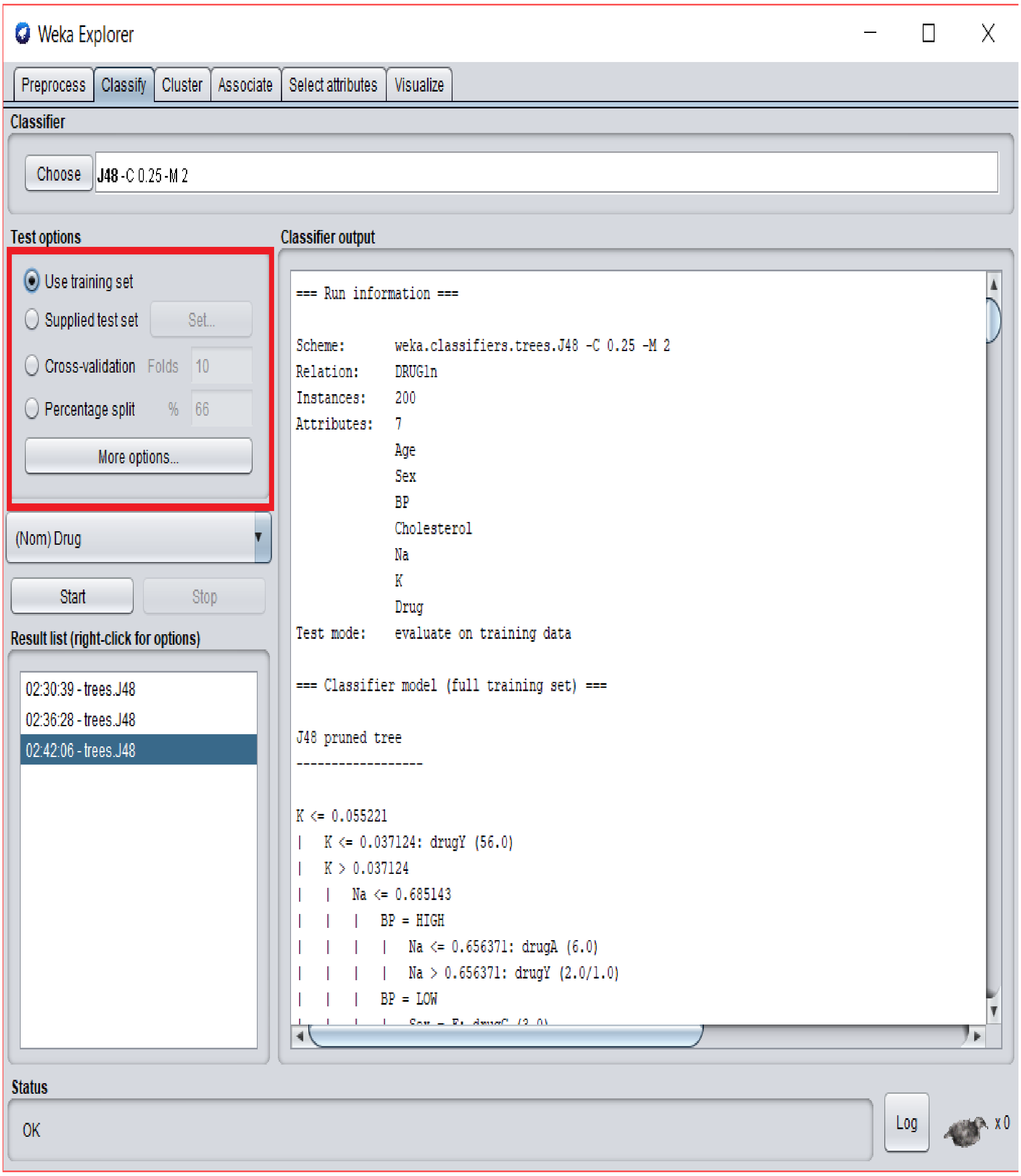

Figura 7.Opciones de testeo

Después de tener ya definida la opción de testeo podemos dar clic en el botón start

y en la parte de resultados nos aparecerá la hora en que se realizó el test y el método usado, adicionalmente en la ventana de salida del clasificador ubicada al lado derecho encontramos los datos relevantes del testeo, en esta ventana podremos encontrar las instancias del test, los atributos, una vista del árbol de decisión con sus condicionales como se muestra en la Figura 7, un resumen en donde se encuentra el porcentaje de acierto, errores y la matriz de confusión como se muestra en la Figura 8, además las predicciones que hizo el árbol sobre nuestros datos, pero estas predicciones se mostrarán en un documento en la sección de anexos anexo 3.

