Máster Universitario en Ciencias de la Salud: Investigación y Nuevos Retos

DISEÑO Y EJECUCIÓN DE LA INVESTIGACIÓN CUALITATIVA

TEMA 7

Análisis de datos a través de técnicas de minería de datos utilizando distintos softwares

Docente

Dra. María Consuelo Sáiz Manzanares
Departamento de Ciencias de la Salud

Índice de contenidos

I. Introducción

En este tema va a trabajar la aplicación de técnicas de tratamiento de datos desde la utilización de distintos softwares como Weka o SPSS para el análisis de datos en formato cualitativo. En concretos Weka permite el tratamiento de variables categorizadas en orden cualitativo desde la aplicación de técnicas complejas de aprendizaje automático, supervisado o no supervisado, ya que incorpora algoritmos específicos para poder procesar los datos en este formato. SPSS permite el tratamiento de estas variables en perfiles de complejidad computacional más baja como es el análisis de frecuencias, de porcentajes y la utilización de tablas cruzadas, que son también relevantes para la realización una primera aproximación descriptiva a los datos.

II. Objetivos

7.1. Conocer herramientas para el análisis de datos a través de la utilización de técnicas de minería de datos utilizando distintos softwares (SPSS, R, librerías).
7.2. Conocer y aplicar algunas Técnicas de aprendizaje automático.
7.2.1. Conocer y aplicar algunas Técnicas de aprendizaje automático supervisadas.
7.2.2. Conocer y aplicar algunas Técnicas de aprendizaje automático no supervisadas.

III. Contenidos específicos del tema

3.1. Preparación de los datos

Siguiendo lo visto en el Tema 5 relativo a la preparación de los datos para la ejecución de una investigación se va a realizar el protocolo para la inserción de una base de datos en distintas herramientas que van a permitir su procesamiento. El esquema queda claramente descrito por García, Luengo, & Herrera (2015), ver Figura 1.

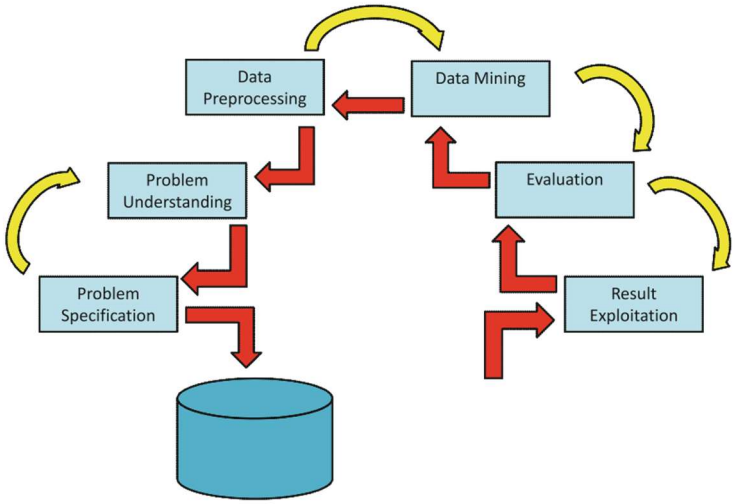


Figura 1. Knowledge Discovery in Databases (KDD), Tomada de García, Luengo, & Herrera (2015) p. 3.

Siguiendo pues el diseño de García, Luengo y Herrera (2015) se pueden diferenciar las fases siguientes:

Fase de preparación de los datos

Los datos antes de ser procesados tienen que prepararse de forma correcta para poder, posteriormente, aplicar distintos algoritmos, a esta acción se la ha denominado *Preprocessing*. Para ello el investigador tiene que responder a las siguientes preguntas:

1. ¿Cómo depurar los datos, la base de datos? Fase de Depuración de los Datos (*Data Cleaning*).

Las bases de datos requieren de un análisis de los valores, y de una detección, en su caso, de los valores que faltan, así como de un análisis de toda aquella información que no es relevante para el propósito de la investigación.

2. ¿Cómo detectar los datos precisos? Fase de Transformación de Datos (*Data Transformation*).

Los datos tienen que preparase para poder aplicar sobre ellos distintos algoritmos de Minería de Datos. Para ello el investigador tiene que saber qué datos necesita y cuál es el objetivo de su investigación.

3. ¿Cómo incluir e incorporar datos? Fase de Integración de Datos (*Data Integration*).

Implica la fusión de datos desde múltiples bases, este proceso se debe realizar cuidadosamente para evitar redundancias e inconsistencias en el resultado final. Las operaciones típicas realizadas dentro de la integración de datos son: identificación y unificación de variables y de dominios. Además, se tiene que efectuar un análisis sobre la correlación de atributos, la duplicación de tuplas[1] y la detección de conflictos entre los valores de datos que se puedan obtener desde diferentes fuentes.

4. ¿Cómo unificar la escala de datos? Fase de normalización de los datos (*Data Normalization*).

La unidad de medida utilizada puede afectar el análisis de datos. Por ello, es recomendable que todos los atributos se expresen en las mismas unidades de medida, para lo que hay que utilizar una escala o rango común que permita la comparativa entre los datos. La normalización de los datos intenta dar a todos los atributos el mismo peso y es particularmente útil en métodos estadísticos de aprendizaje.

5. ¿Cómo manejar los valores perdidos? Fase de eliminación de los valores perdidos (*Missing Data Imputation*).

El objetivo es el de eliminar los valores perdidos, se considera que es mejor realizar estimaciones por distintos procedimientos que dejar los valores en blanco. Se pueden utilizar distintas técnicas, por ejemplo: *media de series* (sustituye los valores perdidos con la media de la serie completa), *media de puntos cercanos* (sustituye los valores perdidos por la mediana de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos, por encima y por debajo del valor perdido, utilizados para calcular la mediana), *mediana de puntos cercanos* (sustituye los valores perdidos por la mediana de los valores válidos circundantes. La amplitud de los puntos adyacentes es el número de valores válidos, por encima y por debajo del valor perdido, utilizados para calcular la mediana), *interpolación lineal* (sustituye los valores perdidos utilizando una interpolación lineal. Se utilizan para la interpolación el último valor válido antes del valor

[1] Una Tupla es una lista ordenada de elementos.

perdido y el primer valor válido después del valor perdido. Si el primer o el último caso de la serie tiene un valor perdido, el valor perdido no se sustituye), y *tendencia línea en el punto* (Reemplaza los valores perdidos de la serie por la tendencia lineal en ese punto. Se hace una regresión de la serie existente sobre una variable índice escalada de 1 a *n*. Los valores perdidos se sustituyen por sus valores pronosticados).

6. ¿Cómo detectar y manejar el ruido? Fase de identificación del ruido (*Noise Identification*).

Este es paso se entiende como un paso de limpieza de datos, también es conocido como una fase de suavizado en la transformación de datos. Su objetivo principal es detectar errores aleatorios o variaciones en una variable medida. Es preferible la detección del ruido en lugar su eliminación. Una vez detectado se puede aplicar un proceso basado en la corrección que podría implicar algún tipo de operación subyacente (ver Figura 2).

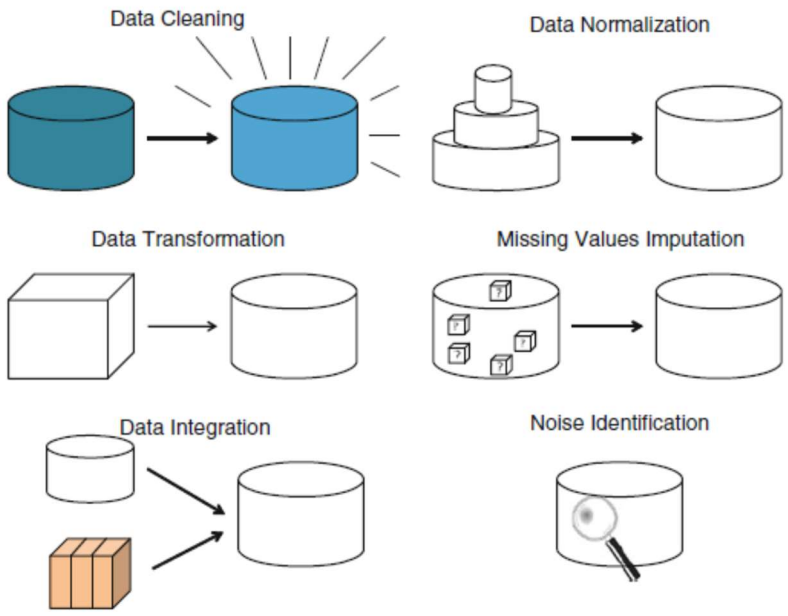


Figura 2. Preparación de los datos. Tomada de García, Luengo, & Herrera (2015) p. 12.

3.2. Fase de reducción de los datos

La reducción de datos hace referencia a un conjunto de técnicas que, de una manera u otra, obtienen una representación reducida de los datos originales. Es importante preparar los datos para poder aplicar técnicas de Minería de Datos. El no realizar este paso puede implicar una aplicación errónea de las Técnicas de Minería de Datos y en último término la obtención de resultados incorrectos.

La reducción de datos pasa por la selección de las características, la selección de instancias, el Data squashing y la discretización (Liu, Hussain, Tan., & Dash, 2002). En concreto la selección de instancias se puede realizar a través de distintos métodos: Sampling, Boosting, Selección de Prototipos o de aprendizaje basado en instancias y aprendizaje activo. Dentro de la selección de Prototipos se pueden distinguir: la selección basada en reglas de *k*nn, Selección basada en la eliminación ordenada, la utilización de algoritmos evolutivos y un muestreo aleatorio que tendría que ver también con el método (Reinartz, 2002) ver Figura 3 y Figura 4.

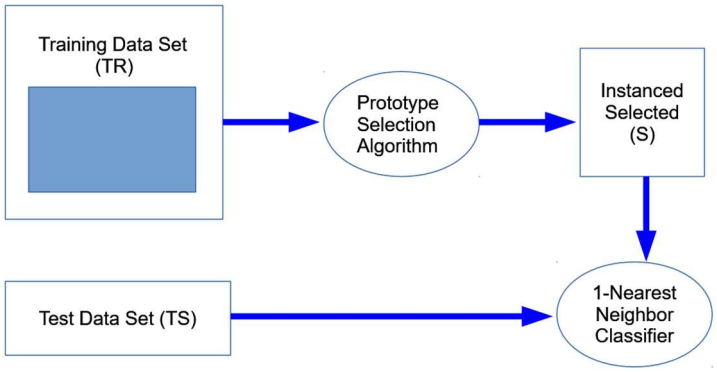


Figura 3. Proceso de selección de instancias. Tomado de Herrera, Riquelme y Ruíz (2004). Reunión nacional de Data Mining y Machine Learning. Madrid.

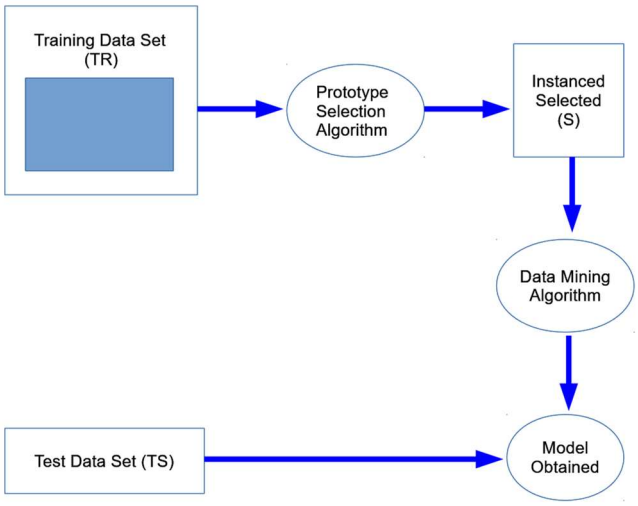


Figura 4. Proceso de selección de instancias. Tomado de Herrera, Riquelme y Ruíz (2004). Reunión nacional de Data Mining y Machine Learning. Madrid.

Fase de extracción de características y generación de instancias

La utilización de Sampling y de Selección de instancias permite la modificación interna de valores que representan a cada atributo. En la extracción de características, aparte de la eliminación de atributos o subconjuntos de atributos, también se puede realizar una fusión de los mismos y se puede contribuir a la creación de atributos artificiales de sustitución. Respecto de la generación de instancias, ésta permite la creación o ajustes artificiales que permitan representar mejor los límites de decisión en aprendizaje supervisado. Dentro de las Técnicas de Machine Learning se pueden distinguir los siguientes elementos: conceptos, instancias y atributos. Relativo a los conceptos se pueden diferenciar cuatro estilos de aprendizaje: aprendizaje de clasificación, aprendizaje de asociación, aprendizaje de agrupación, y aprendizaje de predicción. En relación, a las instancias cada una de ellas proporciona una entrada a la máquina el aprendizaje y se caracteriza por valores en un conjunto fijo o predefinido de características o atributos (Witten y Eibe, 2005). Por ejemplo, en una base de datos las instancias serían las filas y los atributos las columnas. Trasladado al plano de la investigación las instancias serían los sujetos y las columnas (variables independientes). Los atributos, estos se pueden medir en una escala nominal, ordinal o de intervalos.

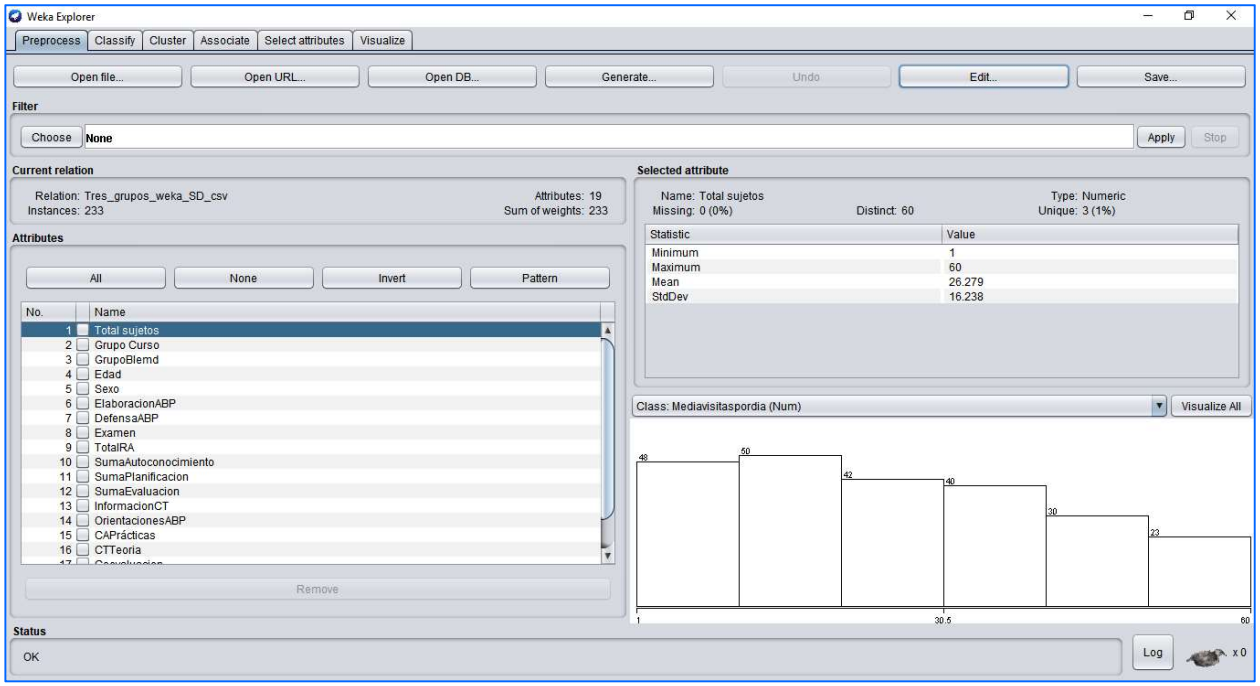


Figura 5. Determinación de atributos e instancias en Weka.

