Guía Metodológica para la Selección y Evaluación de Técnicas de Minería de Datos Enfocadas en el Establecimiento de Asociaciones

Martín Elías Rico Herrera

Directora: Claudia Jiménez Ramírez, Ph.D.
Profesora Asociada, Escuela de Ciencias de la Computación y de la Decisión

Trabajo presentado como requisito para optar al título de Magíster en Ingeniería – Ingeniería de Sistemas

Escuela de Ciencias de la Computación y la Decisión
Facultad de Minas
Universidad Nacional de Colombia
Sede Medellín
2012

Resumen:

El análisis de asociaciones es un estudio tanto de tipo descriptivo como predictivo dentro del proceso de *Descubrimiento de Conocimiento en Bases de Datos*, el cual busca encontrar y establecer relaciones significativas dentro un conjunto de variables o de objetos. Para esto, existen diferentes técnicas estadísticas e informáticas las cuales contienen diferencias significativas, que dependiendo del estudio a realizar, son más convenientes que otras.

Este trabajo está orientado a estudiar y definir un conjunto de etapas y actividades enmarcadas en una guía de desarrollo, la cual busca ser un marco formal para mineros de datos cuando estos necesiten seleccionar la técnica de minería más apropiada para abordar un problema de asociaciones.

Palabras claves: análisis de asociaciones, guía, minería de datos.

Abstract:

The association analysis is both a descriptive study as predictive in the process of Knowledge Discovery in Databases, which intends to find and establish meaningful relationships within a set of variables or objects. For doing this task, there are different statistical and computational techniques which contain significant differences, which depending on the study to be performed are more suitable than others.

This work aims to study and define a set of steps and activities as methodological development guide, which is intended as a formal framework for data miners when they need to select the most appropriate mining technique to tackle a problem of associations.

Keywords: association analysis, guide, data mining.

Índice General

1 Introducción 5

1.1 Definición del problema. 6

1.2 Estado del arte. 7

2 Caracterización de las técnicas. 12

2.1 Regresión Lineal. 12

2.2 Reglas de Asociación. 14

2.3 Cópulas. 18

3 Guía metodológica. 22

3.1 Conceptos. 22

3.2 Definición de la guía. 23

3.2.1 Análisis. 24

3.2.3 Aplicación. 29

3.2.3 Evaluación. 31

4 Aplicación de la guía. 36

Caso de estudio 1. 36

4.1 Etapa de análisis. 37

4.2 Etapa de aplicación. 44

4.3 Etapa de evaluación. 62

5 Conclusiones. 66

Conclusiones. 66

6 Referencias bibliográficas. 67

Índice de tablas

Tabla 1. Relación de artículos y aportes.

Tabla 2. Cópulas Arquimedianas y sus generados.

Tabla 3. Variables incluidas en el diagnóstico de cáncer de mama.

Tabla 4. Estadísticas básicas de las variables de estudio de cáncer de mama.

Tabla 5. Tipos de distribución de las variables de estudio de cáncer de mama.

Tabla 6. Resultados de discretizar las variables de estudio de cáncer de mama.

Tabla 7. Valores del estadístico Wald para las variables de estudio de cáncer de mama.

Tabla 8: Frecuencias observadas y esperadas para los grupos definidos.

1. Introducción

El proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD por sus siglas en inglés), definido como el "*proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y entendibles en los datos*" (Fayyad et al., 1996), se ha convertido en una herramienta de gran acogida y aceptación en diferentes campos de aplicación del mundo real, lo que ha generado un creciente interés en el estudio de las diferentes temáticas involucradas en este proceso.

Dentro de las áreas de estudio se encuentran la estadística y el aprendizaje máquina, las cuales ofrecen diferentes métodos que representan diversos fenómenos naturales y físicos a través de modelos como la Asociación, Clasificación, Agrupamiento, Predicción, Regresión, Análisis de series de tiempo, Visualización (Ngai et al., 2009) a partir de los cuales se pueden realizar inferencias y generar conocimiento. Para obtener cualquiera de estos modelos existe la posibilidad de emplear varios métodos, por lo cual es necesario conocer los diferentes criterios que existen para determinar el más apropiado a emplear sobre un problema en particular y la forma de proceder para utilizarlos adecuadamente.

El proceso de selección de la técnica o método de minería de datos que permita establecer asociaciones con el mejor desempeño posible, no es tarea fácil aún para usuarios expertos, de manera que el grado de dificultad al que se enfrentan analistas no expertos o personas que desean realizar un análisis pero que no están familiarizadas con el área, es altísimo, como el caso de administradores o gerentes. Igualmente, sucede en el entorno académico, investigadores que desean estudiar o realizar una análisis para conocer la forma como se relacionan un conjunto de variables, proceden en muchas ocasiones a trabajar con la más común, dado que no conocen la existencia de otros métodos o no saben cómo proceder a determinar cuál es más apropiado para aplicar a su caso de estudio. Así, es importante para el crecimiento y afianzamiento del proceso de KDD, al igual que para el desarrollo de actividades de investigación, avanzar en mejorar y depurar el proceso de selección de técnicas de minería de datos en el marco del establecimiento de asociaciones.

1.1 Definición del problema

¿Cómo determinar las técnicas de minería de datos que pueden ser aplicadas a un estudio para establecer asociaciones y evaluar sus desempeños?

Es conocido que varios métodos de minería de datos pueden emplearse para realizar un análisis de asociación, dando origen a las siguientes preguntas, ¿cómo escoger el método más apropiado para aplicar a un estudio de asociaciones?, ¿cuáles son los criterios para la selección?, ¿cuándo es más conveniente aplicar los métodos y luego evaluar?. En (Mitra et al., 2002) se establece que no existe universalmente "*el mejor método de minería de datos*", escoger un método o una combinación de métodos es enteramente dependiente de la aplicación en particular y de la interacción humana, que decide sobre lo idóneo de una metodología; de manera similar (Carrier & Povel, 2003, citado en Ngai et al.,2009) sugieren que para escoger una técnica de minería de datos apropiada se deben estudiar las características de los datos y los requerimientos del problema; (Scott & Wilkins, 1999) indican que no existe una metodología base que ayude al desarrollo de este problema.

En la actualidad una variedad de estudios y análisis se han realizado alrededor de los diferentes métodos existentes para realizar tareas de minería de datos, algunos enfocados a comparar diferentes técnicas para un estudio en particular, otros analizan variaciones de los algoritmos de una misma técnica, todas con el objetivo de establecer cual tiene mejor desempeño, es más precisa o arroja mejores resultados (I-Cheng & Che-hui. 2009.) (kurt et al., 2008) (Serna, S. 2009).

Los análisis de estos estudios son realizados teniendo en cuenta las diferentes tareas de minería de datos como son las Asociaciones, el Agrupamiento, las Clasificaciones, Predicciones y el Análisis secuencial. En la gran mayoría de estos se observa que la determinación de la técnica más apropiada se basa en una medida evaluada posteriormente a la aplicación de las técnicas, como la "*tasa de mala calificación*", "*curva ROC*", entre otras. Estos estudios dan poca o ninguna relevancia a los supuestos o propiedades bajo los cuales pueden ser aplicadas estas técnicas y enfatizan más en la aplicación o ejecución, sin detenerse en la caracterización detallada.

Cuando se realiza una evaluación posterior a la aplicación de las técnicas se observa que existen diferentes métodos y medidas de evaluación, para una posterior selección, originando incertidumbre al momento de determinar con cual es más apropiado evaluar. Aunque afrontar esta situación no es difícil, no se conoce de un acercamiento orientado hacia analistas no expertos en la tarea de evaluación de los resultados obtenidos.

Adicionalmente, es importante resaltar la falta de un proceso estructurado a partir del cual se pueda efectuar la selección y evaluación de técnicas, un marco de referencia que defina un conjunto de actividades a seguir, que indique aspectos y características a

tener en cuenta para ejecutar la tarea de forma coherente, una guía que sirva como punto de partida para orientar a mineros de datos a desarrollar dicha actividad con mayor claridad.

En general no existe un acercamiento o una propuesta de diseño orientada a cualquier tipo de usuario para abordar la tarea de seleccionar una técnica de minería de datos, específicamente para el descubrimiento de asociaciones entre variables, por lo cual es necesario trabajar para avanzar en la solución de las deficiencias antes mencionadas.

1.2 Estado del arte

Como se mencionó anteriormente, uno de los principales problemas para que los métodos de aprendizaje máquina no se hayan consolidado en la industria es el grado de complejidad que representa para los gerentes y administradores abordar ciertos aspectos, como lo es el proceso de selección del método a aplicar en determinados estudios.

Como un acercamiento para facilitar el ejercicio a los analistas de datos de seleccionar adecuadamente las técnicas de minería a aplicar a un estudio en particular, (Spott & Nauck, 2006) proponen el desarrollo de una aplicación informática que permita establecer, lo que en el artículo llaman, preferencias, que no es más que requisitos acerca del modelo que se desea obtener, para luego realizar una "mapeo" con las características de los distintos métodos que maneja la aplicación y que como resultado se presenten un conjunto de técnicas que cumplen con las condiciones establecidas y de las cuales es posible seleccionar una para realizar el análisis. Además, ese documento resalta la pertinencia de analizar ciertas características como la adaptabilidad y la facilidad de interpretación sobre los modelos generados a partir de las diferentes técnicas, como parte del proceso de evaluación de las mismas.

El estudio recién mencionado tiene como fortaleza que presenta implícitamente una guía para abordar el proceso de selección de una técnica de minería, representa el mayor acercamiento o semejanza a la propuesta de investigación que aquí se plantea. Sin embargo, tiene como desventaja que no profundiza en los aspectos o pasos dentro del proceso de selección de la técnica. Apenas menciona algunos métodos de minería de datos y criterios de evaluación de modelos como la precisión y simplicidad. Por otro lado, al estar orientado al desarrollo de una aplicación el proceso se convierte en una caja negra para el analista de datos y siendo este un proceso con alto grado de variabilidad y dependencia de diferentes factores como el tamaño de los datos, la naturaleza de los mismos, el análisis de supuestos, entre otros, generaría cierta inconformidad o desconfianza en cuanto a la posibilidad de que se tomen decisiones equivocadas.