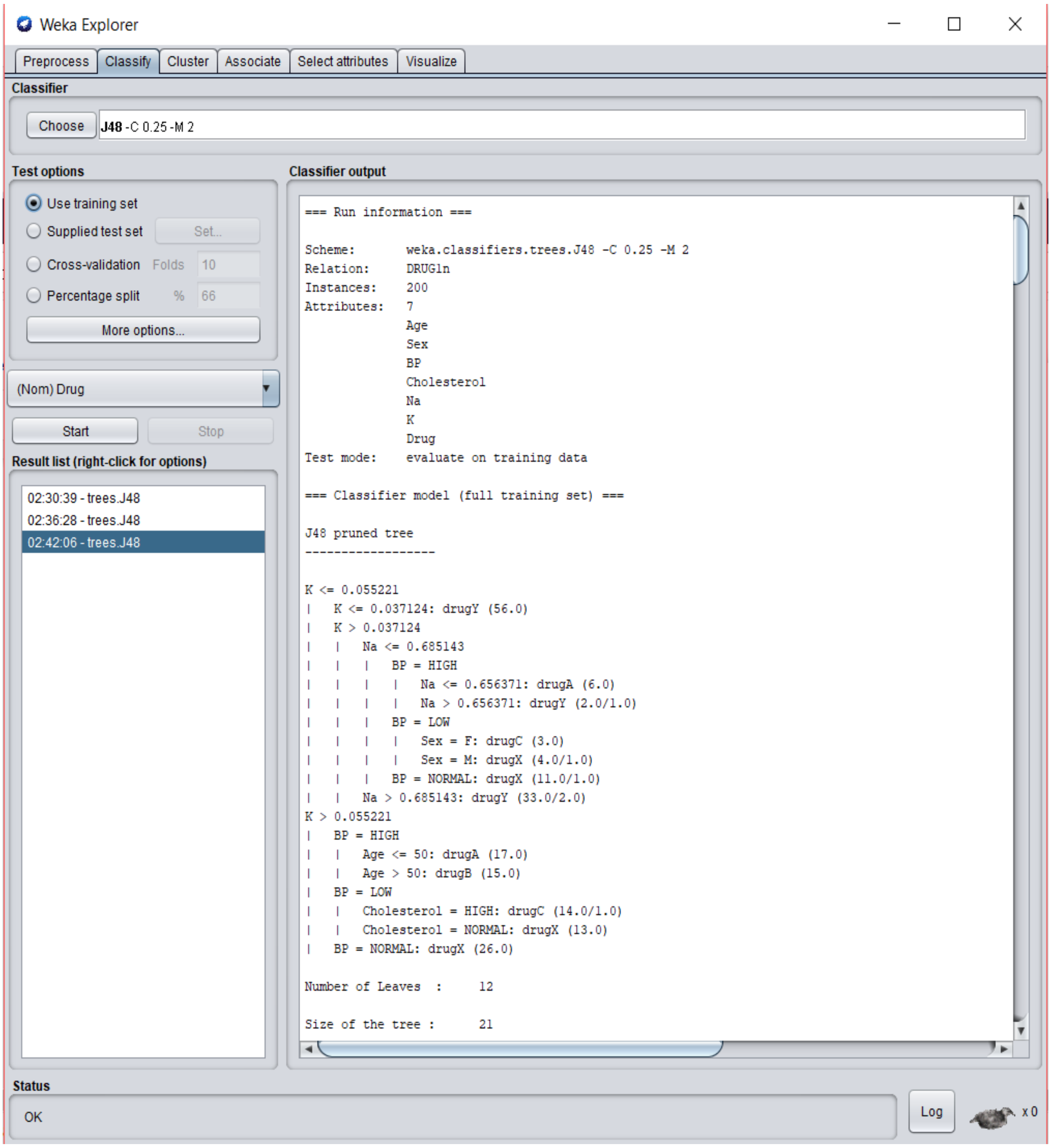

Figura 8.Salida del test 1

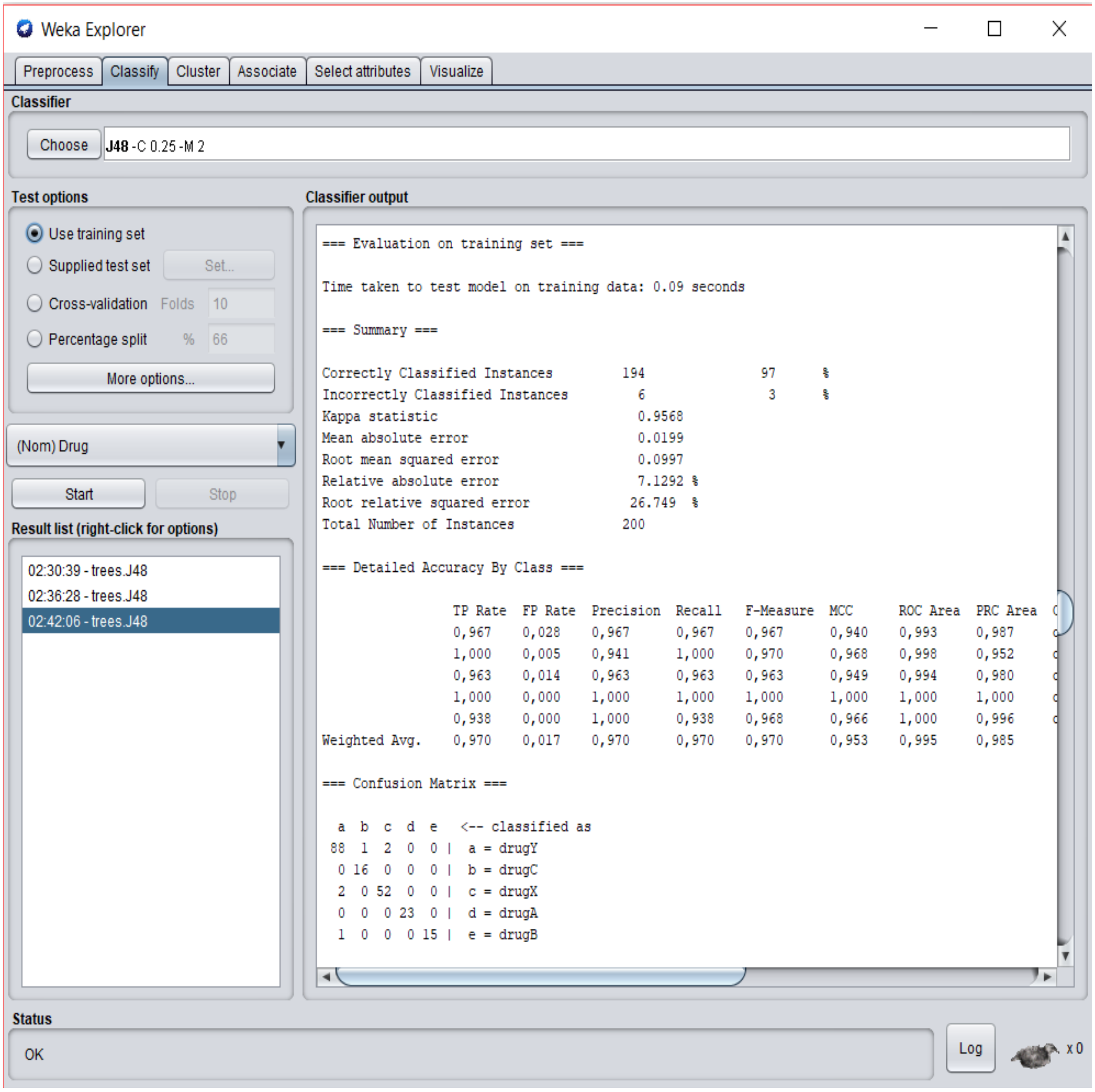

Figura 9.Salida del test 2

Además podremos visualizar el árbol gráficamente para que sea más fácil de comprender dando clic derecho sobre el test que se desea ver el árbol en la parte de lista de resultados y luego dando clic en visualizar árbol, como se muestra en la Figura 9 y el árbol se mostrará en la Figura 10.