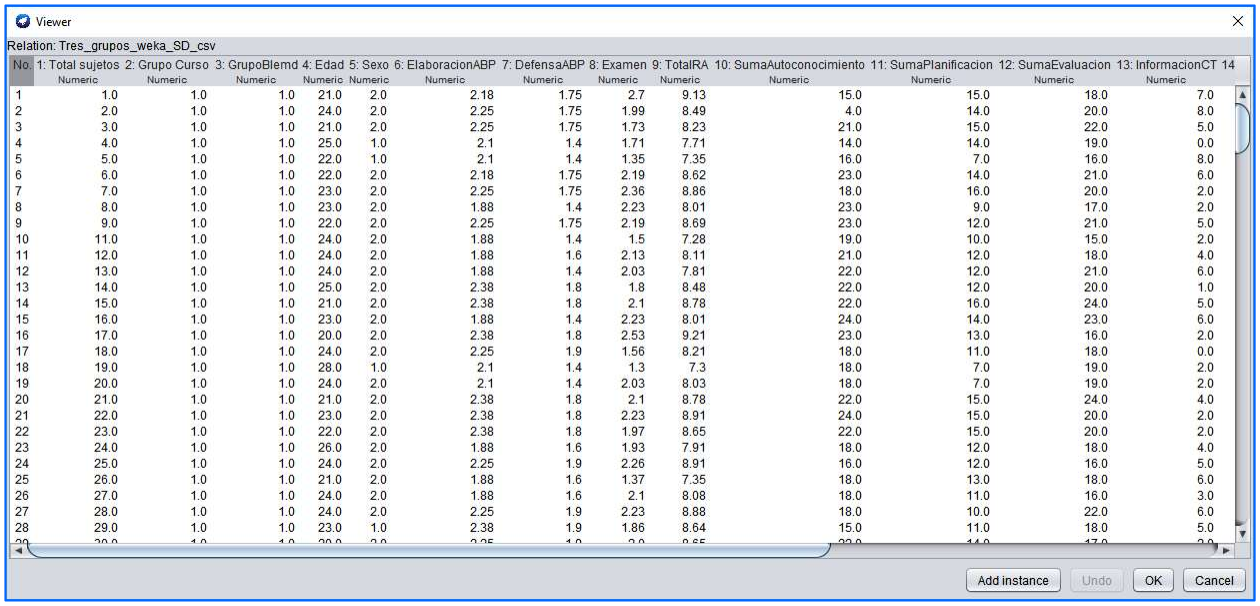


Figura 6. Determinación de atributos e instancias en una base de datos en Weka.

Fase de tratamiento de los datos

Para el análisis de los datos se pueden emplear técnicas estadísticas tradicionales y/ o técnicas de *Data Mining*. Una posible clasificación de las mismas es la propuesta por García, Luengo y Herrera (2015). Las dividen en técnicas de predicción y de descripción. Dentro de las primeras diferencian métodos estadísticos y métodos simbólicos. Dentro de los primeros incluyen: Modelos de Regresión (Regression Models), Redes Neuronales (Neural Networks), Técnicas de aprendizaje bayesiano (Bayesian Learning), Técnicas de aprendizaje basadas en instancias (Instance-based Learning) y Support Vector Machine. En los segundos incluyen Técnicas de aprendizaje de Reglas (Rule Learning) y Árboles de Decisión (Decision Trees). Asimismo, dentro de los métodos de descripción diferencian las Técnicas de Clustering y de las Técnicas de asociación de reglas (Association Rules). En la Figura 7 se puede consultar la diferenciación de técnicas realizada por García, Luengo y Herrera (2015).

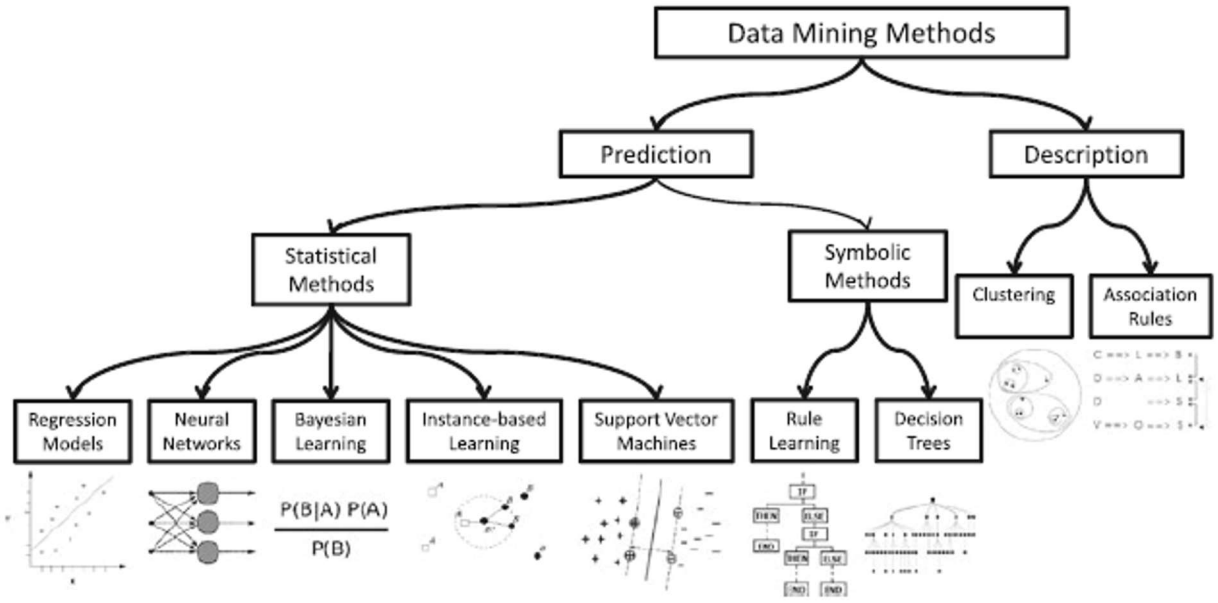


Figura 7. Data Mining Methods Figura Tomada de García, Luengo, & Herrera, F. (2015) p. 4.

3.3. Aplicación de Técnicas de Minería de Datos en el ámbito de las Ciencias Sociales y de la Salud

3.3.1. Clasificación de Técnicas de Machine Learning de Predicción y de descripción.

Seguidamente se van a describir las técnicas más representativas de Machine Learning: Redes neuronales, Técnicas bayesianas, Aprendizaje basado en instancias, Support Vector Machine, Rule Learning, Árboles de Decisión, Clustering y Reglas de asociación, siguiendo la definición realizada por García, Luengo y Herrera (2015).

Técnicas de predicción

Estas se dividen en métodos estadísticos y en métodos simbólicos.

Métodos estadísticos.

Redes Neuronales (Artificial Neural Networks-ANNs): Esta técnica tiene una función esencialmente predictiva. Pueden diferenciarse las técnicas de Multi-Layer Perceptron (Perceptrón Multicapa) MLP, Técnicas Neuronales de Base Radial (Basis Funtion Networks Techniques -RBFNs-). Dichas técnicas requieren atributos numéricos que no tengan valores perdidos. Una vez cumplido este requisito esta técnica es robusta contra valores atípicos y ruido.

Aprendizaje bayesiano (Bayesian Learning): Se basan en el teorema de Bayes, la técnica más utilizada dentro de ellas es el método Naïve Bayes. Dicho método asume que el efecto de un valor de atributo de una clase dada es independiente de los valores de otros atributos. Las definiciones iniciales de estos algoritmos solo funcionan con atributos categóricos, debido a que el cálculo de probabilidad solo puede hacerse con variables discretas. Además, la suposición de independencia entre los atributos es muy sensible a la redundancia y a la utilidad de algunos de los atributos y ejemplos de los datos. Estas técnicas no pueden trabajar con valores perdidos. Además, de Naïve

Bayes, también hay otros modelos complejos basados en estructuras de dependencia como las Redes Bayesianas.

Instancias basadas en aprendizaje (Instancia-based Learning): Aquí, los ejemplos se almacenan textualmente, la distancia de una función se utiliza para determinar qué miembros de la base de datos están más cerca de uno nuevo con un grado de predicción deseable. Utiliza distintos algoritmos para encontrar los ejemplos más cercanos las técnicas más usadas son los Árboles de decisión (KD-Trees) o la técnica del vecino más cercano K- Nearest Neighbor (*k*NN). Este último algoritmo es el más aplicado y útil, aunque tiene varios inconvenientes, como son los altos requisitos de almacenamiento, la baja eficiencia en la respuesta de predicción y el bajo nivel de tolerancia al ruido. Por ello, esta técnica puede mejorar a través de la fase de reducción de datos.

Máquinas de Soporte de Vectores (Support Vector Machine): Son un conjunto de algoritmos que se basan en la teoría del aprendizaje. Esta técnica es similar a las Redes Neuronales, ya que se emplea para la estimación. Dicha técnica funciona muy bien cuando los datos se pueden separar de forma lineal. No obstante, requiere que no haya valores perdidos y parece ser una técnica robusta contra ruido y los valores atípicos.

Métodos simbólicos.

Reglas de aprendizaje (Rule Learning): Estas técnicas buscan una regla que explique parte de los datos y separe estos ejemplos. En términos generales esta técnica requiere de datos nominales o discretizados (aunque esta tarea con frecuencia está implícita en el algoritmo) y procura una selección de atributos interesantes sobre los datos. No obstante, los valores atípicos pueden perjudicar el rendimiento del modelo final. Ejemplos de estos los modelos son los algoritmos más utilizados son AQ, CN2, RIPPER, PART y FURIA.

Árboles de Decisión (Decision Trees): Comprende modelos predictivos formados por iteraciones de una división que conlleva un esquema de decisiones jerárquicas. Funcionan al intentar dividir los datos usando una de las variables independientes en subgrupos homogéneos. La forma final del árbol se puede traducir a un conjunto de reglas If-Then-Else desde la raíz a cada uno de los

nodos hoja. Los algoritmos más utilizados son CART, C4.5 y PUBLIC, un ejemplo gráfico de los mismos se puede observar en la Figura 8 y Figura 18 y Figura 20.

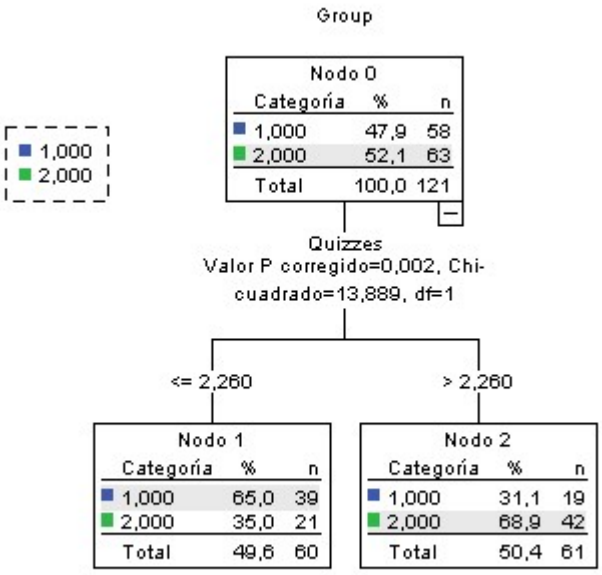


Figura 8. Un ejemplo de árbol de decisión con SPSS.

Métodos de descripción

Técnicas de Clustering: Se utilizan cuando no hay información de clase que predecir, pero los datos se tienen que dividir en grupos naturales o grupos. Dichos grupos incluyen subgrupos de ejemplos que comparten algunas propiedades o tienen algunas similitudes. Trabajan calculando una medida de distancia multivariada entre las observaciones que están más estrechamente relacionadas. Se pueden diferenciar distintas categorías: agrupamiento aglomerativo, agrupamiento divisivo y particionamiento. Los dos primeros son tipos jerárquicos de agrupamiento opuesto uno a otro. El último se basa en la agrupación. El más utilizado es el algoritmo de *k*-medias (*k*-means). Esta técnica comienza con un número *k* de clústeres y agrega o elimina iterativamente ejemplos hasta que logra una mejora basada en una minimización de intra y / o inter-medida de distancia del racimo. Esta técnica es sensible a los valores perdidos, el ruido y los valores atípicos.

Técnicas de Asociación de Reglas (Association Rules): Las reglas de asociación son un conjunto de técnicas que buscan encontrar relaciones de asociación entre los datos. La aplicación típica de estos algoritmos es el análisis de datos de transacciones minoristas.

3.3.2. Clasificación de Técnicas de Machine Learning en Técnicas Supervisadas y no supervisadas.

Otro posible esquema de distribución de las distintas Técnicas de aprendizaje es el que diferencia entre técnicas de Machine Learning de Aprendizaje Supervisado (Supervised Learning) y de Aprendizaje no supervisado (Unsupervised Learning). Dentro de las primeras se encuentran las Técnicas de Clasificación y dentro de ellas se incluyen: Support Vector Machine, Análisis discriminante (Discriminat Analysis), Naïve Bayes y Nearest Neigbbor. Y las Técnicas de Regresión y dentro de ellas se diferencian las técnicas de Regresión lineal (Linnear Regression), Support Vector Machine, Árboles de Decisión y Neural Networks. Asimismo, dentro de las Técnicas de Aprendizaje no supervisado se incluyen las Técnicas de Clustering y a su vez dentro de ellas las técnicas de *k*-means, *k*-medoids, S-Means, Gaussian Mixture y Neural Networks (ver Figura 9).