El estudio realizado por (Enke & Thawornwong, 2005) aborda también la problemática de selección de las técnicas de minería de datos. Ellos indican que muchos estudios

realizados sobre mercados financieros fallan a causa de predecir con técnicas alternativas a las redes neuronales. Este estudio sugiere que algunas técnicas trabajan bajo el supuesto de *"relaciones lineales"* lo que descarta su uso, dado la comprobada inexistencia de linealidad en los mercados financieros actuales. Luego analizan características de las redes neuronales y otras técnicas que permiten trabajar bajo el supuesto de no-linealidad, permitiendo establecer que la primera técnica es más conveniente que las otras, dado que no necesita cumplir con supuestos para el proceso de modelado, además de las oportunidades que ofrece gracias a los diferentes algoritmos de aprendizaje, procesos de evaluación y tipos de arquitectura. Estos son el tipo de características que deben analizarse al momento de seleccionar el método de minería es más apropiado para un estudio particular.

El mencionado trabajo investigativo, tiene como desventaja que se enfatiza en comparar y evaluar modelos generados sólo empleando redes neuronales, sin embargo se realiza una buena descripción de la métrica y del método de evaluación, que para este estudio es la RMSE (Root Mean Square Error y la validación cruzada, respectivamente.

Por su lado (Yeh & Lien. 2009) indican que la mayoría de los estudios de comparación acerca de la precisión de la clasificación se realizan bajo la medida de "tasa de error", explican por qué no es útil en este estudio. Emplean como medida entonces el "gráfico de elevación" que es explicado brevemente. Luego realizan las comparaciones a través de los valores arrojados por las dos medidas, teniendo en cuenta que la muestra de los datos se separó en un grupo de entrenamiento y otro de validación. Adicionalmente, para establecer la precisión de la predicción de no pago se emplea el coeficiente de determinación múltiple (R^2) de los modelos de regresión.

El estudio antes referenciado, genera un aporte puntual al soporte o justificación de la importancia de realizar un análisis previo al momento de seleccionar las técnicas que podrían ser empleadas para aplicar sobre un estudio, al plantear la conveniencia de calcular la probabilidad de no pago de un cliente en lugar de clasificarlos como clientes de riesgo y de no riesgo. Otra ventaja de esta investigación es la ejecución de análisis comparativos sobre la clasificación y la predicción de las técnicas empleadas, cosa que no se observa en otros estudios. Se observa con esto que el análisis previo de supuestos y pre-requisitos no se enfoca exclusivamente a los datos, también incluye las conveniencias particulares del área de estudio. Como desventaja de este estudio y de muchos otros, ocurre que los datos empleados en este no son los más convenientes teniendo en cuenta futuros trabajos donde se pretenda realizar comparaciones, es decir, que para comparaciones entre resultados de investigaciones sería conveniente que se trabaje con el mismo conjunto de datos, de tal manera que trabajar con un conjunto de datos tan particular no representa una ventaja.

En (kurt et al., 2008) se presenta un trabajo más profundo al comparar cinco técnicas para predecir la presencia o ausencia de una enfermedad coronaria analizando la edad, sexo, si fuma, diabetes, hipertensión, masa corporal entre otros. Esta investigación tiene como ventaja la aplicación de análisis de clúster jerárquico y el escalamiento multidimensional, dos técnicas poco empleadas en el establecimiento de medidas de desempeño. Esta investigación también ilustra los conceptos de sensibilidad, especificidad, tasa de predicción positiva, tasa de predicción negativa, como conceptos a evaluar para determinar el desempeño de las diferentes técnicas. Adicionalmente realiza un estudio descriptivo de los datos de algunas variables con las que se trabaja, descripción que no fue realizada pensando en un posterior mapeo con los supuestos y restricciones de las técnicas que permitieran determinar cuál podría ser empleada para abordar el problema, de hecho no existe ninguna descripción de las razones por las cuales dichas técnicas fueron escogidas, siendo esta una de las desventajas que se observa en este trabajo.

Por otra parte (Weaver, D. 2004) realiza una revisión de literatura sobre la aplicación de técnicas de minería de datos para el descubrimiento de fármacos, indica que estos análisis se realizan basados en estructuras de tablas de datos, donde cada fila corresponde al elemento que compone el fármaco y las columnas son una medida experimental de los componentes, a partir de los cuales se podrían generar modelos mediante técnicas de minería de datos, que permitieran predecir valores de propiedades claves de ciertos componentes y obtener ganancias dentro de la relación estructura-actividad de los mismos.

El reporte de esta investigación inicia diferenciando entre técnicas lineales y no lineales, resaltando las ventajas de interpretación y comprensión de modelos lineales y la dificultad para la representación matemática de los modelos no lineales, aportando bajo esta distinción, la importancia de evaluar características inherentes a las técnicas antes de seleccionar alguna para aplicar a un estudio en particular.

Como ventajas que se destacan en el documento relacionado, se encuentran las indicaciones sobre ciertos factores a considerar para iniciar un proyecto de minería de datos, los cuales se observan a través de preguntas como ¿cuál técnica de minería de datos debería ser usada? ¿es un problema que tiene una relación lineal o es probable que no sea lineal?, ¿cómo será validado el modelo de minería de datos?, ¿es necesario medir la precisión de la predicción o clasificación?, ¿existen suficientes datos para una minería efectiva?. Esta última pregunta indica un nuevo aspecto a considerar previo a la aplicación de las técnicas. Es decir, según la pregunta, es importante analizar el tamaño del conjunto de datos, a razón de que el desempeño de algunas técnicas está asociado a este aspecto. Además, confirma la importancia del conjunto de datos con los que se trabaja, al indicar que actualmente no se puede afirmar que una técnica de minería tiene mejor desempeño que otra, a pesar de los cuidados que han tenido los

investigadores y emplear el mismo conjunto de datos para poder establecer comparaciones entre diferentes investigaciones que tienen el mismo objetivo.

A pesar de tan sustanciales aportes el documento se queda corto en relación a la metodología y deja de lado la definición de algunos conceptos importantes como las técnicas de evaluación que se emplearon por los diferentes estudios revisados.

En la literatura científica se encuentran diferentes investigaciones que abordan el estudio de comparación de diferentes técnicas de minería de datos y que representan un acercamiento a la investigación que se pretende realizar.

A continuación, se compendia la información correspondiente a la revisión de la literatura realizada hasta ahora para esta investigación. Los criterios bajo los cuales se analizan los diferentes artículos son los siguientes:

1. El artículo tiene como objeto de estudio el proceso de selección y comparación entre técnicas de minería de datos.

2. El artículo desarrolla la actividad de análisis de supuestos o pre-requisitos para establecer las técnicas que son aplicables.

3. La investigación aporta información sobre alguna de las técnicas que se pretenden analizar en esta propuesta de investigación (Regresión Lineal, Reglas de Asociación, Cópulas).

4. El estudio realiza algún aporte ya sea conceptual o procedimental al objeto de esta investigación.

5. El artículo realiza una buena descripción de los métodos y medidas a considerar para el proceso de evaluación; así como la medición del desempeño de técnicas de minería de datos.

A continuación, en la Tabla 1, se presentan las fuentes bibliográficas más relevantes para este trabajo y se relacionan con aspectos de interés para la investigación realizada.

Tabla 1. Relación de artículos y aportes.

Autores	1	2	3	4	5	Observaciones
Spott, M., Nauck, D. 2006.	x	x		x		
I-Cheng, Y., Che-hui, L. 2009.	x			x	x	Comparación entre seis técnicas para un modelo de predicción.
Baesens et al. 2009.				x		

Autores	1	2	3	4	5	Observaciones
Mitra, S., Mitra, P. 2002.				x		
Weaver, D. 2004.	x		x	x		Realiza estudios entre métodos lineales y no lineales, por lo que incluye la Regresión Lineal.
Ngai et al. 2009.				x		Concepto de asociación
Scott, p., Wilkins, E. 1999.	x			x		
Kim, Y. 2008.	x		x			Comparación empírica variando el tamaño de la muestra.
Serna, S. 2009.	x		x	x	x	
Hung, S., et al	x			x	x	
Delene et al. 2009	x			x		
Chen, w., Du, Y. 2009	x			x	x	
Enke, D., Thawornwong, S. 2005	x	x		x	x	
Liao et al. 2009			x			Reglas de asociación
Lahiri, R. 2006	x		x	x	x	
Frees, E., Valdez, E. 1997			x			Cópulas
Kurt et al. 2008	x	x		x		
Ozisikyilmaz et al. 2008	x		x	x		

CAPÍTULO I

2. Caracterización de las técnicas

Como parte de esta investigación, se considera necesario ilustrar la aplicación de la guía que se define, para lo cual se plantea el uso de la regresión lineal múltiple, reglas de asociación y cópulas como técnicas a emplear para abordar un problema dado. Por tal razón, en este capítulo se procede a caracterizar cada una de estas técnicas, indicando sus fundamentos matemáticos o estadísticos básicos, supuestos bajo los cuales pueden ser aplicadas y consideraciones para su uso.

2.1 Regresión Lineal

En (Ozisikyilmaz, 2008) se define el análisis de regresión lineal múltiple como una técnica estadística para investigar y modelar las relaciones entre variables. Esta técnica permite identificar un conjunto de variables regresoras que influyen en el comportamiento de una variable dependiente.

El modelo de regresión lineal más simple, tiene la forma:

$$y = B_0 + B_1 X + E \quad \text{(Ecuación del modelo de Regresión Lineal Simple)}$$

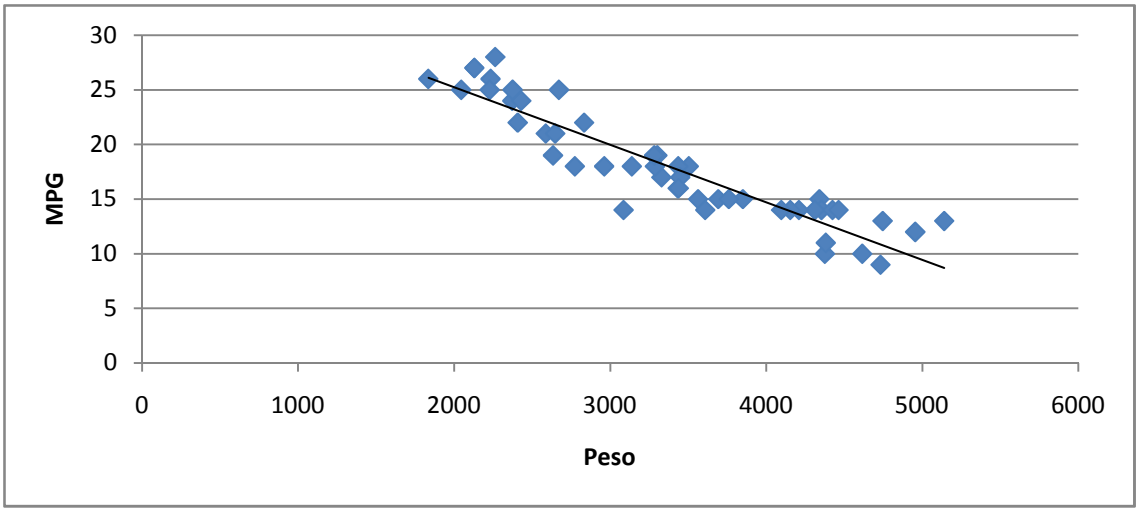
Donde B_0 y B_1 representan los coeficientes usados para describir la respuesta como una función lineal de las variables regresoras, más un error aleatorio. Sin embargo en la mayoría de los problemas del mundo real, el comportamiento de una variable es afectado por varios factores, de tal manera que un estudio de regresión lineal aplicado a estos casos tendría varias variables regresoras. Este tipo de modelos es conocido como regresión lineal múltiple, cuya ecuación es de la forma

$$y = B_0 + B_1 X_1 + \cdots + B_i X_n + E \quad \text{(Ecuación del modelo de Regresión Lineal Múltiple)}$$

En ambos modelos de regresión los coeficientes B_1,.., B_i, son conocidos como los coeficientes de regresión, los cuales se interpretan como "*el cambio esperado en la respuesta Y por cambio unitario* X_n*, cuando todas las demás variables regresoras se mantienen constantes*". Estos coeficientes son determinados a través del método de mínimos cuadrados del error (LSE por su sigla en inglés).