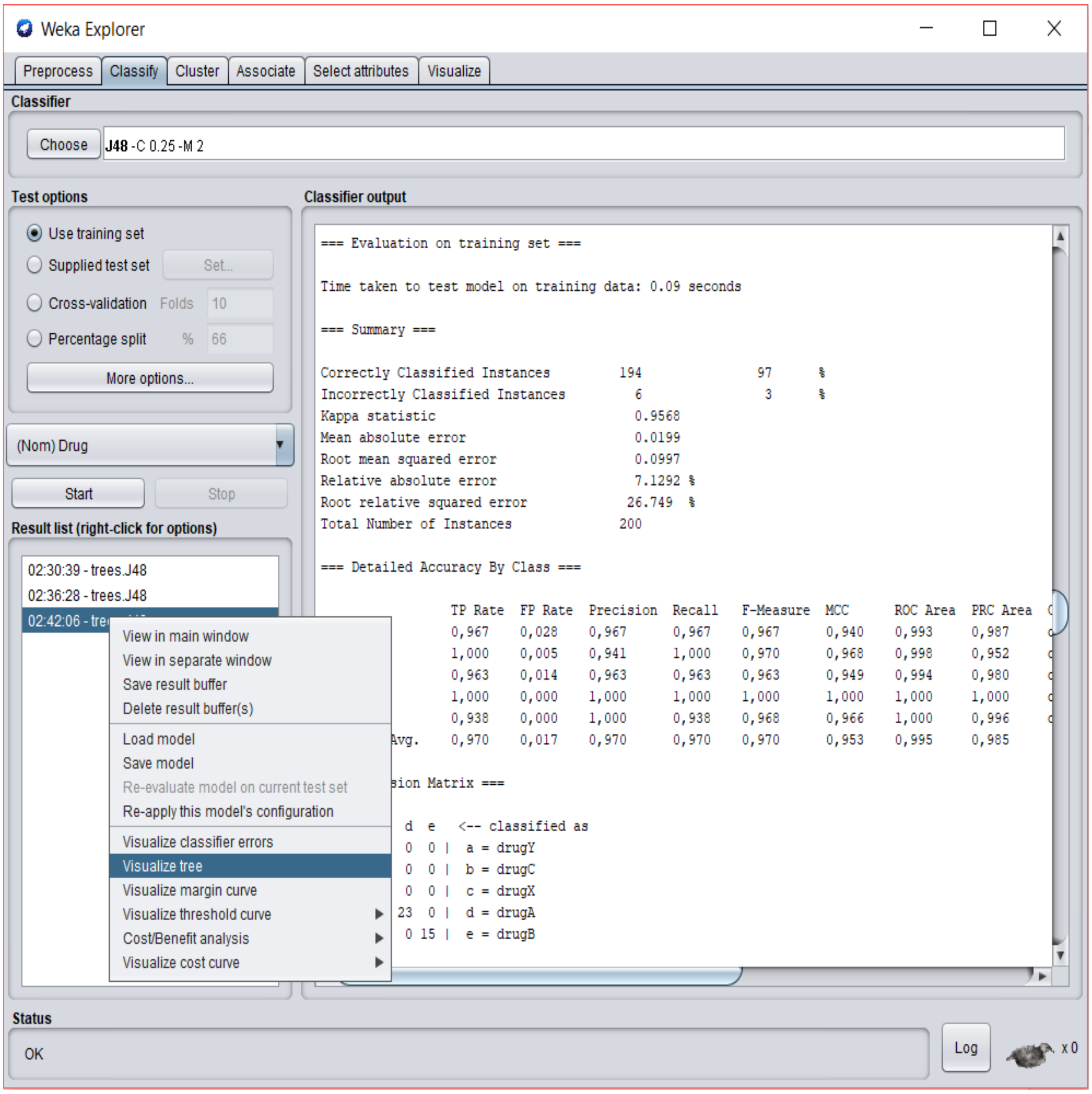

Figura 10.Visualizar el árbol de nuestro clasificador

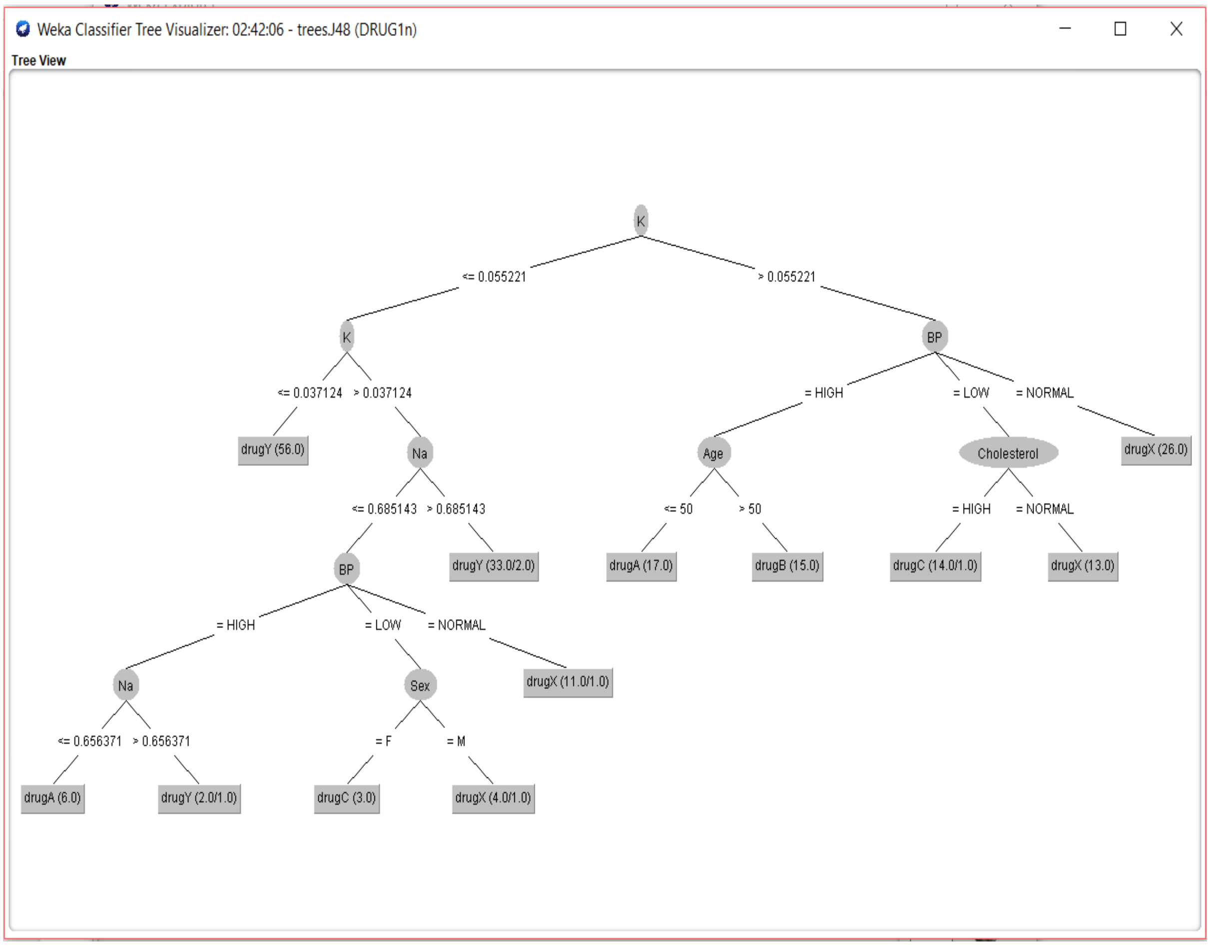

Figura 11.Visualización del árbol de decisión

Igualmente se puede obtener una visualización de los datos del test dando clic en la pestaña visualizar que es la última de izquierda a derecha como se muestra en la Figura 11.