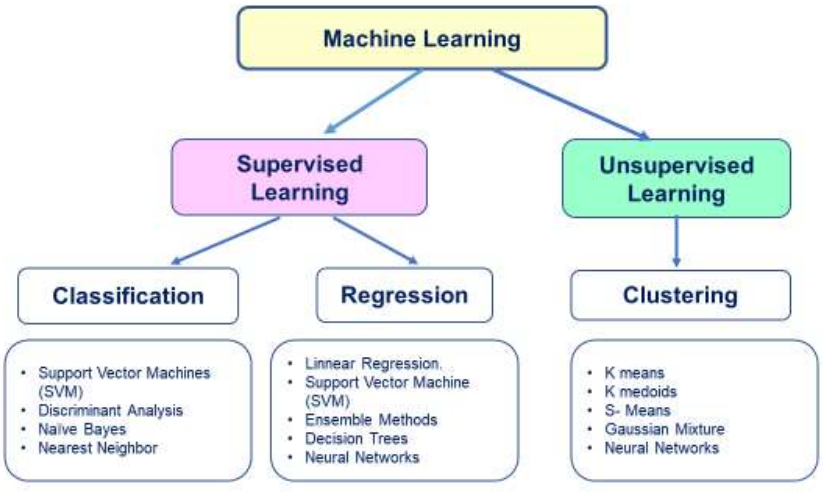


Figura 9. Clasificación de las Técnicas de Machine Learning.

3.3.2.1. Técnicas de Aprendizaje Supervisado.

Pretenden descubrir las relaciones entre los atributos de entrada (a veces llamados variables o características) y un atributo de destino (a veces referido como clase). La relación que se busca se representa en una estructura llamada modelo. Este se describe como datos ocultos que se explicitan en una predicción del valor del atributo de destino, cuando los valores de los atributos de entrada son conocidos. El aprendizaje supervisado se utiliza en campos de las ciencias de la salud y de la ingeniería entre otros.

En un escenario típico de aprendizaje supervisado, se puede usar para predecir ejemplos no vistos. Este entrenamiento el conjunto se puede describir de varias formas. El más común es describirlo por un conjunto de instancias, que es básicamente una colección de tuplas que pueden contener duplicados. Cada tupla se describe mediante un vector de valores de atributo. Cada atributo tiene un asociado dominio de valores que se conocen antes de la tarea de aprendizaje. Los atributos son típicamente uno de dos tipos: nominal o categórico (cuyos valores son miembros desordenados) o numérico (los valores son números enteros o reales, y se asume un orden). Los atributos nominales tienen una cardinalidad finita, mientras que los dominios de atributos numéricos son delimitados por los límites inferior y superior. El espacio de instancia (el conjunto de posibles ejemplos) se define como un producto cartesiano de todos los dominios de atributos de entrada.

Los dos problemas básicos y clásicos que pertenecen a la categoría de aprendizaje supervisado son clasificación y regresión. En clasificación, el dominio del objetivo atributo es finito y categórico. Es decir, hay un número finito de clases o categorías para predecir una muestra y son conocidos por el algoritmo de aprendizaje. Un clasificador debe asignar una clase a un ejemplo no visto cuando está entrenado por un conjunto de datos de entrenamiento. La naturaleza de la clasificación es discriminar unos ejemplos de otros, logrando como una predicción fiable para poder hacer predicciones correctas para nuevas instancias. Sin embargo, cuando el atributo de destino está formado por valores infinitos, como en el caso de predecir un número real entre un intervalo, se debe utilizar un modelo de regresión (García, Luengo y Herrera, 2015).

3.3.2.2. Técnicas de aprendizaje no supervisado

En el aprendizaje supervisado, el objetivo es obtener un mapa entrada y de salida cuyos valores correctos y definidos son proporcionados por un supervisor. Sin embargo, en el aprendizaje no supervisado, no existe tal supervisor y solo hay datos de entrada disponibles. El objetivo en este caso es el de encontrar regularidades, irregularidades, relaciones, similitudes y asociaciones en la entrada. Con las Técnicas de aprendizaje no supervisado, es posible aprender más y aplicar modelos más complejos que con las Técnicas de aprendizaje supervisado. La razón es que en las Técnicas supervisadas el aprendizaje trata de encontrar la conexión entre dos conjuntos de observaciones y la dificultad de la tarea de aprendizaje aumenta exponencialmente con la cantidad de pasos entre los dos conjuntos, esta es la razón por la que el aprendizaje supervisado no puede en la práctica, aprender modelos con jerarquías profundas (García et al., 2015).
En aprendizaje no supervisado dentro de las Técnicas de Clustering se pueden diferenciar siguientes (García et al., 2015):

Pattern Mining:

La minería de patrones incluye patrones extraños o negativos. Las reglas de búsqueda de patrones pueden incluir multinivel, multidimensional, patrones aproximados, inciertos, comprimidos, raros /negativos y de alta dimensión. Tanto en correlación como en excepción.

Detección de los valores perdidos (Outlier Detection):

Facilita la detección de datos con comportamientos que son muy diferentes de las expectativas. La detección de valores atípicos intenta localizar aquellos casos excepcionales que presentan desviaciones de los patrones de la mayoría.

Técnicas de Process mining

Es una técnica de minería de datos que se aplica sobre los registros y trabajan con un flujo de procesos a partir de los registros realizados por muchas aplicaciones (Trcka y Pechenizkiy, 2009) un ejemplo de aplicación diferencial entre el procedimiento tradicional y la técnica de *Process mining* se puede comprobar en la Figura 10.

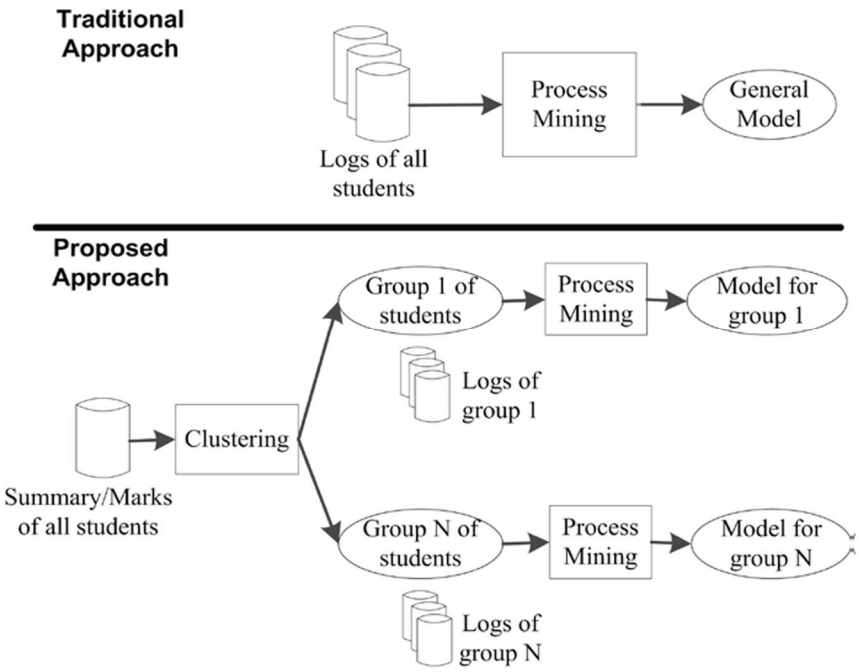


Figura 10. Esquema de análisis de datos con un procedimiento tradicional y con uno de *Process Minin*g. Tomado de Bogarín, Romero y Cerezo (2016) p. 77.

Como ya se ha visto anteriormente, previamente al análisis de los datos se deberá realizar una depuración de las bases de datos con el fin de eliminar datos que no son relevantes para el estudio. Seguidamente, se puede aplicar un algoritmo con Heuristics Miner de ProM, es una red heurística dibujada como un grafo cíclico que muestra el comportamiento más habitual de los estudiantes y también técnicas de ajuste como *Goodness-of-fit indices* o Fitness estas técnicas indican la diferencia entre el comportamiento observado en el registro y el comportamiento descrito por el modelo de proceso. Un ejemplo de análisis de grafos se puede consultar en la Figura 11. En este caso se observa el comportamiento de los estudiantes en tres semanas (1, 2 y 3) en tres variables. Para realizarlo se utilizó el software libre Grafos de Rodríguez Villalobos (2012).

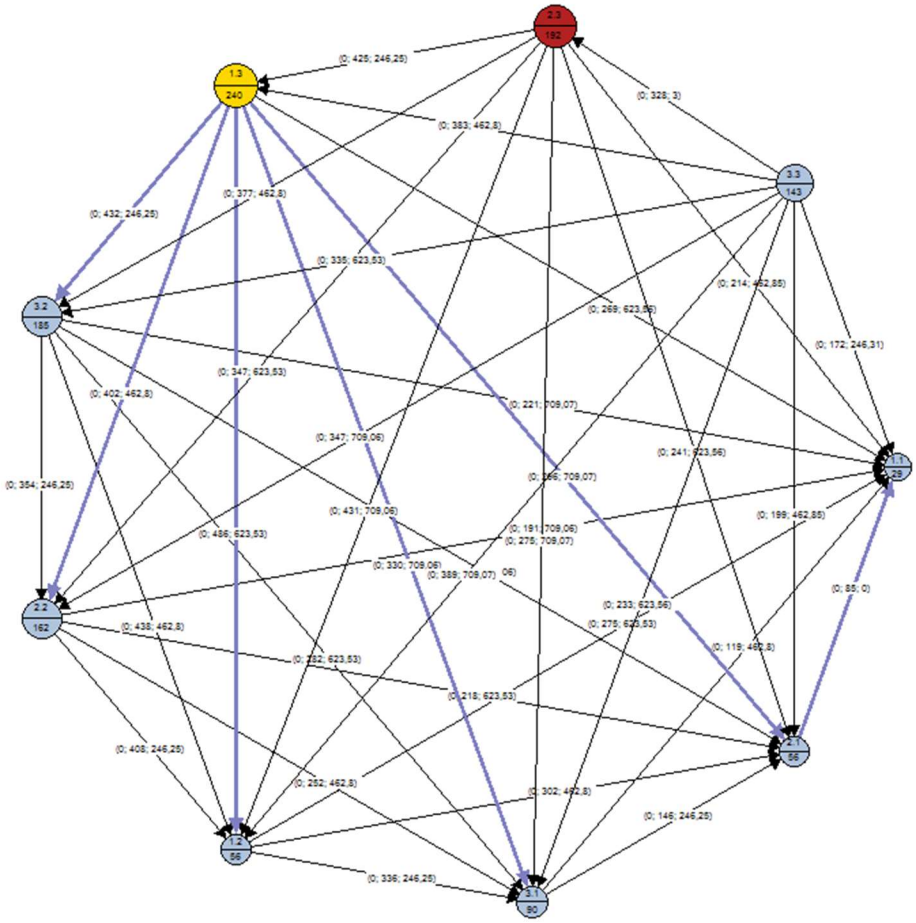


Figura 11. Análisis de patrones de comportamiento de los estudiantes en tres semanas de interacción en la plataforma en distintas variables de observación.

También, podría utilizar Técnicas de Minería de Datos (DM) de aprendizaje supervisado de clasificación y de regresión (Bogarín, Romero y Cerezo, 2017b):

- SPSS en sus paquetes de clustering.
- Weka aplicando la técnica de clustering.
- ProM (Van Der Aalst, 2011) Heuristics Miner, analiza la probabilidad desde el cálculo de frecuencias o relaciones entre las tareas y los constructos de dependencia/frecuencia en tablas y en gráficos.

- *Medidas de la teoría de grafos* (número total de nodos y número total de enlaces) para analizar el nivel de complejidad de los modelos obtenidos.
- *Intention Mining (IM*), es un análisis de procesos que se focaliza en el análisis del razonamiento a través del análisis de las actividades.
- *Alpha Miner (AM)*: Descubre cuál será el mejor algoritmo, si bien las limitaciones de esta técnica es que no usa frecuencias y por ello sólo es adecuado cuando analiza eventos con ruido, es muy poco frecuente en el aprendizaje de datos.
- *Sequence Pattern Mining (SPM)*, es una técnica común en DM y descubre las subsecuencias comunes, encuentra relaciones entre acontecimientos sucesivos. Puede analizar episode mining (EP) se basan en los *t-pattern analysis* de los modelos de Markov. SPM se utiliza para analizar las conductas de aprendizaje de los estudiantes. Esta técnica no es adecuada para hallar patrones de aprendizaje sino cuando se analiza un comportamiento simple.
- *Graph Mining (GM)* también denominada sub-graph mining hay que diferenciar esta técnica de la de network analyisis (SNA) esta última puede ser considerada una parte de GM.
- *Genetic algorithm*: esto proporciona modelos de proceso construido sobre matrices causales (entrada y salida) y dependencias para cada actividad. Este enfoque aborda problemas tales como ruido, datos incompletos, constructos de libre elección, actividades ocultas, concurrencia y actividades duplicadas.
- *Fuzzy miner*: Es un proceso de descubrimiento de algoritmos se utiliza para abordar problemas con un número grande de números y actividades que no están muy estructuradas.
- *Social Network Analysis Technique (SNA)*, es una técnica sociométrica que analiza las redes de interacción social, consiste en nodos que representan la organización en entidades y arcos.
- *Visualización de procesos*, permite desarrollar una interfaz de visualización del análisis de datos.

Esta forma de análisis permite el seguimiento de cada uno de los grupos detectados y por ende la puesta en marcha de orientaciones personalizadas para cada uno de ellos. Asimismo, la combinación de las técnicas de EDM y de PM van a servir al docente para estudiar el proceso de comportamiento de todos los estudiantes desde el inicio del desarrollo de la docencia, con el objetivo

último de poder adaptarla a las necesidades de cada grupo una extracción de los datos e implementación del PM se puede comprobar en la Figura 12 y un posible análisis del proceso en la Figura 13 y un estudio de la red heurística de comportamiento entre los estudiantes que aprueban y los que suspenden (ver Figura 14 y Figura 15).

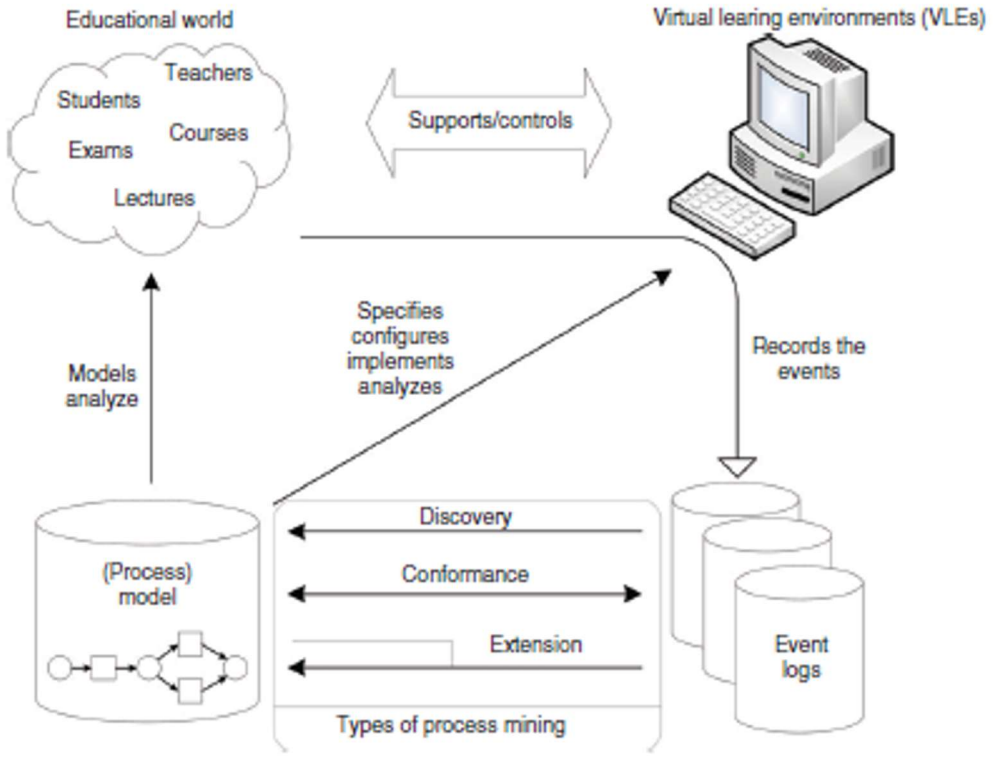


Figura 12. Educational Process Mining. Bogarín, Cerezo, & Romero (2017a) p. 4.