Gráficamente un modelo de regresión lineal lo que pretende es encontrar la función que mejor se ajuste a un conjunto de datos.

Gráfico 1. Ilustración gráfica de un modelo de Regresión Lineal Simple.



Los modelos de regresión lineal múltiple son empleados teniendo en cuenta varios supuestos, entre los que están (Montgomery, 2002):

- La media del error es cero $E(\varepsilon i)=0$
- La varianza del error es constante $Var(\varepsilon i) = \delta^2$
- Los errores no están correlacionados $E(\varepsilon_i \varepsilon_j)\neq 0$ y
- El supuesto de normalidad sobre los errores, el cual se emplea para fines de probar hipótesis y establecer intervalos de confianza.

Todos estos supuestos serán tenidos en cuenta durante el desarrollo de la investigación que se propone, dado que son factores vitales para considerar la selección de la técnica más apropiada para determinado estudio.

Además de los supuestos considerados por la regresión existen otros aspectos a considerar, características propias de los modelos de regresión que influyen al momento de considerar su aplicación para un trabajo o investigación. En (Montgomery, 2002) se indican algunas consideraciones sobre el uso de la regresión, dado que en muchas ocasiones es mal empleada. Entre las consideraciones se encuentran las siguientes:

- Los modelos de regresión tienen por objeto servir de ecuaciones de interpolación dentro del intervalo de las variables regresoras que se usan para ajustarlos. Se debe tener cuidado al extrapolar muy por fuera de ese intervalo.
- El modelo de regresión lineal múltiple permite modelar el comportamiento de una variable específica (variable dependiente) con relación a las otras (variables independientes) y evaluando únicamente el comportamiento en caso de que sólo una de las variables independientes cambie.

- El hecho de que un análisis de regresión haya indicado que existe una fuerte relación entre dos variables, no implica que éstas tengan relación alguna en el sentido causal. El análisis de regresión sólo puede examinar los asuntos de regresión, no puede manejar el asunto de la necesidad. De manera que los modelos de regresión lineal no deben ser empleados para tratar de establecer relaciones causa-efecto.
- Resulta bastante importante considerar la distribución de los valores, dado que valores extremos en los datos tienen influencia en la pendiente de la recta de regresión, en casos de regresión lineal simple y en casos de regresión lineal múltiple también afectan el modelo. En estos casos es importante tomar acciones correctivas.
- En problemas con datos de series de tiempo, la suposición de errores no correlacionados muchas veces no es adecuada, dado que $E(\varepsilon_i \varepsilon_{j+1}) \neq 0$, es decir que estos términos de error están correlacionados.

Además de estas características propias del modelo de regresión, existe otra particularidad que se desprende de la linealidad de este modelo. (Weaver, 2004) presenta un trabajo comparativo entre modelos lineales y no lineales. Entre sus conclusiones se resalta que:

- Los modelos no lineales pueden ser más precisos respecto a los modelos lineales, a expensas del nivel de interpretación del mismo, puesto que los primeros suelen ser matemáticamente más complejos.
- Los modelos lineales se caracterizan por su simplicidad y facilidad de interpretación, quedando de esta manera a consideración del analista y dependiendo de las necesidades del problema escoger entre un modelo lineal y no lineal.

2.2 Reglas de Asociación

Las reglas de asociación son una herramienta empleada principalmente para encontrar relaciones entre ítems o características a partir de un conjunto de datos.

Las Reglas de Asociación se definen formalmente en (Wang et al, 2004) así:

Sea $I = \{i_1, i_2, \ldots, i_n\}$ el conjunto de ítems, donde cada ítem corresponde a un artículo específico. *D* representa una base de datos comercial, en la cual cada transacción *T* representa un conjunto de ítems, es decir que $T \subseteq I$. Adicionalmente cada conjunto de ítems, es un conjunto de sub-ítems no vacío. Con esta anotación, se puede definir una regla de asociación de la forma

$$X \rightarrow Y \text{ donde } X, Y \subseteq I \text{ y } X \sqcap Y \neq \emptyset$$

La forma de las reglas de asociación, en un entorno de comercialización de productos, indica que si los artículos de *X* son comprados, es posible que también se compren los de *Y*.

Para las reglas de asociación se definen dos medidas de evaluación, referidas en el estudio recién mencionado, como son *Soporte* y *Confianza.* La medida de soporte (denotada como *Sup(X, D)*) indica la tasa de comercialización de los artículos de *X* en *D*, en otras palabras, indica la tasa de transacciones en *D* que contienen el conjunto de ítems *X*.

Respecto a las reglas, el *Sup(X U Y, D)* indica la tasa de transacciones en *D* que contienen a *X U Y* (Liao et al, 2009). Si se tienen valores altos de soporte, existe mayor certeza de que las reglas obtenidas son veraces, dado que se está indicando que existen muchas transacciones en la base de datos que respaldan la regla obtenida.

La medida de confianza (denotada como *Conf (X -> Y)*) permite establecer que si la comercialización incluye a *X*, existe una alta probabilidad de que también incluya a *Y*. Esta medida se establece de la siguiente manera:

$$Confianza(X \rightarrow Y) = \frac{Soporte(X\ U\ Y, D)}{Soporte\ (X, D)}$$

De acuerdo a (Agrawale and Shafer, 1996) citado en (Liao et al, 2009) las reglas de asociación pueden ser obtenidas en dos pasos. El primero, es identificar el conjunto de ítems frecuentes en la base de datos cuyo soporte sea más grande que *MinSup* (tasa de referencia definida por el analista de datos) y el segundo paso es generar reglas de asociación con base en un conjunto de ítems identificados. Las reglas que se crean deben cumplir las siguientes dos condiciones.

Sup(*X U Y, D*) ≥ *MinSup*

Conf(X -> Y) ≥ *MinConf*, donde *MinConf*, es una medida de referencia definida por el analista para evaluar las reglas significativas.

2.2.1 Minería de reglas de asociación

En el trabajo de (Kuo and Shih, 2007) se indica que el problema de identificar reglas de asociación se divide en dos sub-problemas:

- Encontrar conjuntos de ítems frecuentes cuyo soporte este por encima del soporte mínimo establecido.

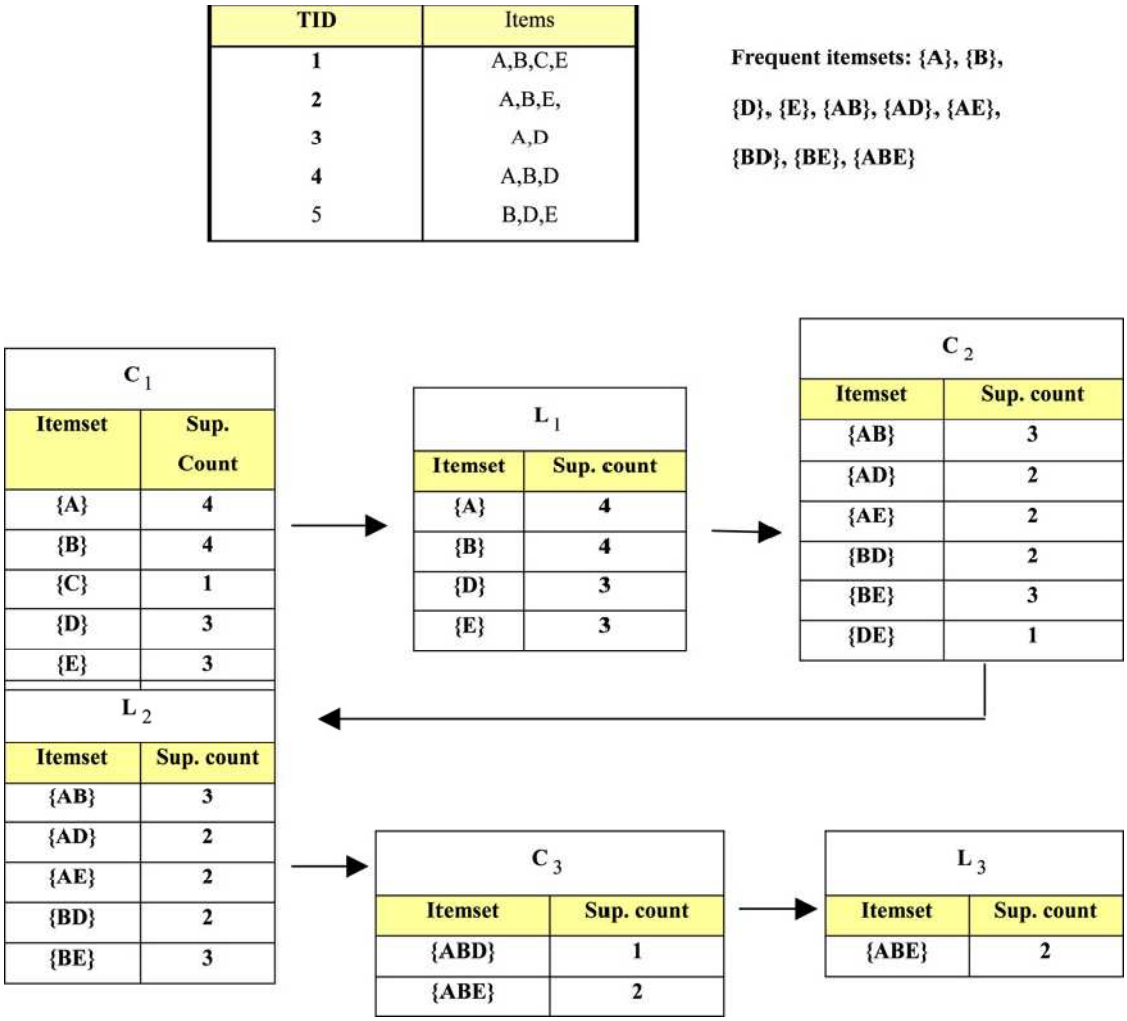
- Usar el conjunto definido en el primer paso para generar reglas de asociación que tengan un nivel de confianza por encima del nivel mínimo establecido para esta medida.