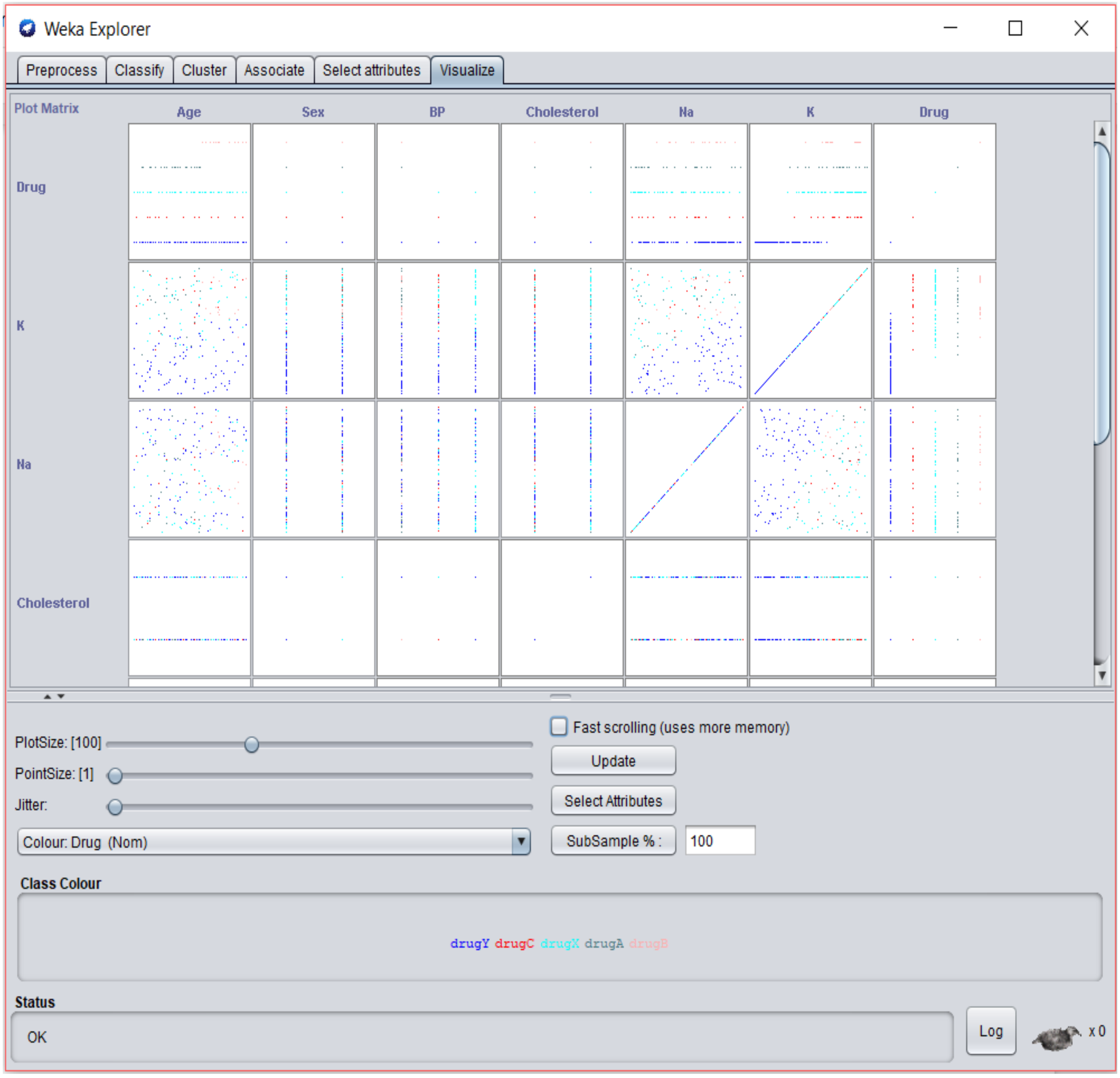

Figura 12.Visualización de todos los datos del test

Conclusiones del caso práctico

Al visualizar los datos que arrojó el test de forma detallada, se puede concluir que la droga más efectiva es la DrogaY ya que se puede administrar con éxito en 88 pacientes que son casi la mitad.