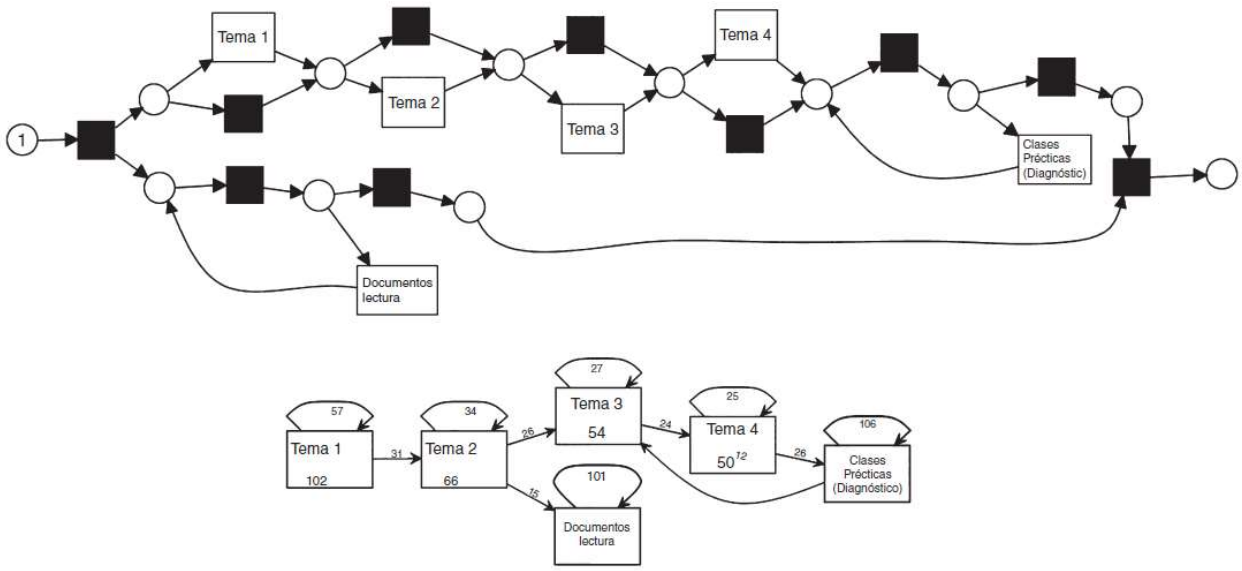


Figura 13. Ejemplos de análisis de un patrón de conducta en la plataforma aplicando Petri y Heuristic Net. Tomado de Bogarín, Cerezo, & Romero (2017a) p. 8.

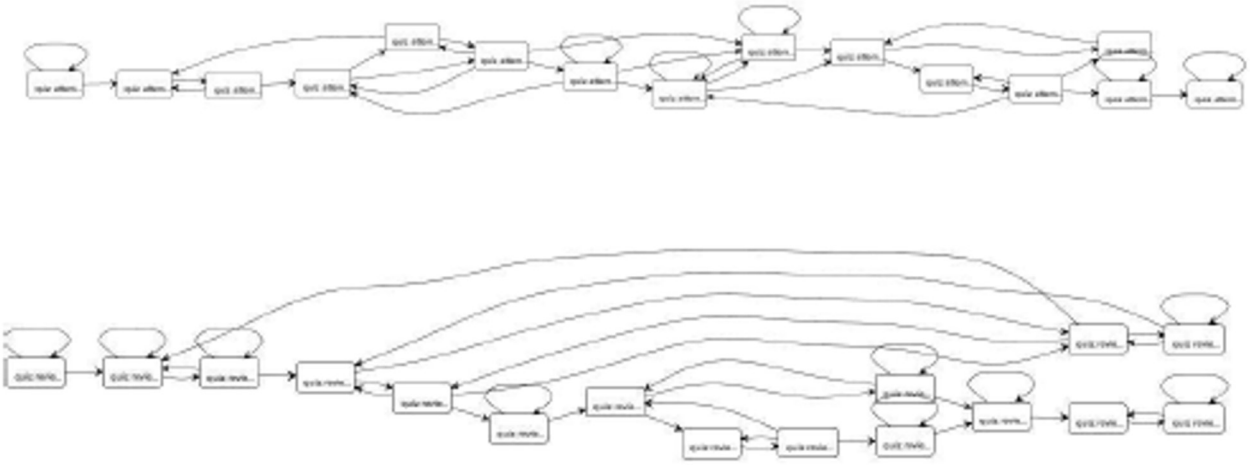


Figura 14. Ejemplo de un análisis heurístico de los estudiantes que aprueban. Tomado de Bogarín, Romero, & Cerezo (2016) p. 88.

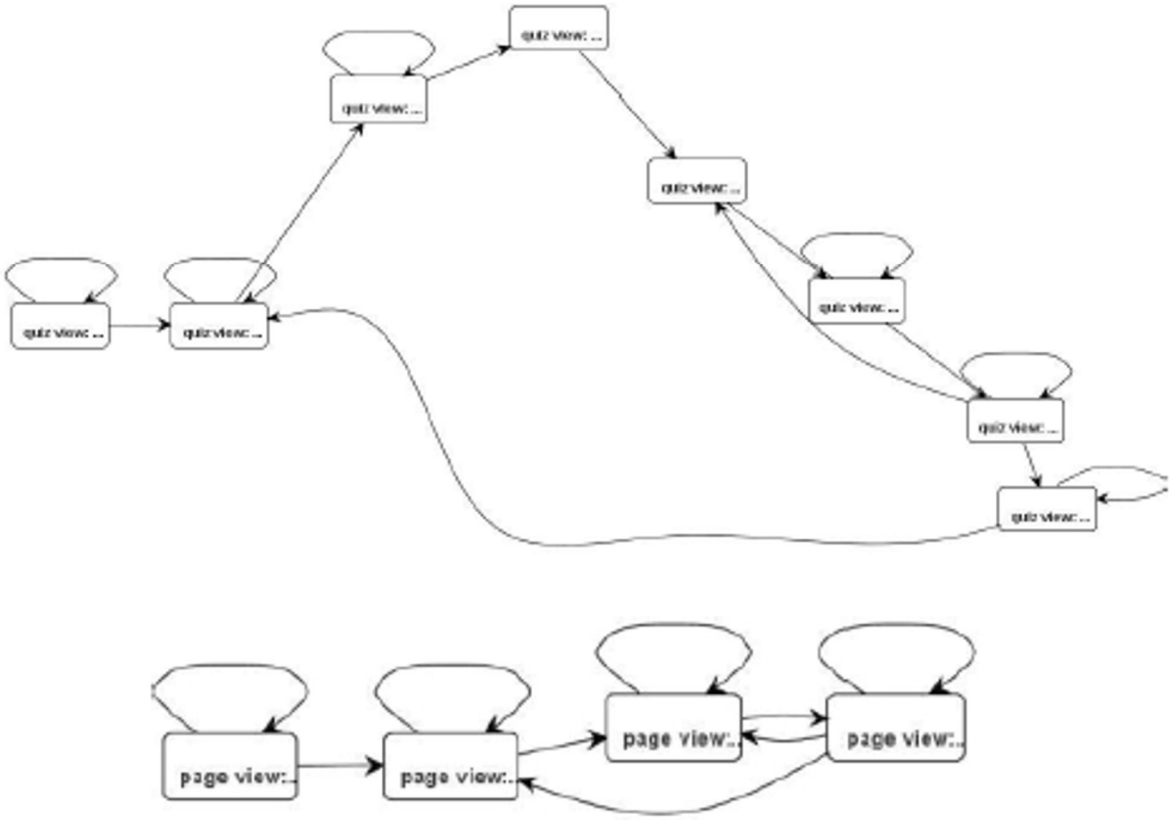


Figura 15. Ejemplo de un análisis heurístico de los estudiantes que suspenden. Tomado de Bogarín, Romero, & Cerezo (2016) p. 88.

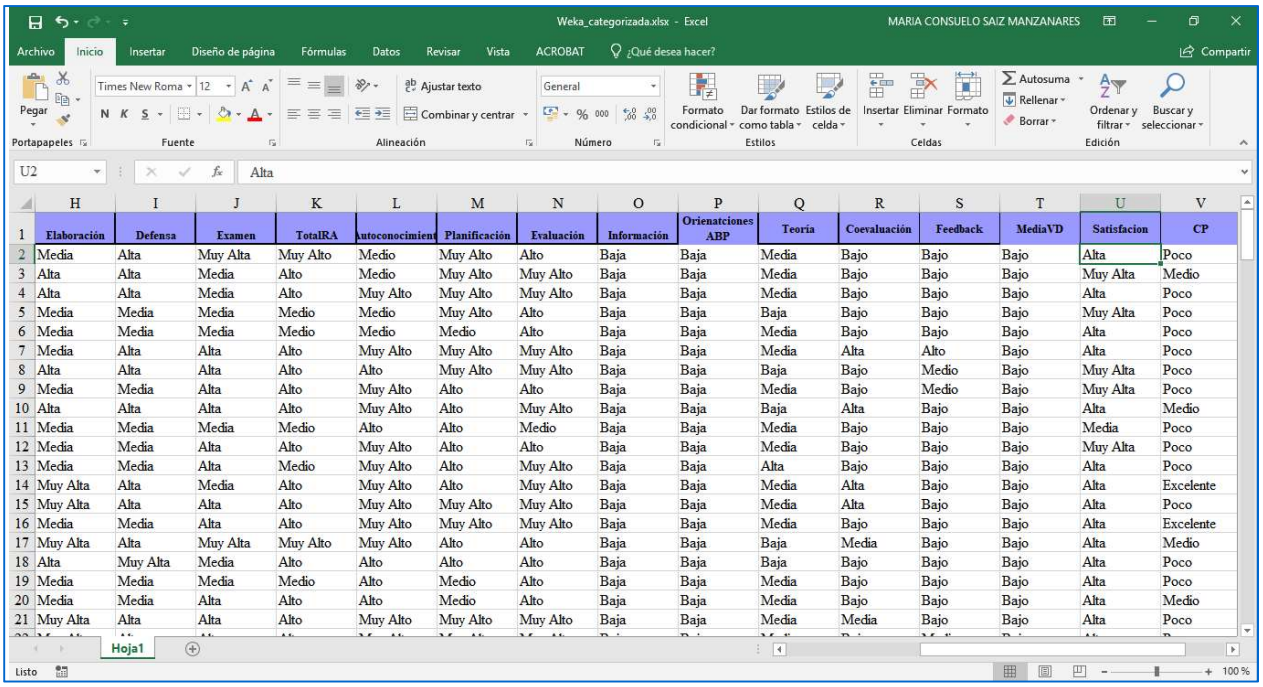
3.3.3. Ejemplos de aplicación de Técnicas de Machine Learning en el ámbito educativo.

3.3.3.1. Tratamiento de los datos con Weka en formato texto.

Weka, también permite el tratamiento y análisis de los datos en formato de calificación cualitativa en variables ordinales en formato alfanumérico. Seguidamente, se realizará un ejemplo de este tipo de análisis.

Paso 1. Preparación de la base de datos en formato texto ordinal, en este caso se aplicaron tres valores de gradación. Primero se dieron valores a las distintas variables de la base de datos con la que se estaba trabajando (ver Tabla 7, Tabla 8, Tabla 9).

En la Figura 16 se presenta un ejemplo de categorización en la base numérica original.



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Elaboración	Defensa	Examen	TotalRA	Autoconocimient	Planificación	Evaluación	Información	Orienatciones ABP	Teoría	Coevaluación	Feedback	MediaVD	Satisfacion	CP
2	Media	Alta	Muy Alta	Muy Alto	Medio	Muy Alto	Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Poco
3	Alta	Alta	Media	Alto	Medio	Muy Alto	Muy Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Muy Alta	Medio
4	Alta	Alta	Media	Alto	Muy Alto	Muy Alto	Muy Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Poco
5	Media	Media	Media	Medio	Medio	Muy Alto	Alto	Baja	Baja	Baja	Bajo	Bajo	Bajo	Muy Alta	Poco
6	Media	Media	Media	Medio	Medio	Medio	Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Poco
7	Media	Alta	Alta	Alto	Muy Alto	Muy Alto	Muy Alto	Baja	Baja	Media	Alta	Alto	Bajo	Alta	Poco
8	Alta	Alta	Alta	Alto	Alto	Muy Alto	Muy Alto	Baja	Baja	Baja	Bajo	Medio	Bajo	Muy Alta	Poco
9	Media	Media	Alta	Alto	Muy Alto	Alto	Alto	Baja	Baja	Media	Bajo	Medio	Bajo	Muy Alta	Poco
10	Alta	Alta	Alta	Alto	Muy Alto	Alto	Muy Alto	Baja	Baja	Baja	Alta	Bajo	Bajo	Alta	Medio
11	Media	Media	Media	Medio	Alto	Alto	Medio	Baja	Baja	Media	Bajo	Bajo	Bajo	Media	Poco
12	Media	Media	Alta	Alto	Muy Alto	Alto	Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Muy Alta	Poco
13	Media	Media	Alta	Medio	Muy Alto	Alto	Muy Alto	Baja	Baja	Alta	Bajo	Bajo	Bajo	Alta	Poco
14	Muy Alta	Alta	Media	Alto	Muy Alto	Alto	Muy Alto	Baja	Baja	Media	Alta	Bajo	Bajo	Alta	Excelente
15	Muy Alta	Alta	Alta	Alto	Muy Alto	Muy Alto	Muy Alto	Baja	Baja	Media	Alta	Bajo	Bajo	Alta	Poco
16	Media	Media	Alta	Alto	Muy Alto	Muy Alto	Muy Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Excelente
17	Muy Alta	Alta	Muy Alta	Muy Alto	Muy Alto	Alto	Alto	Baja	Baja	Baja	Media	Bajo	Bajo	Alta	Medio
18	Alta	Muy Alta	Media	Alto	Alto	Alto	Alto	Baja	Baja	Baja	Bajo	Bajo	Bajo	Alta	Poco
19	Media	Media	Media	Medio	Alto	Medio	Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Poco
20	Media	Media	Alta	Alto	Alto	Medio	Alto	Baja	Baja	Media	Bajo	Bajo	Bajo	Alta	Medio
21	Muy Alta	Alta	Alta	Alto	Muy Alto	Muy Alto	Muy Alto	Baja	Baja	Media	Media	Bajo	Bajo	Alta	Poco

Figura 16. Ejemplo de categorización en variables cualitativas de la base numérica.