Teniendo en cuenta este enfoque para la creación de reglas de asociación, las investigaciones alrededor de esta temática se han enfocado en el desarrollo de algoritmos eficientes para identificar el conjunto de ítems frecuentes y generar las reglas.

Entre los algoritmos más conocidos se encuentran:

Algoritmo Apriori: Algoritmo que realiza búsquedas sobre grandes conjuntos de ítems. Este genera *k*-conjunto de ítems candidatos a partir de (*k-1*) conjuntos de ítems frecuentes encontrados previamente, evaluando posteriormente el soporte mínimo definido de los conjuntos candidatos para formar *k*-conjunto de ítems frecuentes. Este algoritmo tiene como desventaja la necesidad de recorrer varias veces los datos, afectando su eficiencia, por lo que se le considera como un acercamiento muy costoso, sobre todo en grandes volúmenes de información.

Gráfico 2. Ilustración del algoritmo A priori. Extraído de (Kuo and Shih, 2007)



Identificación de ítem frecuentes con soporte = 2

Donde C_i corresponde al conjunto de ítems frecuentes obtenidos en cada iteración que se realiza sobre los L_{i-1}, que corresponde a las transacciones que cumplen con las condiciones de soporte y confianza mínimos definidos evaluadas sobre C_{i-1}. Es decir C_1 es un conjunto de ítems frecuentes, L_1, es el conjunto de ítems que cumplen con las condiciones de soporte y confianza mínimos definidos y a partir del cual se obtiene el conjunto de ítems frecuentes C_2.

Algoritmo Parallel Mining: Otra técnica empleada para mejorar el algoritmo clásico de minería de reglas de asociación (Apriori), este considera la existencia de múltiples procesadores en el entorno computacional. La idea central del algoritmo es separar las tareas de minería en muchas tareas pequeñas, de tal manera que varias de estas puedan ser ejecutadas simultáneamente sobre varios procesadores. De esta manera se mejora la eficiencia en la extracción de reglas de asociación.

Algoritmo Sampling: Algoritmo que aplica una técnica de muestreo aleatorio para identificar reglas de asociación y disminuir la actividad en la base de datos. La idea es seleccionar una muestra aleatoria y encontrar usando esta muestra todas las reglas que probablemente existan en la base de datos, las cuales se verificarían posteriormente con el resto de la base de datos. En el caso donde el método de muestreo no produzca todas las reglas de asociación, las reglas faltantes pueden ser encontradas en un segundo paso. De esta forma, en la mayoría de los casos, este algoritmo requiere solo un paso completo por la base de datos y solamente dos casos en el peor de los casos.

Algoritmo Partition*:* La propuesta de este algoritmo es dividir la base de datos en secciones lo suficientemente pequeñas para ser manejadas en la memoria principal, de tal manera que una vez que se tenga una parte en el disco, la generación y evaluación de conjuntos candidatos para esta parte es ejecutada en memoria principal sin ninguna actividad en base de datos.

Este algoritmo computa todos los conjuntos frecuentes en dos pasos sobre la base de datos. El primero consiste en identificar en cada parte, la colección de todos los conjuntos frecuentes locales, para luego en el segundo paso unir las colecciones de conjuntos frecuentes locales y usarlos como los conjuntos candidatos.

2.2.2 Características de las Reglas de Asociación

Entre las características más destacadas se tiene que:

- No es posible expresar reglas con atributos numéricos, dados que estas reglas no pueden ser descubiertas por los métodos existentes (Qodmanan et al, 2011). En caso de que se desee trabajar con variables cuantitativas será necesario realizar procesos de discretización.

- Las reglas de asociación tienen asociadas problemas de efectividad y eficiencia, dada la cantidad de reglas inútiles que pueden ser generadas con cualquier método y el tiempo que toma generar estas reglas.

- Facilidad de interpretación de resultados. Las reglas de asociación son fáciles de interpretar dada la naturaleza sencilla de su estructura, facilitando el trabajo de los analistas a la hora de evaluar los resultados y generar conocimiento.

- Posibilidad de trabajar con datos difusos. Algunos investigaciones han avanzado en el desarrollo de algoritmos que permitan trabajar con datos que por su naturaleza son difusos o porque han sido procesados para proveerlos con cierto grado de imprecisión.

- Trabajar con reglas de asociación implica un proceso de selección para determinar el algoritmo que será empleado para generar las reglas.

2.3 Cópulas

Una definición formal y un acercamiento conceptual a las cópulas es presentado en (Nelsen, 1999), donde se refieren a cópulas como "*funciones que juntan o acoplan funciones de distribución multivariante a sus funciones de distribución marginales univariantes*" o como "*funciones de distribución cuyas marginales unidimensionales son uniformes en el intervalo [0,1]*".

La definición formal de cópulas que se presenta se fundamenta en el teorema de Sklar.

2.3.1 Teorema de Sklar

Sea *H* una función de distribución conjunta con marginales *F* y *G*. Entonces existe una cópula '*C*' tal que para todo x, y en '*R*' se cumple que:

$$H(x, y) = C(F(x), G(y))$$

Si *F* y *G* son continuas entonces *C* es única. De otra forma *C* es únicamente determinada sobre *RanF* x *RanG*. Recíprocamente, si *C* es una cópula y *F* y *G* son funciones de distribución, entonces la función *H* es una función de distribución conjunta, con marginales *F* y *G.*

En la literatura existe poco material de referencia vinculado a esta temática. Sin embargo en (Frees & Valdes, 1997) se presenta además del concepto asociado, algunas consideraciones bajo las cuales es posible emplear las cópulas en lugar de los métodos tradicionales. La anotación más relevante está asociada a la distribución de los datos. Este artículo indica que la distribución normal multivariada tiene un amplio dominio en el estudio distribuciones, pero se reconoce la necesidad de examinar alternativas a la distribución normal y emplear distribuciones como la bivariante Pareto y la bivariante Gamma. Sin embargo estas distribuciones tienen como inconveniente que:

- Una familia diferente de cópulas es necesaria para cada distribución marginal
- Análisis para casos diferentes a los bivariantes no son claros, y que
- Las medidas de asociación frecuentemente aparecen en distribuciones marginales.

Inconvenientes que no ocurren al emplear distribuciones normales, dado que para estas, las distribuciones marginales también son normales.

Teniendo en cuenta la necesidad de trabajar con distribuciones alternativas a la distribución normal, es necesario considerar métodos alternativos a los tradicionales y es aquí donde aparece la figura de cópulas.

2.3.2 Familias de cópulas

En (Frees & Valdes, 1997) se describen los métodos "Archimedean" y "Compounding" empleados para especificar familias de cópulas, las cuales se caracterizan por presentar una estructura determinada para modelar problemas. Este documento explica que la generación de una familia de cópulas arquimedianas depende de una función generadora la cual se identifica en la ecuación que define a este tipo de cópulas, ilustrada a continuación:

$$C_\Phi = \Phi^{-1}\ (\Phi(u) + \Phi(v))\ \text{ para } u,v \in (0,1]$$

Donde Φ es llamado el generador de la cópula C_Φ y cuyo valor determina el tipo de familia que se origina.

Tabla 2.Cópulas arquimedianas y sus generadores. Inspirado en (Frees & Valdes, 1997)

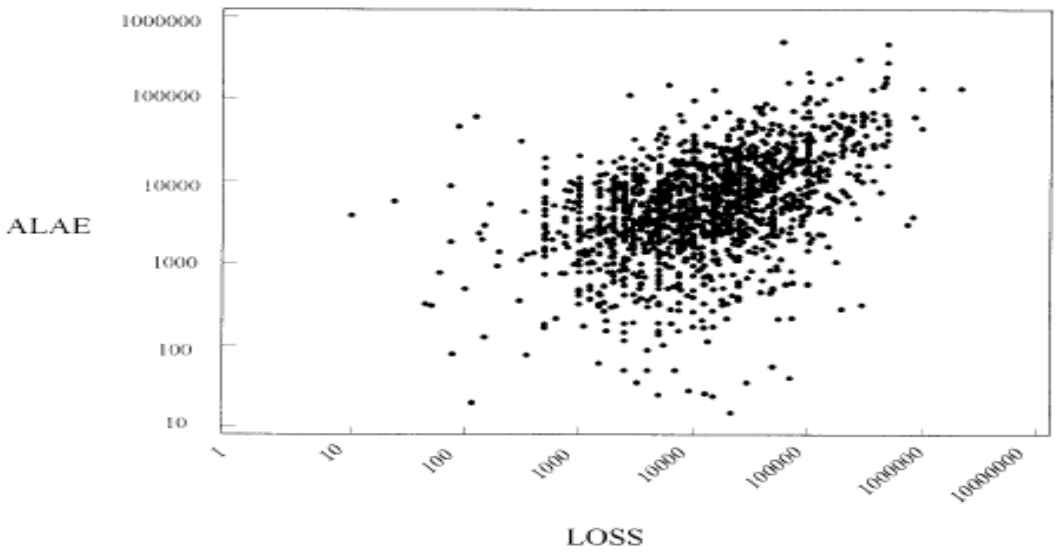
Familia	Generador Φ(t)	Dependencia del parámetro espacio	Cópula Bivariada $C_\Phi(u,v)$
Independence	- ln t	No aplica	Uv
Clayton(1978)	$t^{-\alpha}-1$	$\alpha > 1$	$(u^{-\alpha} + v^{-\alpha} -1)^{-1/\alpha}$
Gumbel (1960), Hougaard(1986)	$(-\ln t)^{\alpha}$	$\alpha \geq 1$	$\text{Exp}\{- [(-\ln u)^{\alpha} + (-\ln v)^{\alpha}]^{1/\alpha}\}$
Frank (1979)		$-\infty < \alpha < \infty$	$1/\alpha \ln (1 +)$

A causa de la existencia de la variedad de familias de cópulas, trabajar con estas implica la tarea de seleccionar la que mejor se adapte a los datos con los que se trabaja, originando un grado de complejidad adicional al trabajo con esta técnica.