También es posible concluir que el árbol de decisión tiene un gran porcentaje de acierto que corresponde al 92% en adelante con cualquiera de los test usados, en este caso un 97% con el test que realizamos nosotros permitiéndonos afirmar que el árbol de decisión fue entrenado de manera correcta.

8. CONCLUSIONES

- En esencia, la analítica de datos nos deja observar un conjunto de tecnologías y arquitecturas diseñadas para conseguir un mejor rendimiento de grandes volúmenes de información.

- Con el anterior estudio nos encontramos con que la analítica de datos es un área en expansión, si hablamos de analítica de datos hablamos de Big Data y las posibilidades de desarrollo de algoritmos para nuevos datos y aplicaciones reales, nos encontramos con un campo que es un nicho de investigación y desarrollo en los próximos años.

- Los cuatro tipos de analítica están enfocados a responder las siguientes preguntas: ¿Qué está pasando? mediante la analítica descriptiva, ¿Por qué está pasando? mediante la analítica diagnostica, ¿Qué es lo más probable que pase? mediante la analítica predictiva, y finalmente ¿Qué necesito hacer? mediante la analítica prescriptiva.

- Al realizar las investigación de las herramientas libres más comunes de la analítica de datos, encontramos las 4 más populares que fueron expuestas en el documento RapidMiner, WEKA, Tableau, R de las cuales al realizar la comparativa, podemos afirmar que WEKA es la mejor herramienta para aquellos que deseen incursionar por primera vez en el mundo de la analítica de datos, ya que su facilidad de uso, la velocidad de funcionamiento, los pocos requerimientos y su interfaz amigable con el usuario, la hacen las mejor opción.

- En el caso práctico podemos evidenciar la facilidad de uso de WEKA, y aunque es fácil de utilizar no significa que sea menos potente, ya que para nuestro caso nos presentó un 97% de precisión en la predicción de los datos, permitiendo concluir que el entrenamiento del árbol de decisión fue un éxito.

- El factor clave para la obtención de beneficios de la analítica de datos, no depende de la capacidad tecnológica sino de la capacidad humana para una correcta interpretación de la información obtenida.

- Un trabajo a futuro de este proyecto podría ser un análisis más a profundidad sobre algoritmos de árboles de decisión en WEKA, como se estructura y su respectivo funcionamiento.

9. GLOSARIO

- **GUI**: La interfaz gráfica de usuario, conocida también como **GUI** (del inglés **graphical user interface**), es un programa informático que actúa de interfaz de usuario, utilizando un conjunto de imágenes y objetos gráficos para representar la información y acciones disponibles en la interfaz [21].

- **CLI**: interfaz de línea de comandos o interfaz de línea de órdenes (en inglés, **command-line** interface, **CLI**) es un método que permite a los usuarios dar instrucciones a algún programa informático por medio de una línea de texto simple [20].

- **SQL:** (por sus siglas en inglés Structured Query Language; en español lenguaje de consulta estructurada) es un lenguaje específico del dominio que da acceso a un sistema de gestión de bases de datos relacionales que permite especificar diversos tipos de operaciones en ellos [19].

- **CSV:** Los archivos **CSV** (del inglés *comma-separated values*) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal: Chile, Perú, Argentina, España, Brasil...) y las filas por saltos de línea.
  El formato CSV es muy sencillo y no indica un juego de caracteres concreto, ni cómo van situados los bytes, ni el formato para el salto de línea. Estos puntos deben indicarse muchas veces al abrir el archivo, por ejemplo, con una hoja de cálculo [18].