En la Figura 17, se puede apreciar la visualización de los datos después de la categorización de los datos en nomenclatura cualitativa, en tres una escala con tres gradientes. En la Figura 18 se presenta toda la distribución de las variables estudiadas en los tres grados de categorización, y en la Figura 19 la distribución de los tres gradientes de análisis en las distintas variables.

Elaboración ABP		Defensa ABP		Examen		RT	
C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica
Media	(1,75-2,25)	Media	(1,00-1,6)	Media	(1,00-1,99)	Media	6-7,9
Alta	(2,26-2,37)	Alta	(1,7-1,8)	Alta	(2-2,5)	Alta	8-8,9
Muy Alta	(2,38-2,45)	Muy Alta	(1,9-1,95)	Muy Alta	(2,53-3)	Muy Alta	9-10

Autoconocimiento		Planificación		Evaluación	
C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica
Bajo	4-16	Media	6-8	Media	9-15
Alto	17-20	Alta	9-13	Alta	16-19
Muy Alto	21-24	Muy Alta	14-16	Muy Alta	20-24

Consultas Información C		Consultas Orientación ABP		Consultas Teoría		Co-evaluación		Feedback		MVD	
C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica	C. Cualitativa	C. numérica
Baja	(0-19)	Media	(0-10)	Media	(0-10)	Media	0-20	Bajo	0-40	Media	0-2,98
Alta	(20-40)	Alta	(11-19)	Alta	(11-30)	Alta	21-40	Alto	41-70	Alta	3-9,99
Muy Alta	(40-100)	Muy Alta	(20-32)	Muy Alta	(31-100)	Muy Alta	41-100	Muy Alto	71-249	Muy Alta	10-17

Figura 17. Transformación de las variables cuantitativas en variables cualitativas.

Paso 2. Seguidamente, se introducen los valores categorizados de forma cualitativa en la base de datos en formato .csv en Weka, se puede ver un análisis por elementos de gradación en las distintas variables recogidas en la base de datos de forma conjunta (Figura 14) y de forma individualizada en cada una de las variables (Figura 15). En este caso se tiene una base con 23 atributos y 176 sujetos.

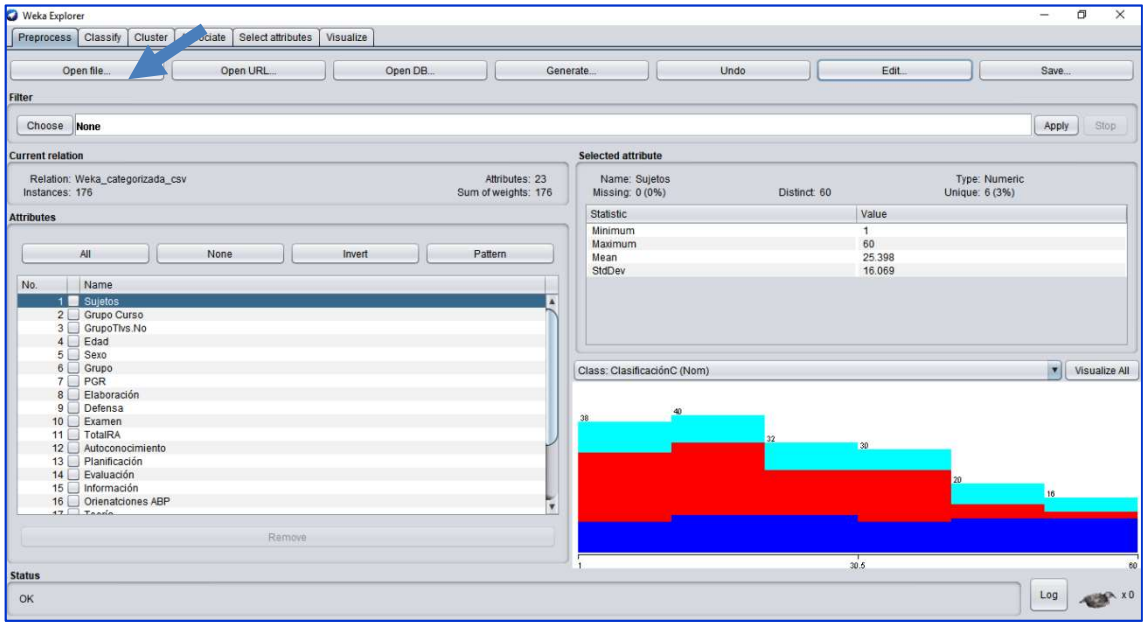


Figura 18. Inclusión en Weka de una base de datos categorizada en elementos de gradación cualitativa ver Figura 13.

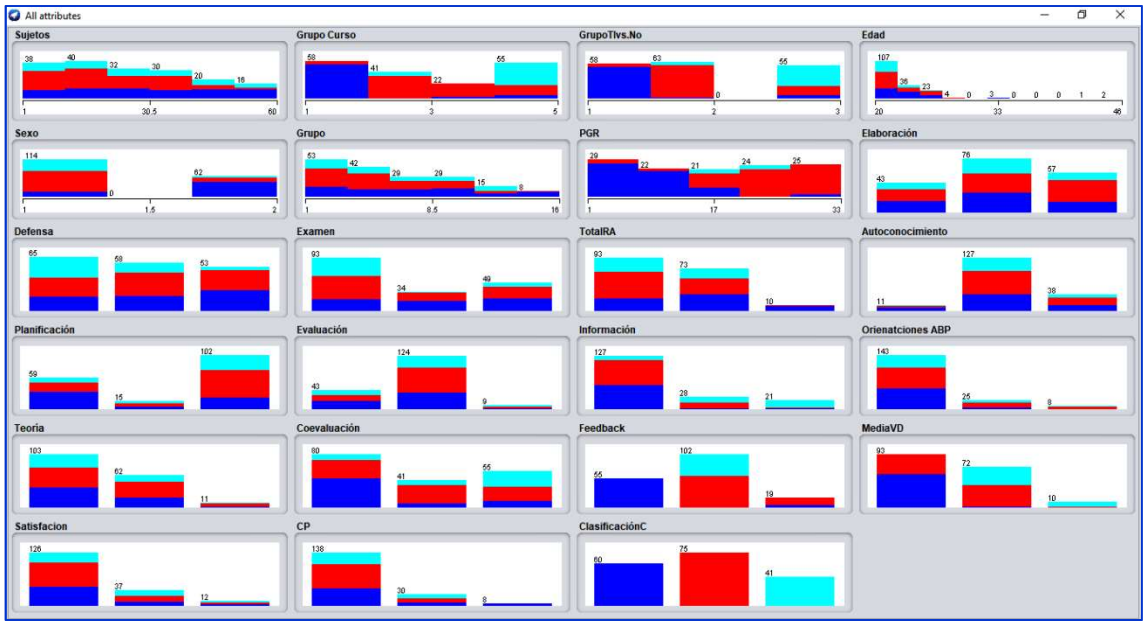


Figura 19. Análisis en Weka de cada una de las variables recogidas en la base de datos ver Figura 13.

Paso 3. Aplicación de las **técnicas de clasificación**.

Se aplicaron los algoritmos de árboles de decisión J48, genera un C4 podado o sin podar (ver Figura 20, Figura 21 y Figura 22) y LMT (logistic model trees), el cual realiza una clasificación de árboles con una regresión logística en las hojas (ver Figura 23 y Figura 24).

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

Correctly Classified Instances         145               82.3864 %
Incorrectly Classified Instances        31               17.6136 %
Kappa statistic                          0.7294
Mean absolute error                      0.149
Root mean squared error                  0.3017
Relative absolute error                 34.4663 %
Root relative squared error             64.9066 %
Total Number of Instances              176

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,917    0,017    0,965      0,917    0,940      0,911    0,975     0,956     Excelente
               0,787    0,149    0,797      0,787    0,792      0,639    0,858     0,725     Suficiente
               0,756    0,104    0,689      0,756    0,721      0,632    0,878     0,614     Bueno
Weighted Avg.  0,824    0,093    0,829      0,824    0,826      0,730    0,903     0,778