La forma de abordar un problema empleando el modelo cópulas es brevemente descrito en el documento que se mencionó previamente, donde se muestra el ajuste de una cópula al problema de *reclamos de indemnización de una compañía de seguros*, donde cada reclamo consta de una valor de indemnización (pérdidas) (*x1*) y unos gastos asignados a las pérdidas (*x2*). El objetivo es describir la distribución conjunta de pérdidas y gastos.

El primer paso es realizar una análisis gráfico empleando un diagrama de dispersión de las dos variables para visualizar si existe una posible correlación entre estas. El gráfico, junto con una medida estadística, sugiere una correlación entre las variables.

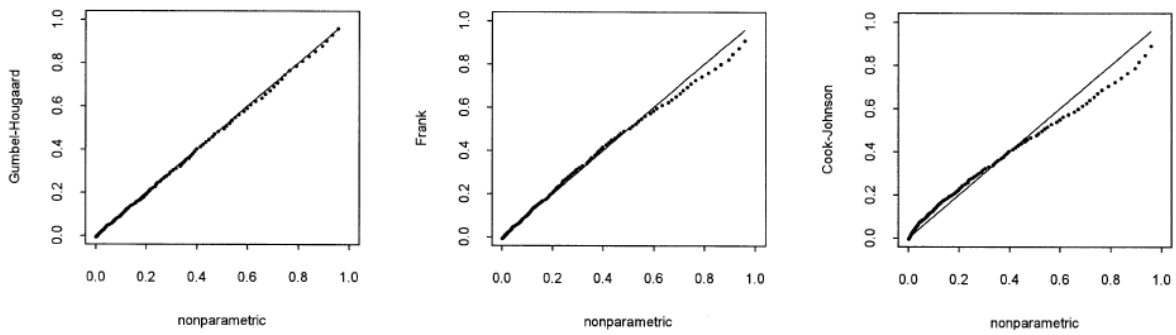
Gráfico 3. Gráfico de dispersión que sugiere la relación entre pérdida y gastos. Extraído de Frees & Valdes, 1997.



De esta manera el paso siguiente, teniendo presente la ecuación del modelo de Sklar, es determinar las funciones de distribución marginal de las variables, que para *x1* y *x2* corresponden a una distribución Pareto y una distribución Kaplan-Meier respectivamente. Además de las funciones de distribución marginal $F(x1)$ y $G(x2)$ la ecuación también indica la necesidad de ajustar la cópula "C", para lo cual se determina la forma de la cópula y luego se estima usando el método de máxima verosimilitud.

Para determinar la cópula más apropiada se hace uso del gráfico *q-q* y se comparan las versiones paramétricas y no paramétricas de las distribuciones, donde el factor determinante para la selección es el grado de cercanía entre estas dos versiones

Gráfico 4. Comparación de versiones paramétricas y no paramétricas, para determinar cuál familia de cópula modela de mejor manera los datos. Extraído de Frees & Valdes, 1997



Como se observa en el gráfico 4, la imagen más a la izquierda permite sugerir el uso de la cópula Gumbel-Hougaard para este caso.

De esta manera se determina cual familia de cópula se emplea para abordar el objetivo de determinar la función de distribución conjunta de pérdida y gastos para el caso de *reclamos de indemnización de una compañía de seguros*.

2.3.3 Características de las cópulas:

- La técnica de cópulas no tiene restricciones para trabajar con cualquier tipo de distribución marginal, es decir con cualquier tipo de distribución de las variables bajo estudio.
- Permite el análisis de funciones de distribución conjunta a través de sus funciones de distribución marginal, facilitando la labor de hallar la probabilidad de que ciertos eventos ocurran simultáneamente.
- Esta técnica no está orientada a abordar problemas donde se tiene solo una variable de salida o dependiente, esta permite explorar problemas en los cuales se debe analizar el comportamiento conjunto de varias variables de salida (Frees & Valdes, 1997)
- Dentro del proceso de trabajo con esta técnica es necesario realizar actividades para seleccionar el tipo de cópula que más se adapta al problema que se aborda.
- Modelo con alta complejidad estadística y matemática, aplicado principalmente dentro de procesos de análisis multivariante.
- Insensible a transformaciones (Schweizer & Wolff, 1981, citado en Frees & Valdes, 1997). Schweizer & Wolff demostraron que las variables transformadas de *g1(x1)* y *g2*(*x2*) de *x1* y *x2* respectivamente tiene la misma cópula que *x1* y *x2*. De esta forma, la manera en la que *x1* y *x2* se mueven juntas es capturada por la función cópula independiente de la escala en la cual cada variable es medida.

CAPITULO II

3. Guía metodológica

3.1 Conceptos

Como parte de la tarea de definición de la guía que se propone para facilitar el trabajo de los analistas de datos y proporcionar criterios científicos para la actividad de seleccionar y evaluar técnicas de minería de datos enfocadas en el descubrimiento de asociaciones, es necesario definir los principales conceptos que se utilizarán en la guía con el objetivo de evitar confusiones y ambigüedades y proporcionar suficiente claridad al momento de emplearla.

A continuación, se listan los principales conceptos. Aunque ya se han mencionado antes, es importante presentar una definición más formal:

Guía: La real academia española define guía como "*Aquello que dirige o encamina*", "*Tratado en que se dan preceptos para encaminar o dirigir en cosas, ya espirituales o abstractas, ya puramente mecánicas.*" entre otras definiciones.
Es justo la primera definición la que se pretende implementar con esta investigación. Lo que se desea es proporcionar a los analistas o mineros de datos un marco de referencia que dirija las actividades y consideraciones necesarias a llevar a cabo para abordar el problema de seleccionar la técnica de minería de datos más apropiada para establecer asociaciones.

Método: Definido por la real academia española como "*modo de decir o hacer con orden*", "*modo de obrar o proceder, hábito que cada uno tiene y observa*", "*procedimiento que se sigue en la ciencias para hallar la verdad y enseñarla*" entre otras definiciones.
Este trabajo de investigación a través de esta guía pretende presentar un método que sirva como punto de partida o referencia al momento de realizar el proceso de selección objeto de esta investigación.

Técnica de minería de datos: Herramientas estadísticas y de la inteligencia artificial empleadas para la extracción de patrones potencialmente útiles. En el caso específico de esta investigación son herramientas para el descubrimiento de asociaciones entre variables.
En este trabajo investigativo se estudian las reglas de asociación, la regresión lineal múltiple y las cópulas como técnicas de minería de datos.

Asociación y dependencia: (Ngai et al, 2009) define una asociación dentro del contexto de minería de datos, como un modelo cuyo objetivo es establecer relaciones entre ítems que existen de manera conjunta en un almacén de datos. Las relaciones

establecidas o encontradas se establecen en muchos casos con base a una medida de dependencia que determina el grado de asociación entre las variables.

Criterios de evaluación y métricas: Son las herramientas empleadas para evaluar y medir el desempeño de la aplicación de las diferentes técnicas de minería a un problema determinado y verificar que los resultados sean óptimos y adecuados en cuanto a lo que se esperaba de ellos. Durante el desarrollo de la guía se especificaránn algunas métricas que se emplean para medir el desempeño, en este caso de las reglas de asociación, las cópulas y la regresión lineal múltiple.

Dataset: Conjunto de datos sobre los cuales se aplicaran las técnicas y se validará el modelo de guía propuesto.

3.2 Definición de la guía

Teniendo en cuenta que el problema que aborda esta investigación es la falta de formalismo y criterio científico para afrontar la tarea de selección de una técnica de minería de datos, en este caso enfocada en el establecimiento de asociaciones, esta guía pretende definir una serie de actividades y tareas a seguir para afrontar este problema.

Es importante resaltar que el proceso de selección de una técnica va más allá de identificar las que pueden ser aplicables al problema que se afronta. Para esta investigación la actividad de selección incluye también la aplicación de las técnicas que pueden emplearse para el desarrollo del análisis en cuestión y la evaluación del desempeño de las mismas, dado que dependiendo del resultado del desempeño, se determina cuál modelo es más apropiado o sobre cual apoyar la toma de decisiones.

Adicionalmente, es necesario aclarar que la guía y sus actividades no pretenden ser un instrumento definitivo para seleccionar o descartar la aplicación de determinadas técnicas a un problema. La decisión final y definitiva la realiza el analista de datos o minero de datos. Se recuerda que este proceso tiene mucha influencia humana que es decisiva. Esta guía sólo pretende orientar sobre el desarrollo de actividades concernientes al proceso de selección, además de proporcionar un marco de trabajo formal.

La guía consta de tres etapas macros que abarcan las principales actividades que se definirán, dichas etapas se determinaron con base en el tiempo de ejecución de las diferentes actividades. Considerando esto, la primera etapa es llamada **Análisis**, la segunda se denomina **Aplicación** y la última es la fase de **Evaluación.**

3.2.1 Análisis

Esta etapa y sus actividades se derivan de la necesidad evidente y develada por algunos autores, de analizar una variada cantidad de aspectos técnicos y del negocio u organización que influyen en el proceso de selección.

(Carrier & Povel, 2003, citado en Ngai et al.,2009) sugieren que para escoger una técnica de minería de datos apropiada se deben estudiar las características de los datos y los requerimientos del problema, sugerencia que es aplicada en el trabajo de (kurt et al., 2008) donde se realiza un estudio descriptivo de algunas de las variables involucradas en el ejercicio desarrollado en su trabajo. Adicionalmente, conocido el hecho de que algunas técnicas de minería de datos poseen características particulares y supuestos o condiciones de aplicación, es necesario analizarlas, además de otros factores.

A continuación, se describen las actividades a desarrollar en esta etapa.

3.2.1.1 Análisis del negocio y objetivos de la solución

Esta actividad tiene como propósito dar a conocer al analista el objetivo que se pretende cumplir, el cual, al igual que en muchos otros casos, se emplea como orientación para el desarrollo de la solución y en el que se encuentran subyacentes características de la solución.

¿Por qué es importante?

Sin duda esta actividad es necesaria y determinante en cualquier tipo de proyecto, particularmente en un proyecto de minería de datos, hace parte de un proceso previo que hay que abordar antes de llegar a la etapa de selección y evaluación de técnicas de minería. Considerando que existen modelos de asociación, clasificación, predicción, agrupamiento ¿cuál es más apropiado?. Para atender esto es necesario conocer los objetivos que se desean alcanzar, dado que puede desearse un modelo para clasificar personas o cosas, o un modelo para predecir el comportamiento de personas o cosas, sin embargo este problema no es el que se atiende en esta investigación dado que para esta, el modelo de trabajo es la asociación.