- **XML:** proviene de **extensible Markup Language** (**"Lenguaje de Marcas Extensible"**). Se trata de un **metalenguaje** (un **lenguaje** que se utiliza para decir algo acerca de otro) extensible de etiquetas que fue desarrollado por el **Word Wide Web Consortium** (**W3C**), una sociedad mercantil internacional que elabora recomendaciones para la **World Wide Web**.
  El **XML** es una adaptación del SGML (Standard Generalized Markup Language), un lenguaje que permite la organización y el etiquetado de documentos. Esto quiere decir

que el **XML** no es un lenguaje en sí mismo, sino un sistema que permite definir lenguajes de acuerdo a las necesidades [10].

- **K-means:** es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos [11].

- **ARFF:** Aunque WEKA acepta en teoría ficheros csv para obtener los datos de entrada e incluso soporta consultas a base de datos, yo he tenido problemas para cargar datos en los dos formatos, por lo que recomiendo crear nosotros mismos un fichero con formato ARFF (el formato propio de WEKA) [13].

- **Machine Learning:** es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. **La máquina que realmente aprende es un algoritmo** que revisa los datos y es capaz de predecir comportamientos futuros. *Automáticamente*, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana [12].

- **Clustering**: El término clúster (del inglés clúster, que significa grupo o racimo) se aplica a los conjuntos o conglomerados de ordenadores unidos entre sí normalmente por una red de alta velocidad y que se comportan como si fuesen una única computadora. La tecnología de clústeres ha evolucionado en apoyo de actividades que van desde aplicaciones de supe cómputo y software para aplicaciones críticas, servidores web y comercio electrónico, hasta bases de datos de alto rendimiento, entre otros usos [14].

- **GNU:** es un sistema operativo de tipo Unix desarrollado por y para el Proyecto GNU, y auspiciado por la Free Software Foundation. Está formado en su totalidad por software libre, mayoritariamente bajo términos de copyleft [15].

10. REFERENCIAS

[1] RapidMiner [Online].Disponible: https://rapidminer.com/.

[2] RapidMiner [Online]. Disponible: https://en.wikipedia.org/wiki/RapidMiner.

[3] RapidMiner Manual [Online]. Disponible: http://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf.

[4] WEKA [Online]. Disponible: http://isa.umh.es/asignaturas/crss/turorialWEKA.pdf

[5] WEKA Introduction [Online].Disponible: http://users.dsic.upv.es/~jorallo/docent/doctorat/weka.pdf

[6] Fundamentos de Big Data (Arcitura Education Inc. www.arcitura.com) Versión 1.7

[7] Tableau [Online]. Disponible: https://cetatech.ceta-ciemat.es/2015/11/introduccion-a-tableau/

[8] Herramienta R [Online].Disponible: ttps://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf

[9] Definición de análisis de datos [Online].Disponible: http://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos

[10] Definición de XML [Online] Disponible: https://definicion.de/xml/

[11] Definición de K-means [Online] Disponible: https://es.wikipedia.org/wiki/K-means

[12] Definición de Machine learning [Online] Disponible: http://cleverdata.io/que-es-machine-learning-big-data/

[13] Definición de Archivos ARFF [Online] Disponible:https://txikiboo.wordpress.com/2014/01/16/archivos-arff-weka/

[14] Definición de Clúster [Online] Disponible: https://es.wikipedia.org/wiki/Cl%C3%BAster_(inform%C3%A1tica)

[15] Definición de GNU [Online] Disponible: https://es.wikipedia.org/wiki/GNU

[16] Definición de Big Data [Online]. Disponible : http://searchdatacenter.techtarget.com/es/definicion/Analisis-de-big-data

[17] Data Mining [Online] Disponible: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/introduccion-DM.pdf

[18] Definición de CSV [Online] Disponible: https://es.wikipedia.org/wiki/Valores_separados_por_comas

[19] Definición de SQL [Online] Disponible: https://es.wikipedia.org/wiki/SQL

[20] Definición de CLI [Online] Disponible: https://es.wikipedia.org/wiki/Interfaz_de_l%C3%ADnea_de_comandos

[21] Definición de GUI [Online] Disponible: https://es.wikipedia.org/wiki/Interfaz_gr%C3%A1fica_de_usuario