=== Confusion Matrix ===

  a  b  c   <-- classified as
 55  5  0 |  a = Excelente
  2 59 14 |  b = Suficiente
  0 10 31 |  c = Bueno
```

Figura 20. Arquitectura del árbol de decisión con la técnica J48 en Weka.

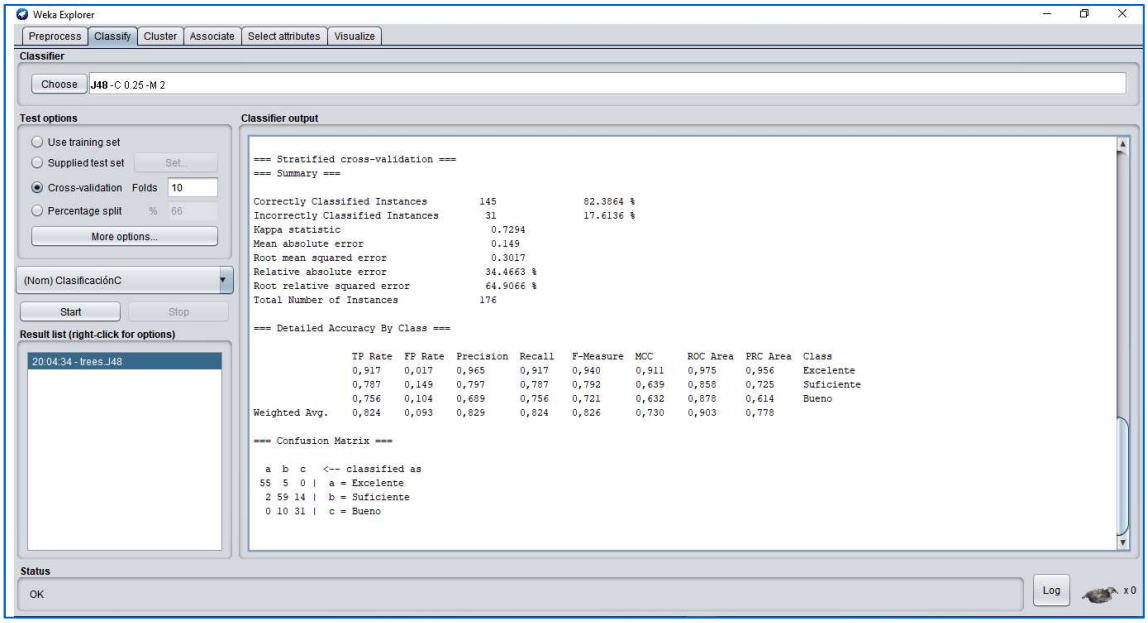


Figura 21. Resultados aplicando la técnica de árbol de decisión J48 en Weka.

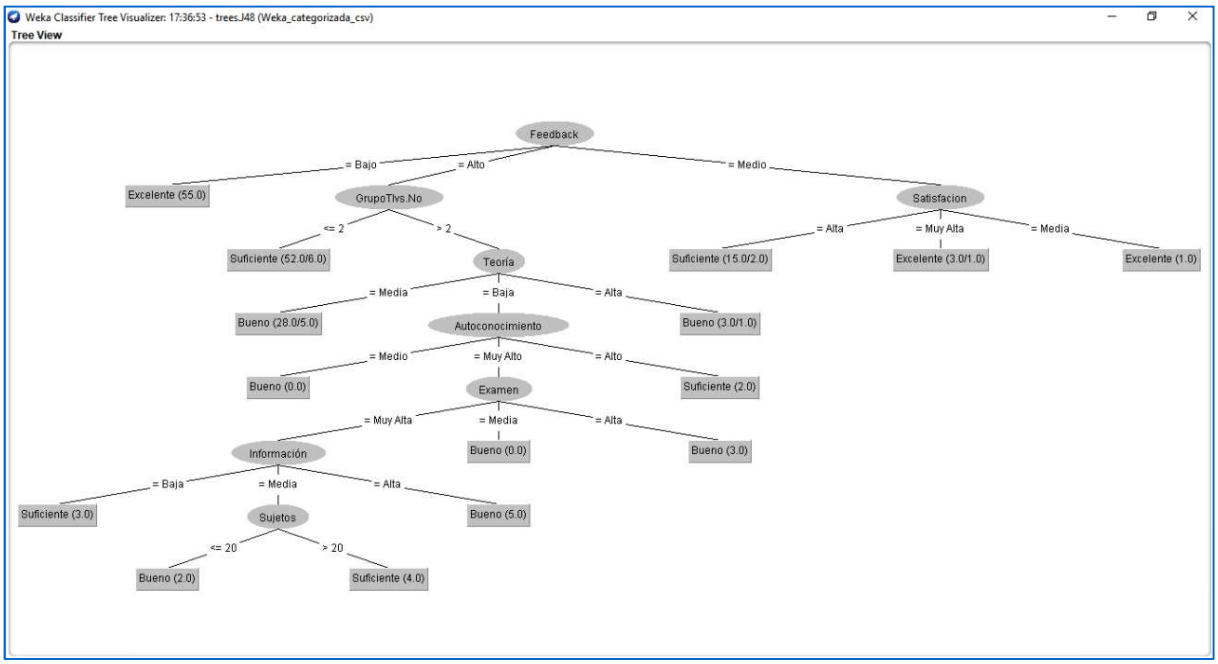


Figura 22. Visualización del Árbol de decisión la técnica de árbol de decisión J48 en Weka.

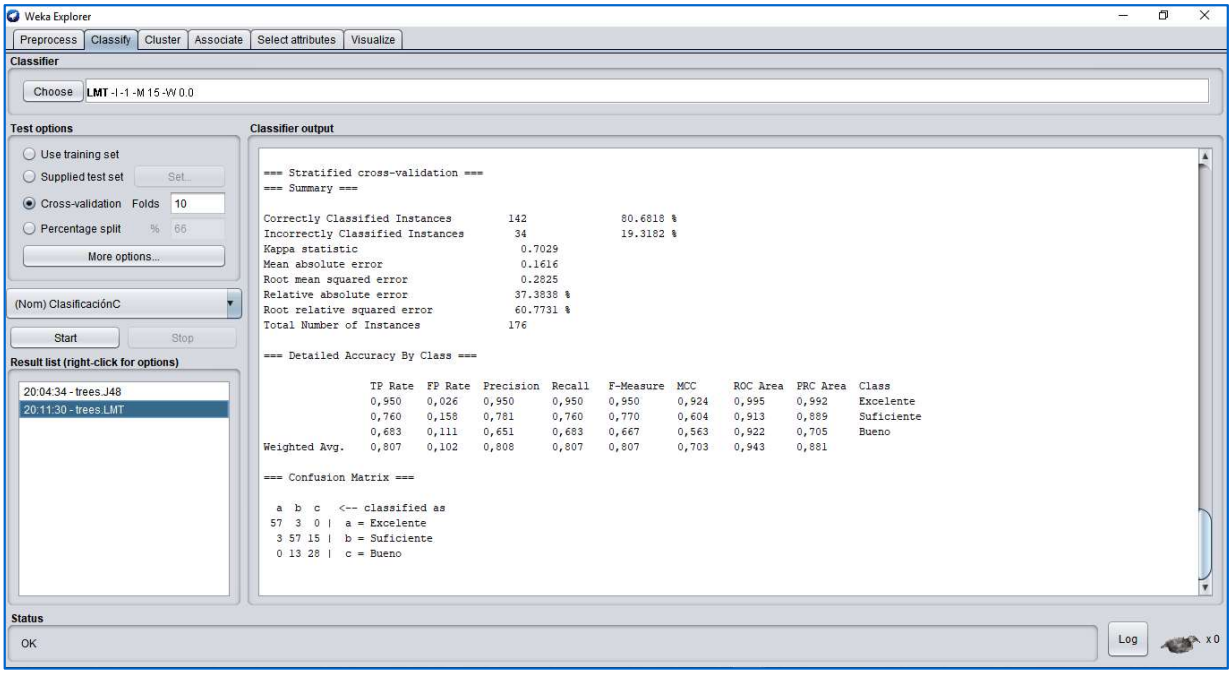


Figura 23. Resultados aplicando la técnica de árbol de decisión LMT en Weka.

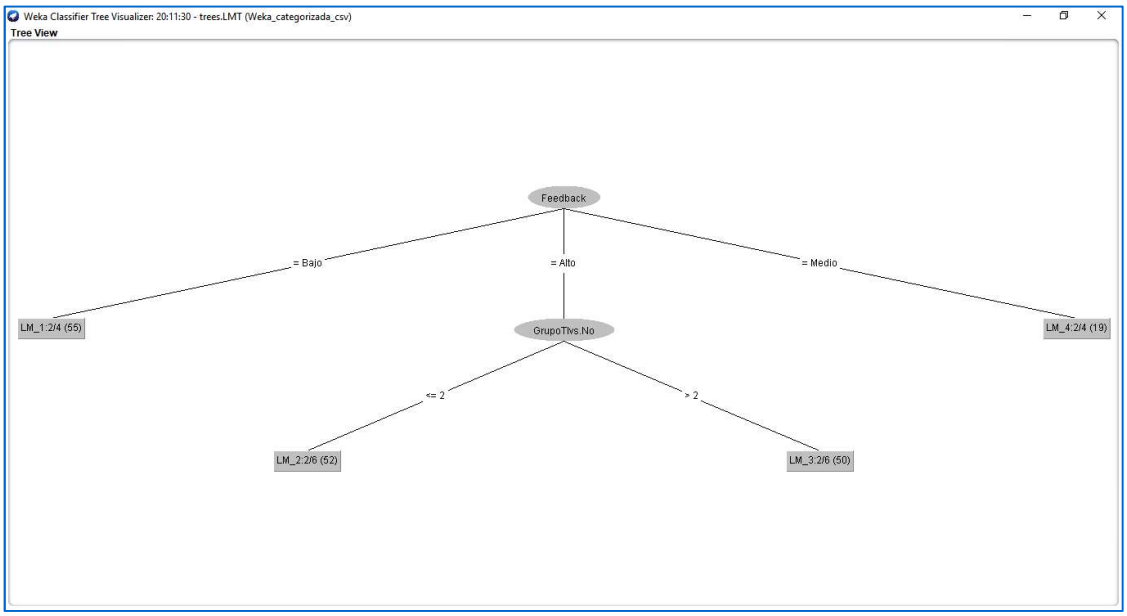
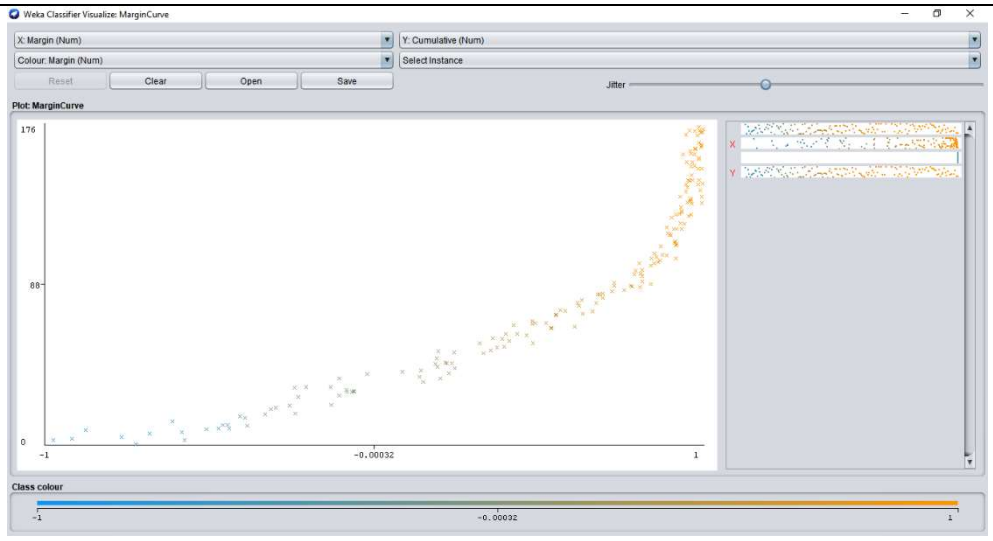


Figura 24. Visualización del Árbol de decisión la técnica de árbol de decisión LMT en Weka.

Una Visualización del margen de la curva con ambos algoritmos se observar ver en la Figura 25.

LMT	

J48

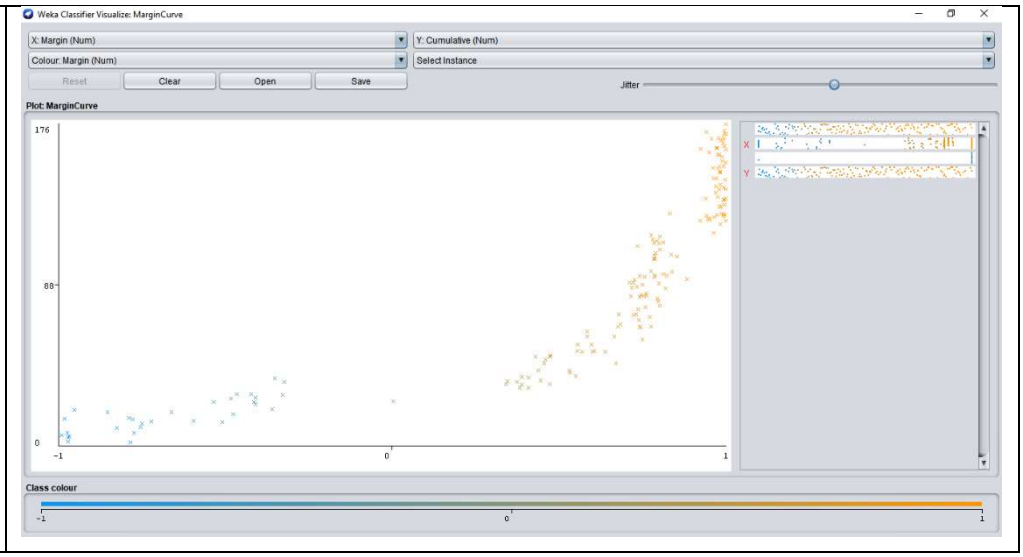


Figura 25. Visualización del margen de la curva en LMT y J48 en Weka.

En la Tabla 1, se presenta una comparativa del ajuste de las distintas técnicas de clasificación realizadas en Weka, en este caso la prueba que mejor ajusta es RBFClassifier. En la Tabla 2, se puede comprobar la asignación de sujetos a los clústeres denominados: Excelente, Suficiente y Bueno desde la aplicación de distintas técnicas de clasificación.

Tabla 1

Resultados de distintas pruebas de clasificación en Weka.

	Multilayer Perceptron	RBFClassifier	Simple Logistic	Support Vector Machine (SMOreg)	Trees. J48	Trees. LMT
Correctly Classified Instances	138	148	142	144	145	142
Incorrectly Classified Instances	38	28	34	32	31	34
Kappa statistic	.667	0.755*	0.703	0.720	.729	.703
Mean absolute error	0.159	0.249	.161	.263	.149	.161
Root mean squared error	0.353	0.310	.282	.338	.3017	.282
Relative absolute error	36.717%	57.517%	37.384%	60.760%	34.466%	37.384%
Root relative squared error	75.999%	66.583%	60.773%	72.794%	64.906%	60.773%
Total Number of Instances	176	176	176	176	176	176

Tabla 2

Resultados de distintas pruebas de clustering en Weka.

Grupos origen			Categorías	Método de clustering																
				Multilayer Perceptron			RBFClassifier			Simple Logistic			Support Vector Machine (SMOreg)			Trees. J48			Trees. LMT		
			De la clasificación																	
C_1	C_2	C_3		a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
59	63	55	Excelente	57	4	0	57	2	1	57	3	0	58	2	0	55	5	0	57	3	0
			Suficiente	4	56	15	1	62	12	3	57	15	1	59	15	2	59	14	3	57	15
			Bueno	0	16	25	0	12	29	0	13	28	0	14	27	0	10	31	0	13	28

Paso 4. Aplicación de técnicas de clustering, se utilizaron técnicas de EM [Simple EM (Expectation-Maximization) class] ver Figura 26 y Figura 27 y de *k*mean ver Figura 28 y Figura 29.

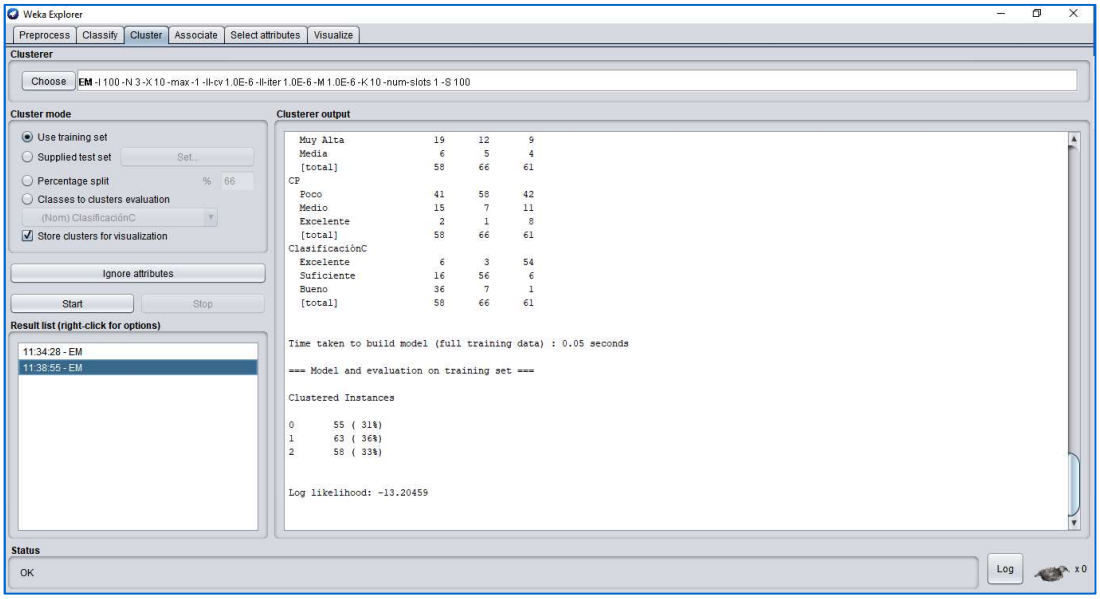


Figura 26. Resultados aplicando la técnica de clustering con la técnica de EM en Weka.

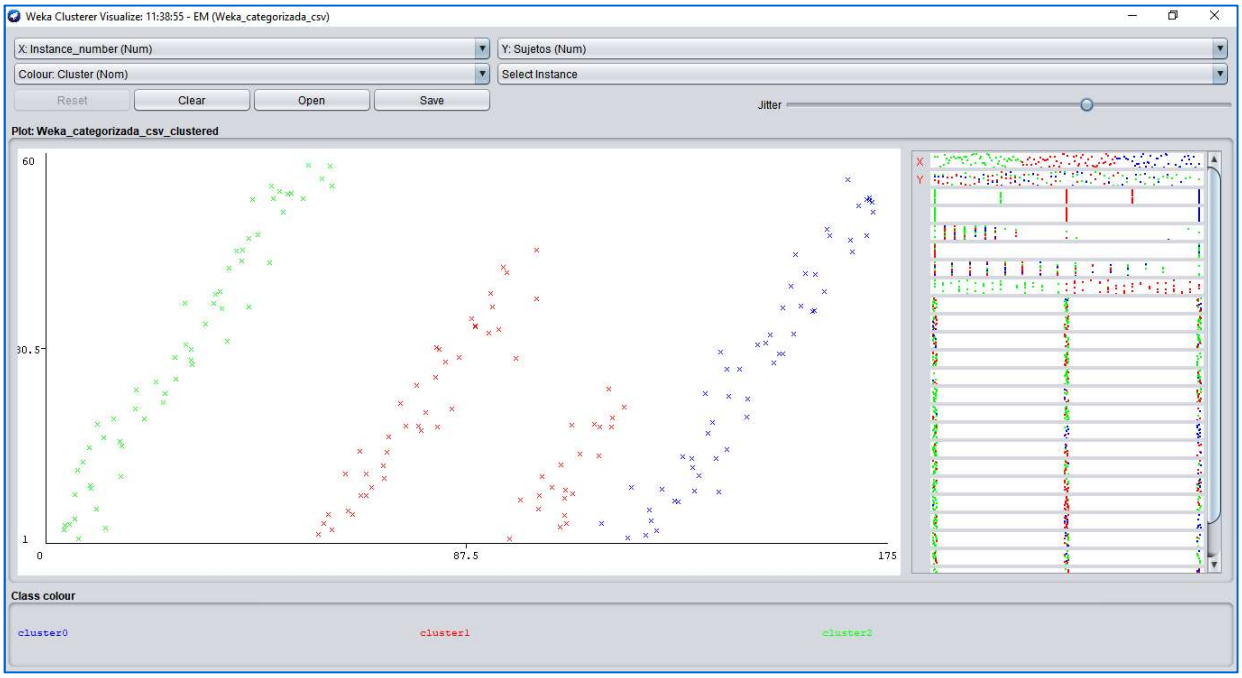


Figura 27. Visualización de la asignación de clústeres con la técnica de EM en Weka.

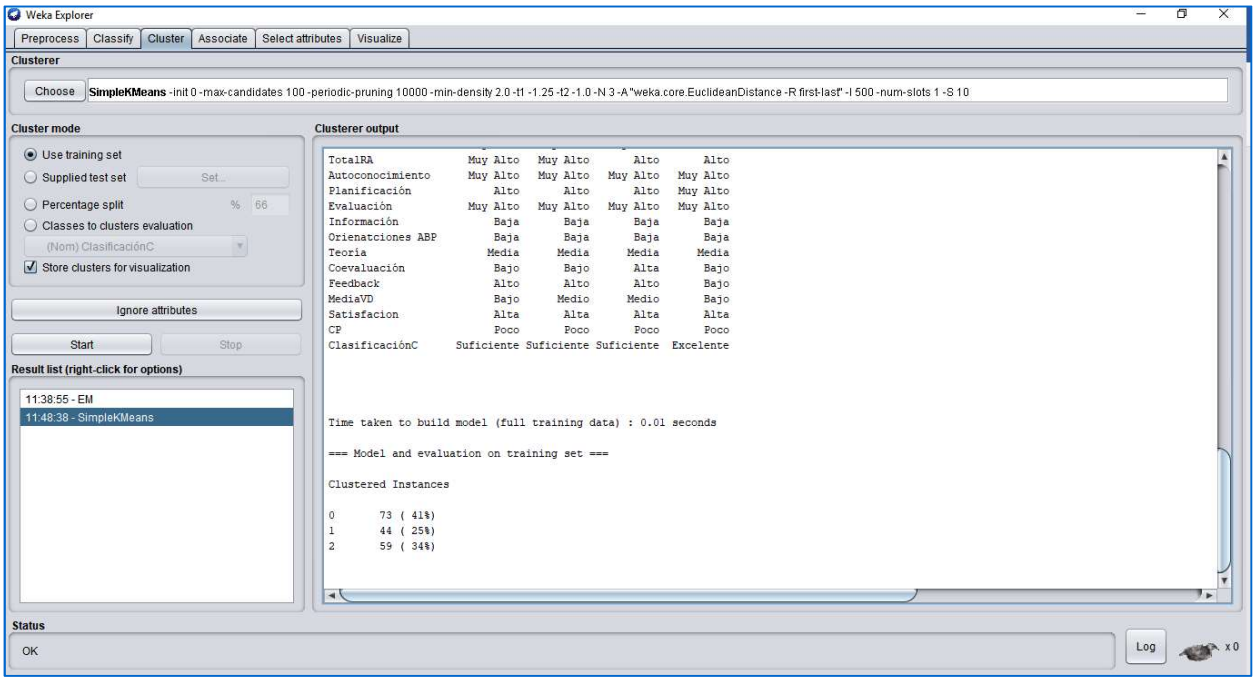


Figura 28. Resultados aplicando la técnica de clustering con la técnica de *k*mean en Weka.

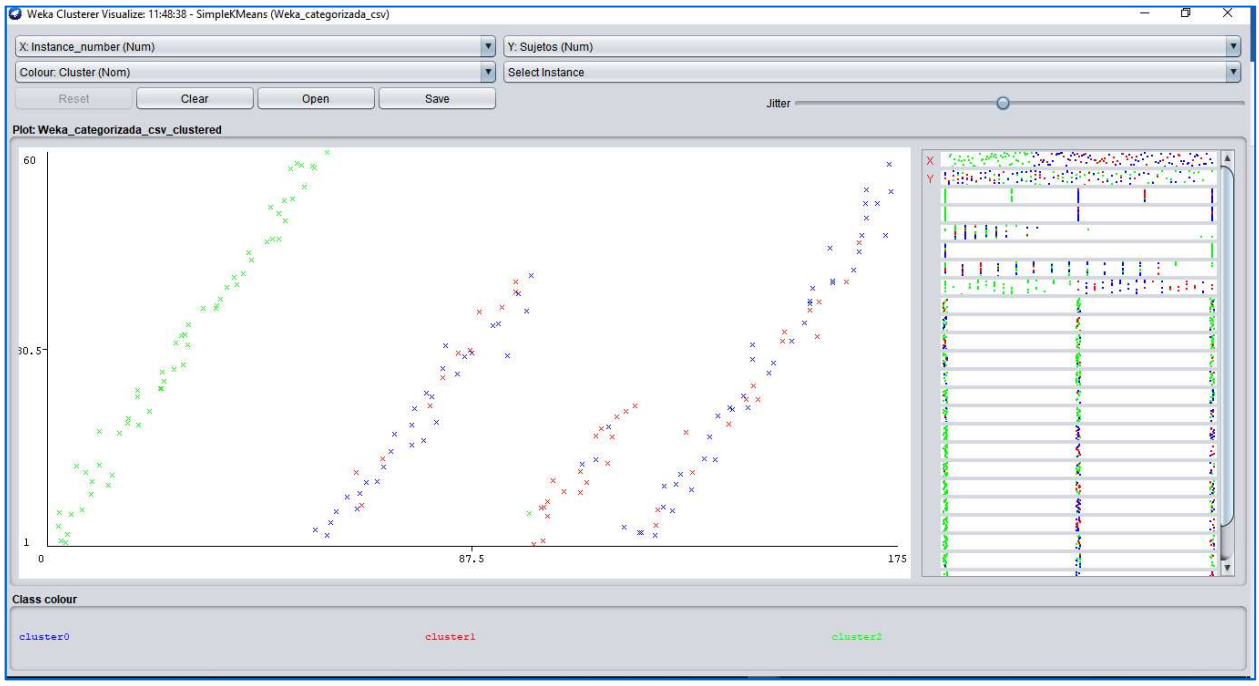


Figura 29. Visualización de la asignación de clústeres con la técnica de *k*mean en Weka.

La comparativa de los resultados con ambas técnicas de clustering se puede consultar en la Tabla 3.

Tabla 3

Resultados de distintas pruebas de clustering en Weka.

Grupos	Método de clustering			
	EM	%	Kmean	%
n = 59	$C_0 = 55$	31	$C_0 = 73$	41
n = 63	$C_1 = 63$	36	$C_1 = 44$	25
n = 55	$C_2 = 58$	33	$C_2 = 59$	34

Paso 5. Aplicación de técnicas de selección de atributos con Weka.