Dentro del contexto específico de esta investigación, dado que ya se conoce que se trabajará con el modelo de asociación, es importante validar nuevamente que el objetivo que se persigue si es posible cumplirlo a través de este modelo.

En el trabajo de (Yeh & Lien. 2009), cuyo objetivo es realizar análisis de riesgo, que en este caso es la probabilidad de un retraso en el reembolso de un crédito concedido, se observa como deciden calcular las probabilidades de que un cliente no pague a tiempo en lugar de clasificarlos en clientes de riesgo y de no riesgo, considerando el hecho de

que se posee más información con las probabilidades para tomar decisiones sobre estos.

Es así como conocer el objetivo que se persigue y el entorno del negocio es importante para el proceso de seleccionar una técnica de minería. Si la decisión hubiera sido la de clasificar a los clientes, los métodos para el análisis de datos a considerar hubieran sido los de clasificación y las medidas de evaluación estarían asociadas a la precisión de la clasificación.

Otra ventaja de entender el contexto y los objetivos es conocer las prioridades a atender en el problema. En (Baesens et al., 2009.) se ilustra la situación del "*Base1 II Capital Accord"* el cual estimula a las entidades financieras a desarrollar modelos de minería de datos para estimar el riesgo de pérdida, riesgo de incumplimiento y riesgo de exposición, con la problemática asociada de que los reguladores financieros son reacios a aprobar el uso de modelos complejos y de caja negra, dado que para estos es de mayor relevancia la claridad y transparencia en los patrones de datos.

De esta forma se manifiesta nuevamente la importancia de conocer el contexto del negocio y el objetivo que se pretende alcanzar, dado que para el negocio manifestado en el ejemplo al que se hace referencia anteriormente, tiene mayor prioridad la claridad y transparencia de los datos, influyendo esto en la selección de las técnicas de minería a ser empleadas, dado que las técnicas o modelos de caja negra serían la opción de menor consideración.

Intuitivamente se observa que el análisis de característica de las técnicas es importante para el proceso de selección, sin embargo esta actividad será descrita con mayor detalle más adelante.

3.2.1.2 Caracterización de datos

Cualquier proceso de análisis de información involucra un estudio de las variables que hacen parte del dataset de trabajo. Este estudio involucra actividades como el análisis de correlación, análisis de cluster, análisis descriptivo de variables, entre otros. Los resultados de estos, permiten a los analistas tomar decisiones que afectan los resultados de los modelos generados, escoger la técnica de trabajo o determinar las transformaciones necesarias para ejecutarlas, teoría que se observa en el trabajo de (Spott & Nauck, 2006), el cual indica que para determinar cuáles métodos de minería de datos pueden ser aplicados se parte de los requerimientos y de los datos.

La actividad de análisis de datos en este punto ya ha sido iniciada, se recuerda que previamente al proceso de selección y ejecución de técnicas de minería de datos, existe la etapa de pre-procesamiento, donde se realiza un estudio de los datos con el propósito de garantizar la calidad de los mismos y generar confianza sobre los

resultados obtenidos, de manera que la actividad de análisis que aquí se describe retoma los estudios y resultados obtenidos en dicha etapa.

En el trabajo de (Moreno et al., 2008), cuyo propósito es obtener reglas de asociaciones a partir de un conjunto de datos cuyas variables son continuas, se especifica la importancia de esta actividad, al resaltar que la aplicabilidad y relevancia de las asociaciones descubiertas depende principalmente de cómo son discretizados los datos. El proceso de discretización aunque es una actividad de transformación, parte de un análisis del tipo de variables que se encuentran dentro del conjunto de datos de trabajo y del análisis de supuestos bajo los que trabaja la técnica de minería, en este caso las reglas de asociación.

Otro referente de la importancia de esta actividad se observa en (Frees and Valdez, 1997) donde se menciona el dominio que ha tenido la distribución normal dentro del estudio de análisis multivariante, al igual que reconoce la necesidad de examinar alternativas a la distribución normal, considerando que muchas de las variables del entorno real poseen distribuciones distintas a la normal. Cuando se tienen distribuciones no normales algunos analistas realizan transformaciones para normalizar los datos y trabajan considerando que se tienen distribuciones normales o por el contrario trabajan con técnicas que permiten establecer dependencias sin restricciones sobre el tipo de distribución, como las cópulas. De esta manera se observa como características de los datos, en este caso la distribución, tienen importancia al momento de seleccionar la técnica para aplicar a un estudio.

Por experiencia también es posible destacar la significancia de esta etapa, donde se retoma o se realiza la definición y descripción de variables, tal como se observa en el trabajo de (Hung et al., 2006) el cual tiene un apartado que describe las variables de interés asociadas a la problemática que se estudia. Aunque dentro de este trabajo no se realiza ningún análisis específico con esta información, es posible emplearla para tomar decisiones con repercusiones sobre los modelos.

Como ejemplo de estas decisiones se tiene el caso de las variables *FechaNacimiento* y *Edad,* cuya característica particular es que a partir de cualquiera de ellas se puede inferir la otra, de tal manera que si las dos son incluidas dentro de un análisis existiría información duplicada, que además de no aportar información adicional valiosa, genera posibles inconsistencias, mayor complejidad dentro del modelo o sistema, restándole simplicidad, facilidad de interpretación, entre otras características, incluso afectando negativamente las medidas de evaluación del mismo.

La caracterización de los datos, dependiendo de las circunstancias, puede tener repercusiones directas sobre la selección de la técnica o puede servir como referente para analizar aspectos que afectan el modelo a generar, como se observa en las apreciaciones descritas anteriormente. Sin embargo, no hay duda de la necesidad del

desarrollo de esta actividad para tener mayor certeza a la hora de seleccionar y evaluar una técnica de minería para elaborar modelos de asociaciones.

3.2.1.3 Descripción y análisis de técnicas

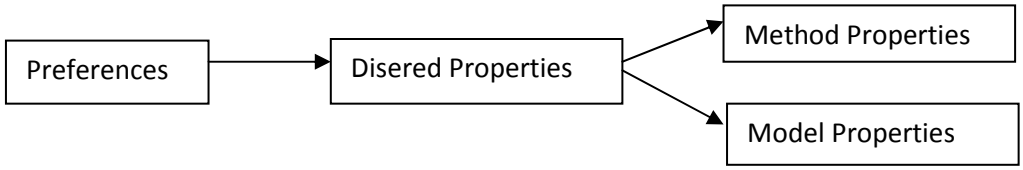
Como se mencionó en el apartado de *Análisis del negocio y objetivo de la solución* (Capítulo II, etapa de Análisis), el análisis de características de las técnicas es importante para el proceso de selección. Esta actividad tiene como objetivo identificar las características de cada una de las técnicas consideradas para establecer modelos de asociación y cómo afectan la decisión de trabajar o no con la técnica, en un caso particular.

Se inicia con la descripción detallada de las técnicas que son consideradas para generar los modelos de asociación, permitiendo conocer las ventajas y desventajas de las mismas, junto con los supuestos que se deben cumplir. Además, identificar si estas técnicas son de tipo supervisado o no supervisado, los tipos de datos con los que trabaja entre otros aspectos. A partir de este conocimiento se genera una estrecha relación con las actividades descritas anteriormente, dado que una vez se conocen los supuestos, se ha realizado un análisis sustancial de los datos y se tienen claros los objetivos que se persiguen, lo que resta es realizar un mapeo de toda la información hasta ahora identificada, obteniendo como resultado las primeras conclusiones respecto a cuáles técnicas podrían ser empleadas para abordar el problema de minería que se presenta.

El detalle de las técnicas puede incluir las ventajas y desventajas conocidas a partir de conocimiento establecido por experiencias previas, el análisis de supuestos bajo las cuales la técnica puede ser aplicada y consideraciones sobre su aplicación que eviten incurrir en errores. Ejemplos de la forma de realizar el detalle se encuentra en los trabajos de (Meyfroidt et al., 2009), (Yeh & Lien, 2009), (Enke & Thawornwong, 2005) y en el primer capítulo de esta investigación donde se describen las reglas de asociación, la regresión lineal múltiple y las cópulas.

El proceso de mapeo se observa explícitamente en el estudio de (Spott & Nauck, 2006) que muestra un esquema general para mapear requerimientos con propiedades de los métodos. De igual manera menciona características de los modelos que son consideradas dentro del proceso de mapeo, consideración que no aplica para nuestra investigación dado que el modelo de asociación es nuestro modelo de investigación ya definido.

Gráfico 5: Esquema de actividad de mapeo. Extraído de (Spott & Nauck, 2006)

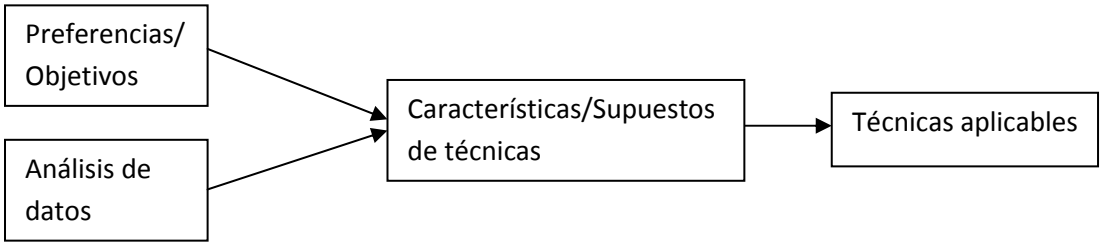


La investigación de (Spott & Nauck, 2006) también lista un conjunto de propiedades a evaluar durante el desarrollo de esta actividad de descripción y análisis de técnicas:

- **Facilidad de explicación:** facilidad de entender, aplicar e interpretar los resultados arrojados por la técnica.
- **Tipo de resultado arrojado por la técnica**: concerniente a la forma de arrojar los resultados ya sean reglas o funciones.
- **Adaptabilidad a nuevos datos**: capacidad para incluir nuevos datos dentro del estudio sin necesidad de volver a realizar gran parte del trabajo hasta ahora realizado.
- **Balance entre precisión y simplicidad**: virtud de la técnica de poder ser precisa sin afectar drásticamente la simplicidad de los resultados y viceversa.

A continuación se muestra gráficamente el proceso de mapeo que se propone.

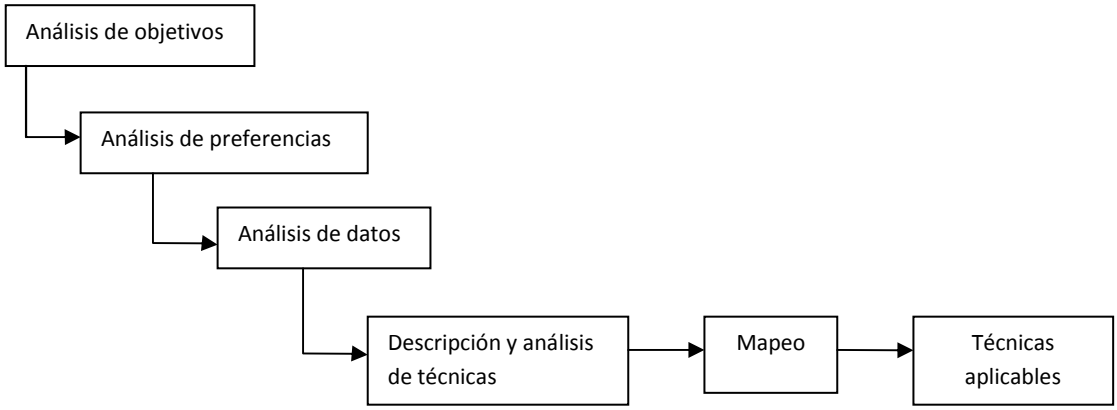
Gráfico 6: Esquema de actividad de mapeo propuesto



En resumen, la etapa de análisis propuesta considera los aspectos más destacados en los trabajos estudiados y los determinados por la experiencia propia. Se puede observar que se contemplan aspectos técnicos, consideraciones propias del problema, se analizan características de los datos y se percata lo relevante de factor humano que con base en sus preferencias toma decisiones.

A continuación se resume gráficamente la etapa de Análisis definida para la guía metodológica.

Gráfico 7: Esquema de proceso de Análisis de la guía propuesta



3.2.2 Aplicación

Una vez terminada la etapa de análisis, donde se han establecido las técnicas que pueden emplearse para abordar el problema, donde se han identificado las preferencias sobre los resultados que se desean, el paso siguiente es aplicar dichas técnicas con el propósito de obtener los modelos, funciones o reglas a partir de las cuales se inicia el proceso de evaluación.

En esta etapa se conoce, por concepto y por experiencia, que puede ser necesario ejecutar diferentes actividades, entre las que se encuentran:

3.2.2.1 Transformaciones

Dadas las características propias de las técnicas, algunas requieren que sean realizadas ciertas transformaciones sobre los datos para poder trabajar con ellos. Entre los cambios más comunes se encuentran los procesos de discretización descrito en (Moreno et al., 2008) y normalización, el primero empleado para convertir una variable de tipo cuantitativa a una variable de tipo categórica y el segundo para escalar los valores de una variables en un rango pequeño como entre -1 y 1 o entre 0 y 1.

Las transformaciones que se realizan son consecuencia de los procesos de análisis descritos para la etapa previa. Esta actividad se realiza a través de diferentes aplicaciones que facilitan el manejo de la gran cantidad de datos que son necesarios convertir. Sin embargo, es un proceso que tiene gran influencia del factor humano, específicamente el de discretización, dado que de acuerdo al conocimiento y preferencias pueden variar los resultados de las transformaciones.

3.2.2.2 Selección de algoritmos

Algunas técnicas de minería de datos han tenido procesos de evolución para mejorar sus características y de esta manera afectar de manera positiva el desempeño de las mismas. Como consecuencia de estos procesos existe una variada cantidad de algoritmos que pueden ser empleados para generar el modelo que la técnica ofrece. Como ejemplo de modelos que han mejorado se encuentran las reglas de asociación, cuyo progreso se observa en el trabajo de (Kuo & Shih, 2007) donde se describen las variaciones que se han realizado a partir del algoritmo A priori; así como las redes bayesianas y con los algoritmos para realizar agrupamientos.

Con base a esta situación en algunos casos será necesario escoger el algoritmo que se considere más apropiado. La selección del algoritmo varía de acuerdo a las ventajas ofrecidas, al tipo de datos de trabajo, entre otras consideraciones.

3.2.2.3 Selección de modelos internos

El estudio constante alrededor del tema de minería de datos ha permitido que para algunas técnicas se hayan desarrollado varios tipos de arquitectura, ocasionando que para cada una exista más de un modelo que permita trabajar con ella. En (Enke & Thawornwong, 2005) se referencia el caso de tres tipos de redes neuronales, la aplicación de cada una y su posterior evaluación.

El trabajo de (Frees and Valdez, 1997) también permite identificar la necesidad de seleccionar entre diferentes modelos de cópulas. Se observa cómo se comparan y se selecciona la que más se ajusta a los datos de trabajo.

La forma de evaluar y seleccionar modelos propios de una misma técnica varía dependiendo de la misma. Esta actividad no se describe con gran detalle dado que no todas las técnicas tienen la característica de poseer un conjunto de modelos internos de donde es necesario escoger uno, adicionalmente para el desarrollo de esta actividad podría ser necesario la construcción de otra guía, dado que puede llegar a ser compleja.

3.2.2.4 Datos de entrenamiento

Las redes neuronales y los árboles de regresión y clasificación son dos técnicas que desarrollan modelos de entrenamientos, los cuales son verificados y luego empleados para llevar a cabo diferentes tareas de predicción y clasificación. Para llevar a cabo este proceso es necesario definir un conjunto de datos de entrenamiento y un conjunto de datos de validación. Los primeros se emplean para obtener el modelo que permite afrontar las necesidades del estudio y el segundo para evaluar el desempeño del modelo obtenido. La mayoría de las ocasiones estos dos grupos se escogen de manera aleatoria, como sugiere (Yeh & Lien. 2009), y la cantidad de elementos de cada grupo depende del tamaño del conjunto de datos de trabajo. Es muy importante que

el conjunto de entrenamiento sea lo más grande posible para que el modelo resultante abarque todos los posibles casos que se pretendan evaluar con el conjunto de validación.

De manera que en la etapa de aplicación de las técnicas, es posible que sea necesario identificar estos dos conjuntos.

3.2.2.5 Selección de herramientas de aplicación

Una actividad fundamental en este proceso es la selección de la herramienta bajo la cual se aplicará la técnica, en la actualidad existen muchas aplicaciones computacionales que permiten la ejecución de un conjunto variado de estas, cada una con diferentes grados de complejidad, diferentes niveles de fiabilidad, acceso y soporte que afectan la decisión sobre cual emplear. Los sistemas más potentes son los desarrollados por Oracle y Microsoft y su área de inteligencia de negocios, diseñados para abordar procesos de gran envergadura y con muchas ventajas de fiabilidad y soporte, pero con limitaciones respecto a su acceso dado el alto costo que es necesario pagar a estas empresas. Sin embargo existen otros sistemas que son empleados con mayor frecuencia en ámbitos académicos y empresariales que lo permitan.

Las aplicaciones van desde Excel de Microsoft Corporation que permite realizar cubos OLAP y desarrollar modelos de regresión lineal, MatLab que contiene una gran variedad de técnicas incorporadas y cuenta con una gran fiabilidad, el software R, que es una lenguaje que permite desarrollar algunos algoritmos asociados a diferentes técnicas, ofreciendo gran versatilidad, WEKA que cuenta con un grupo de algoritmos para modelar asociaciones y es de libre distribución, entre otras aplicaciones.

3.2.3 Evaluación

La evaluación de los modelos, las funciones o reglas obtenidas una vez se ha aplicado una técnica, es una actividad presente en todos los estudios y procesos de minería de datos. El objetivo de esta etapa es determinar la confianza que se puede depositar sobre los modelos, para tomar decisiones o decidir emplearlos como herramientas para predecir, describir o clasificar. Es en este punto donde nuevamente se realiza la actividad de selección, dado que como consecuencia de los resultados de la evaluación, es posible descartar algunos modelos y por ende, algunas técnicas para abordar la solución de determinado problema.

Como en cualquier proceso de evaluación es necesario definir unos indicadores de evaluación a partir de los cuales se realizan las comparaciones, dichos indicadores varían dependiendo del modelo que se pretenda evaluar y de los objetivos de los investigadores. Para modelos de agrupamiento se está más interesado en medir la estabilidad del grupo, mientras que el interés en modelos de clasificación está en

medir la estabilidad de los datos, capacidad de discriminación y/o calibración de la probabilidad (Baesens et al., 2009), (Yeh & Lien, 2009).

Cuando se habla de evaluar un modelo, en palabras más concretas, se habla de evaluar que tan bueno es su desempeño, cuán fácil es leer e interpretar sus resultados. Cada uno de estos aspectos a considerar tiene sus criterios y medidas, que pueden ser tan formales y rigurosos como una fórmula matemática o estadística o por el contrario muy informal como el criterio de expertos.

3.2.3.1 Desempeño

El desempeño se refiere a que tan bien, el modelo obtenido realiza la tarea para la cual fue construido, que tan bueno es para predecir, clasificar o describir unos valores, clases o datos. Esta medida involucra el estudio de otros aspectos como la precisión, tiempo de ejecución y cantidad de variables empleadas (parsimonia). Un buen desempeño es un adecuado balance entre estos aspectos. Es decir, un buen desempeño implica alta precisión, poco tiempo de ejecución para obtener los resultados y el menor número de variables posibles.

Sobre el desempeño es necesario destacar que con el paso del tiempo se va degradando evidenciando la necesidad de realizar ajustes al modelo para evitar que quede obsoleto. El efecto de pérdida de desempeño se debe a tres aspectos; el primero, la variabilidad de la muestra la cual representa un visión limitada de la población que crece y varía con el tiempo. El segundo corresponde a cambios de estrategias o exploración de nuevos segmentos de mercados, los cuales no tienen cabida bajo el modelo de trabajo actual. Por último las influencias macro-económicas cambiantes que podrían afectar de manera positiva o negativa el modelo (Baesens et al., 2009).

Algunas medidas como *confianza* y *soporte* afectan directamente el desempeño, en este caso, de las reglas de asociación, dado que limitan el número de reglas que se pueden generar. La *complejidad*, es otra medida que podría afectar el desempeño de un modelo, esta se mide dependiendo de la forma de la función obtenida, por ejemplo, se podría contar la cantidad de nodos en un árbol de decisiones o cantidad de condiciones en un conjunto de reglas (Scott & Wilkins, 1999).

(Baesens et al., 2009) señala que el desempeño de un modelo, además de los aspectos ya mencionados, también se ve afectado por la actividad de *entrenamiento* (capítulo II - Aplicación - Datos de entrenamiento). Entre más se entrene el modelo y mayor sea la cantidad de datos de entrenamiento mejor será el desempeño del mismo.