Weka permite hallar la selección de atributos dentro de una base de datos, con la aplicación de distintas técnicas. La elección de la técnica dependerá del tipo de variables a modo de ejemplo se aplicó la técnica CfsSubsetEval (Evalúa el valor de un subconjunto de atributos al considerar la capacidad predictiva individual de cada característica junto con el grado de redundancia entre ellas). En este caso se eligieron tres variables tipo de Blend, co-evaluación y *Feedback* (ver Figura 30 y Figura 31). En la Figura 32 se puede ver la relación entre las variables acceso al *Feedback* del docente y tipo de Blend y en la Figura 33 la relación entre las variables participación en acciones de coevaluación y el tipo de Blend.

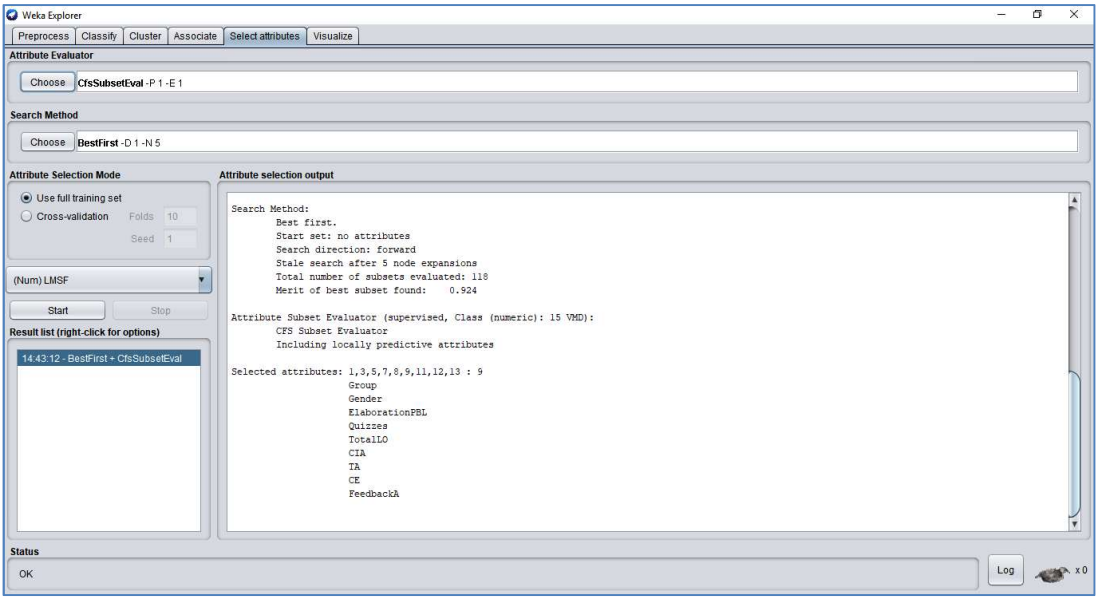


Figura 30. Técnica de selección de atributos CfsSubsetEval en Weka.

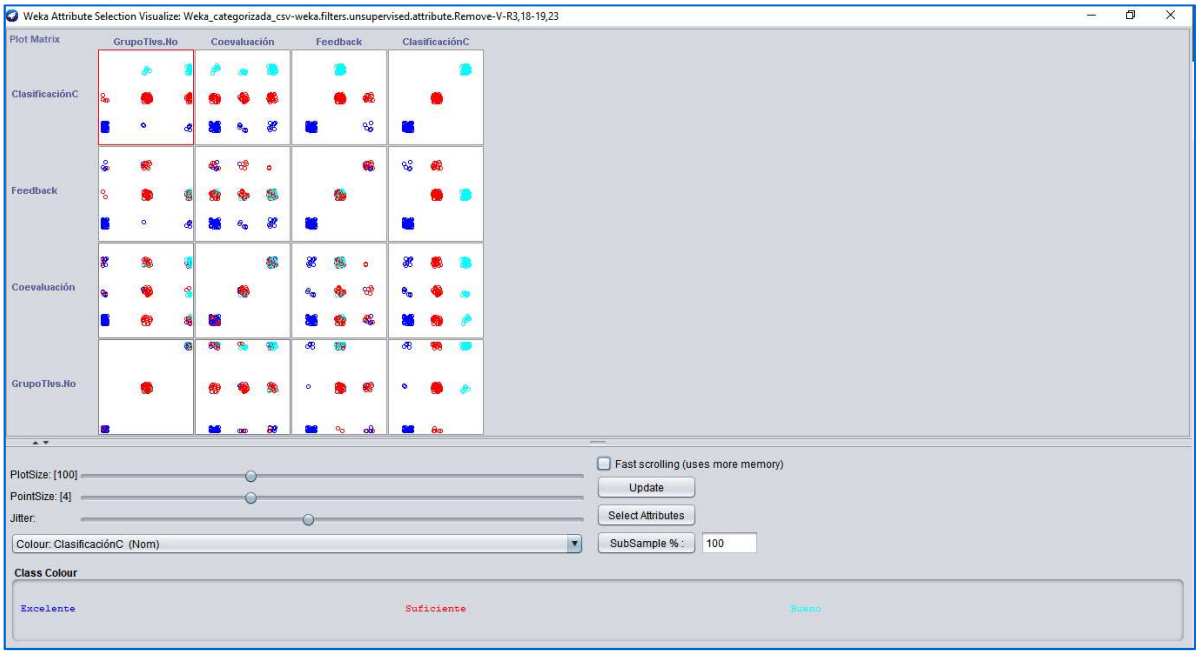


Figura 31. Visualización de las relaciones entre las variables seleccionadas con la técnica CfsSubsetEval en Weka.

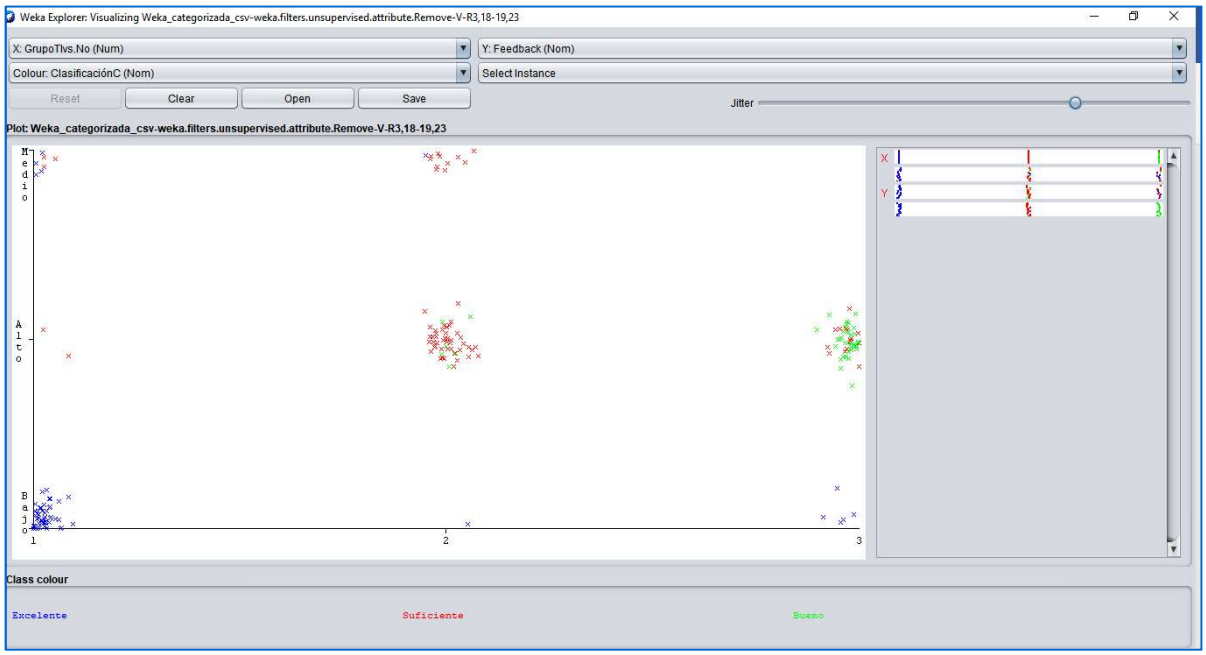


Figura 32. Visualización de las relaciones entre las variables acceso al feedback del docente y el tipo de Blend con la técnica CfsSubsetEval.

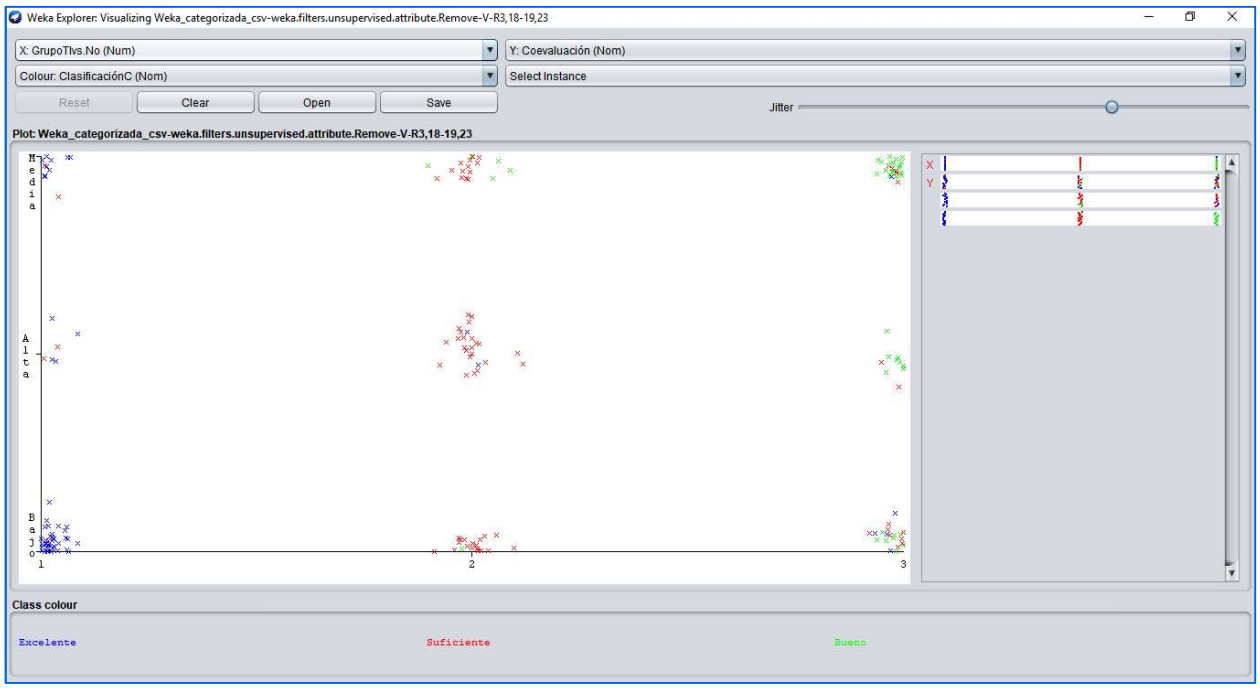


Figura 33. Visualización de las relaciones entre las variables participación en acciones de coevaluación y el tipo de Blend con la técnica CfsSubsetEval.

3.3.3.2. Tratamiento de los datos con SPSS en formato texto

En la Figura 34 se puede comprobar la inserción de los datos en SPSS utilizando una base de datos cualitativos.

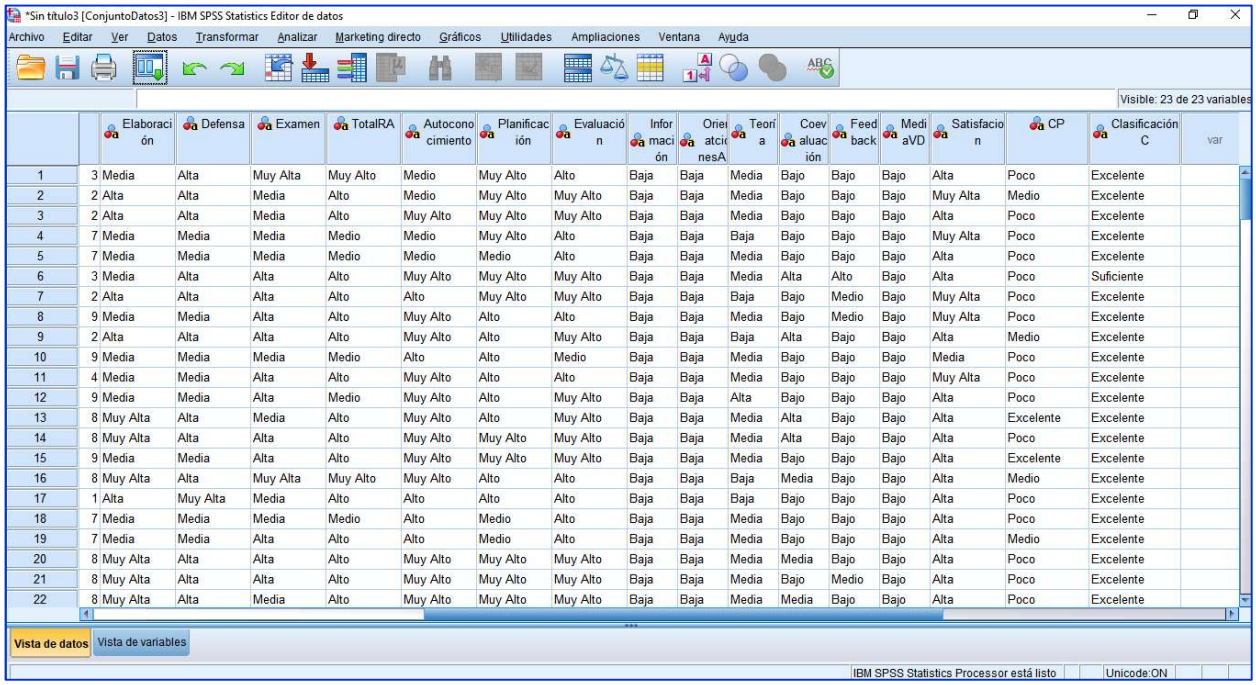


Figura 34. Inserción de una base de datos categorizada de forma cualitativa en SPSS.

Se pueden analizar las frecuencias de cada variable (ver Figura 35), estos datos nos darán un indicador del estado de la cuestión en cada una de las variables.

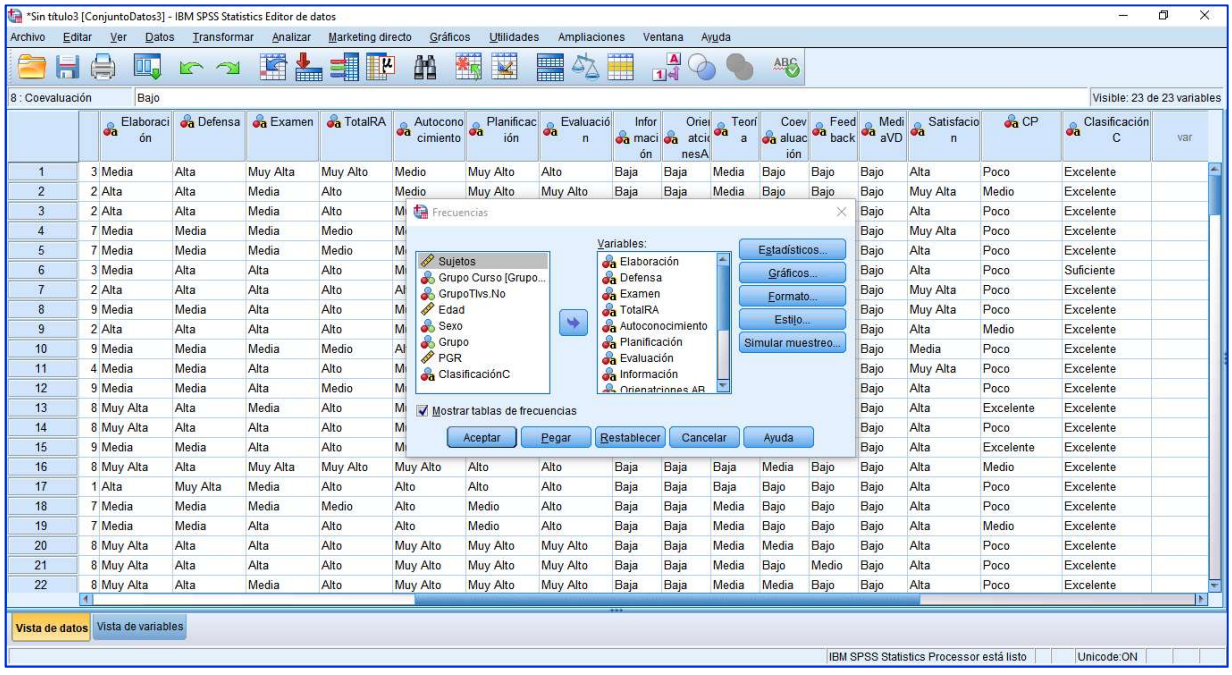


Figura 35. Análisis de frecuencias en SPSS v.24 de las variables categorizadas.

A modo de ejemplo se presenta la frecuencia y los porcentajes en las variables. resultados de aprendizaje (ver Figura 36).

Resultados de aprendizaje en la prueba de elaboración

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Alta	76	43.2	43.2	43.2
	Media	43	24.4	24.4	67.6
	Muy Alta	57	32.4	32.4	100.0
	Total	176	100.0	100.0	

Resultados de aprendizaje en la prueba de Defensa

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Alta	65	36.9	36.9	36.9
	Media	58	33.0	33.0	69.9
	Muy Alta	53	30.1	30.1	100.0
	Total	176	100.0	100.0	

Resultados de aprendizaje en la prueba Examen

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Alta	49	27.8	27.8	27.8
	Media	34	19.3	19.3	47.2
	Muy Alta	93	52.8	52.8	100.0
	Total	176	100.0	100.0	

Resultados de aprendizaje Totales

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Alto	73	41.5	41.5	41.5
	Medio	10	5.7	5.7	47.2
	Muy Alto	93	52.8	52.8	100.0
	Total	176	100.0	100.0	

Figura 36. Frecuencias y porcentajes en las variables: resultados de aprendizaje.

También se pueden hallar tablas cruzadas en las variables que se elijan. a modo de ejemplo en la Figura 37 se presentan tablas cruzadas de la variable tipo de Blend empleado y los resultados de aprendizaje en las distintas pruebas.

Tabla cruzada Tipo de Blender y prueba defensa

Recuento

		Defensa			
		Alta	Media	Muy Alta	Total
Tipo de Blender	1	14	16	28	58
	2	14	29	20	63
	3	37	13	5	55
Total		65	58	53	176

Tabla cruzada Tipo de Blender y Elaboración

Recuento

		Elaboración			
		Alta	Media	Muy Alta	Total
Tipo de Blender	1	29	17	12	58
	2	20	18	25	63
	3	27	8	20	55
Total		76	43	57	176

Tabla cruzada Tipo de Blender y Examen

Recuento

		Examen			
		Alta	Media	Muy Alta	Total
Tipo de Blender	1	25	19	14	58
	2	17	14	32	63
	3	7	1	47	55
Total		49	34	93	176

Tabla cruzada Tipo de Blender y Resultados de aprendizaje Totales

Recuento

		Resultados de aprendizaje Totales			
		Alto	Medio	Muy Alto	Total
Tipo de Blender	1	31	9	18	58
	2	21	1	41	63
	3	21	0	34	55
Total		73	10	93	176

Figura 37. Tablas cruzadas en SPSS.

Resumen

Como se vio en el Tema 5, la preparación de los datos es un tema esencial en el ámbito de la investigación tanto cualitativa como cuantitativa. En este tema se ha aplicado la preparación de los datos desde una perspectiva cualitativa para su posterior tratamiento tratamiento con softwares como Weka y SPSS. En el primero se han utilizado técnicas de aprendizaje automático supervisado y no supervisado, ya que dicho entorno contiene algoritmos que permiten el trabajo con datos cualitativos en variables ordinales. También, se han visto ejemplos del procesamiento con datos cualitativos con SPSS, en este caso permite la aplicación de técnicas de frecuencia, porcentajes y tablas cruzadas que dan mucha información en diseños descriptivos.

Glosario

Atributo: Un atributo se entiende como la categorización que hacen las técnicas de machine learning desde la fijación de valores y la predefinición de las características o de los atributos (Witten y Eibe, 2005).

Boosting: Es un meta-algoritmo de conjunto de aprendizaje automático que se utiliza para reducir principalmente el sesgo, y también la varianza en el aprendizaje supervisado. Boosting está basado en el cuestionamiento planteado por Kearns y Valiant (1988, 1989).

Data squashing: Es un tipo de compresión con pérdida que intenta preservar la información estadística. Para ser eficiente, debe mejorar la estrategia común de tomar una muestra aleatoria del gran conjunto de datos.

Instancia: Una instancia son las cosas que hay que clasificar, asociar o clustorizar desde los datos de la entrada (Witten y Eibe, 2005).

Sampling: Es una técnica que toma una porción de la muestra objeto de estudio y aplicar sobre ella distintas técnicas de análisis.

Tupla: Lista ordenada de elementos.

Bibliografía

Bibliografía básica

Bogarín, A., Romero, C., & Cerezo, R. (2016). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. *Revista de Educación Mediática y TIC*, *5*(1), 73-92

Bogarín, A., Cerezo, R., & Romero, C. (2017a). A survey on educational process mining. *WIREs Data Mining Knowl Discov*, *230*. doi: 10.1002/widm.1230

Bogarín, A., Cerezo, R., & Romero, C. (2017b). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema*, *30*(3). doi: 10.7334/psicothema2018.116

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M.P., Núñez, J.C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42-54. doi: 10.1016/j.compedu.2016.02.006

Cerezo, R., Esteban, M., Sánchez-Santillán., & Núñez, J.C. (2017). Procrastinating Behavior in Computer-Based Learning Environments to Predict Perfromance: A Case Study in Moodle. *Frontiers in Psychology*, *8*, 1-11.doi: 10.3389/fpsyg.2017.01403(Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2016)

Dutt, A., Ismail, M.A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*, 15991-16005. doi: 10.1109/ACCESS.2017.2654247

González-Pienda, J.A., Bernardo, A., Núñez, J.C., & Rodríguez, C. (2017). *Factors Affecting Academic Performance*. New York: Nova.

Rodríguez Villalobos, A. (2010). *Grafos: Software para la construcción, edición y análisis de grafos*. España: Bubok Publishing.

Romero, C., Cerezo, R., Bogarín, A., Sánchez-Santillán, M. (2016). *Educational Process Mining: A tutorial and case study using Moodle data sets*. En S. Elatia, D. Ipperciel., & O.R. Zaïane (Eds.), Data Mining and Learning Analytics (pp. 3-28). New Jersey: Wiley Online Library. doi: 10.1002/9781118998205.ch1

Romero, C., Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews*, *40*(6). doi: 10.1109/TSMCC.2010.2053532

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2016). Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106. doi: 10.3102/1076998616666808

Witten, I.H., & Eibe, F. (2005). Data Mining: Practical Machine Learning Tools and Techniques

Recursos

Web

Data Mining and Learning Analytics	Enlace
Weka	Enlace
Problemas en Weka	Enlace
Árboles de decisión en Weka	Enlace
Curso de minería de datos	Enlace
Grafos	Enlace