3.2.3.2 Legibilidad y facilidad de interpretación

Dentro del contexto de la inteligencia de negocios, uno de los problemas manifiestos es la consolidación de la misma en áreas administrativas, siendo la complejidad asociada a los modelos de trabajo una de las razones por los cuales algunos sectores aún no están conformes con el área. De tal manera que la facilidad que ofrecen los modelos para ser empleados por las personas que toman las decisiones o alimentan el mismo y la facilidad que ofrecen para la interpretación de los resultados, será vital a la hora de determinar qué tan bueno o conveniente es el modelo obtenido. Es mucho más sencillo, aplicar e interpretar los resultados de las reglas de asociación, que aplicar y entender el uso de los modelos cópulas que poseen una rigurosidad matemática y estadística compleja.

Sin embargo, en medio de estos aspectos a evaluar existe una paradoja; lograr un modelo más interpretable, frecuentemente viene en detrimento del desempeño del mismo, ocasionando entonces dos preguntas destacadas para esta etapa de evaluación (Baesens et al., 2009).

¿Cuál es la representación preferida de modelos, en términos de simplicidad, por ejemplo, modelos basados en reglas, modelos lineales o modelos gráficos?

¿Cómo se puede escoger el balance óptimo entre un modelo interpretable y un modelo con buen desempeño?

Sin duda, la respuesta a ambas preguntas está asociada a los objetivos, requerimientos y preferencias establecidas para el problema que se aborde.

A continuación se describen algunas de las medidas más empleadas para evaluar los modelos de minería de datos.

Soporte y Confianza

Entre los problemas principales de los modelos de reglas de asociación se encuentra la generación de patrones que resultan poco interesantes, la gran cantidad de reglas descubiertas y el bajo desempeño de los algoritmos (Moreno et al, 2008), siendo los dos primeros, referentes para determinar el desempeño del modelo de reglas de asociación. Si se tiene una gran cantidad de reglas descubiertas y sólo un porcentaje pequeño resulta ser útil e interesante para los usuarios, entonces el desempeño del modelo será cuestionado por su bajo rendimiento.

Dentro de las medidas empleadas para filtrar las reglas y determinar las que posiblemente son determinantes y útiles, se encuentran el *soporte* y *confianza*, que permiten al usuario, de acuerdo a sus necesidades y rigurosidad, establecer un nivel mínimo de cumplimiento sobre estas medidas y garantizar que las reglas obtenidas

satisfacen sus condiciones. Una mayor descripción de las medidas se tiene en el primer capítulo de este trabajo.

Coeficiente de correlación

El coeficiente de correlación es una medida comúnmente usada para conocer la magnitud y la dirección de la relación estadística que puede existir entre un conjunto de variables. Los posibles valores del coeficiente oscilan entre [-1, 1], donde valores cercanos a menos uno (-1), indica que existe una fuerte relación entre las variables y que a medida que una aumenta la otra disminuye; valores cercanos a uno (1) indica una fuerte relación entre las variables, pero a diferencia del caso anterior, cuando una aumenta la otra también aumenta. Valores cercanos a cero (0) indica que no es posible afirmar que existe una relación entre las variables que se estudian.

- **Coeficiente de correlación de Pearson (*r*):** Es el método más comúnmente usado para medir la dependencia entre dos variables. Este coeficiente que relaciona las variables *X* y *Y*, se define como:

$$\frac{\sum(x_i - xprom)(y_i - yprom)}{\sqrt{\sum(x_i - xprom)^2}\sqrt{\sum(y_i - yprom)^2}}$$

Donde X_i y Y_i son un par de puntos, $-X_{prom}$ Y $-Y_{prom}$ son la media de cada variable.

Este coeficiente es apropiado para mostrar la relación lineal entre variables. No es invariante a transformaciones no lineales monótonas, es decir, la correlación entre *X* y *Y* no es igual a la correlación entre Ln(x) y Ln(y).

- **Coeficiente de correlación de Spearman's (r_s):** Este a diferencia del anterior se calcula usando rangos de valores en lugar de los valores reales, por lo que se indica que este es empleado para medir la relación entre dos conjuntos de grupos de datos. Se define como:

$$\frac{\sum(r_x - r_x\,prom)(r_y - r_y prom)}{\sqrt{\sum(r_x - r_x prom)^2}\sqrt{\sum(r_y - r_y prom)^2}}$$

Donde R_x y R_y son los rangos de *X* y *Y;* R_{xprom} y R_{yprom} son la media del rango de las variables. A diferencia del coeficiente de Pearson, este no requiere que la relación entre las dos variables sea lineal, ni existe ninguna suposición sobre la distribución de las variables, adicionalmente este coeficiente es invariante a transformaciones no lineales sobre las variables.

- **Kendall's tau**: Este coeficiente, al igual que el coeficiente de correlación de Sperman, no requiere ninguna suposición sobre la distribución de las variables y no requiere que la relación entre las variables sea lineal. Se define como:

$$\frac{\#(pares\ concordantes) - \#(pares\ discordantes)}{\frac{n(n-1)}{2}}$$

Donde *n* es el tamaño de la muestra y los pares concordantes indican el número de pares que se están moviendo en el mismo sentido y los pares discordantes son aquellos que se están moviendo en direcciones opuestas a cada uno de los otros.

En el trabajo de (Al-Harthy et al, 2007) se observa la aplicación de estas medidas sobre modelos generados a partir de Cópulas, Iman-Conover entre otros.

Raíz del error cuadrático (RMSE): Medida de evaluación comúnmente usada (Kim, 2008), (Viscarra & Behrens, 2010) para determinar la precisión de los modelos considerando la diferencia entre los valores predichos y los observados. Formalmente se define como:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$$

Donde X_i es el valor predicho y Y_i es el valor observado. A menor valor de la medida, mejor es la precisión de los modelos evaluados.

Criterio de información de Akaike (AIC): Criterio empleado para evaluar directamente el desempeño de la técnica considerando la Raíz del Error Cuadrático y el cantidad de variables empleadas para la obtención del modelo o resultado, es decir, considerando la precisión y la parsimonia. Se define como:

$$AIC = nln\ RMSE + 2p$$

Donde "*n*" es la cantidad de elementos de la muestra y "*P*" es el número de variables que se emplearon en la obtención del modelo. Entre más alto sea el AIC mejor será el modelo que representa.

CAPITULO III

4. Aplicación de la guía

Este capítulo tiene como objetivo mostrar la aplicación de la guía propuesta, a través de un caso de estudio que giran en torno al proceso de selección y evaluación de técnicas de minería de datos para modelar asociaciones. Para esto, se evaluaron diversos conjuntos de datos con características distintas que permiten emplear la guía bajo distintas circunstancias.

Los conjuntos de datos evaluados provienen del *UC Irvine Machine Learning Repository* (http://archive.ics.uci.edu/ml/) y se caracterizan por ser fácilmente accesibles a cualquier persona y orientados a validar o comparar técnicas de minería de datos.

Caso de estudio 1:

Este caso de estudio de evaluación de la guía, emplea el conjunto de datos denominado "**Breast Cancer Wisconsin (Diagnostic) Data Set"**, con las siguientes características:

Información general del dataset

Fecha de disponibilidad: 1995-11-01

Área de estudio: Salud

Tipo de atributos: Multivariable

Cantidad de registros: 569

Cantidad de variables: 11 (Id, diagnóstico, 9 variables relacionadas)

Valores faltantes: Sí (cantidad: 16 - variables: Bare Nuclei)

Referencias de uso:

W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.

W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April 1995.

Planteamiento del problema:

El conjunto de datos contiene los valores de diferentes aspectos evaluados sobre una masa mamaria y el resultado del diagnóstico sobre ésta, indicando si es benigna o maligna. Considerando esto, el problema que se plantea es determinar el grado de asociación que existe entre las variables planteadas como objeto de estudio y el tipo de tumor, con el propósito de evaluar la relevancia de algunas de ellas dentro del resultado del diagnóstico.

Solución:

Dado que el problema planteado busca encontrar asociaciones entre variables y el resultado de un diagnóstico, la solución radica en emplear una técnica de minería de datos que permita generar modelos de asociación. Para nuestro problema se considera el uso de las reglas de asociación, regresión lineal múltiple y cópulas como herramientas para modelar asociaciones.

Para seleccionar una o varias de estas técnicas que permitan soportar las decisiones respecto al diagnóstico del cáncer de mama, se emplea la guía metodológica propuesta en esta investigación mostrando a continuación el desarrollo de cada una de las tareas y actividades que se definen en la misma.

4.1 Etapa de análisis

- **Análisis del negocio y objetivo de la solución**

El problema que se estudia gira en torno al diagnóstico del cáncer de mama a partir del estudio del tumor presente en el paciente. Este trabajo parte del análisis de la información que se tiene de diferentes características presentes en los tumores y el resultado del diagnóstico asociado a dicho tumor.

El objetivo es determinar la relación que existe entre las variables evaluadas sobre el tumor y el resultado del diagnóstico, con el propósito de valorar la relevancia de dichas variables al momento de concluir si el tumor es maligno o benigno.

Respecto al problema, no existen detalles específicos que permitan tomar decisiones diferentes hasta las ahora consideradas, como lo es buscar un modelo de asociación, no de clasificación y la consideración del uso de técnicas conocidas como reglas de asociación, regresión lineal múltiple y cópulas para modelar las asociaciones descubiertas. De igual manera, no existe especificación que indique si se desea tener mayor nivel de interpretación de los resultados o mayor exactitud en las asociaciones. No se tiene información de expertos, ni se tiene trabajo o estudio para la comparación.

- **Caracterización de datos**

El conjunto de datos disponibles para este problema consta de once variables, de las cuales una corresponde a la identificación del registro y otra al resultado del diagnóstico, mientras que las otras nueve son los aspectos evaluados sobre el tumor. A continuación se describen las variables involucradas en el problema.

Tabla 3. Variables incluidas en el diagnóstico de cáncer de mama para este problema

Número atributo	Atributo	Traducción
1	**Samplecodenumber**	**Código identificador**
2	ClumpThickness	Espesor del tumor
3	Uniformity of CellSize	Uniformidad del tamaño celular
4	Uniformity of CellShape	Uniformidad de la forma celular
5	Marginal Adhesion	Adhesión marginal
6	Single EpithelialCellSize	Tamaño individual de la célula epitelial
7	BareNuclei	Núcleo desnudo
8	BlandChromatin	Textura del núcleo
9	Normal Nucleoli	
10	Mitoses	Mitosis
11	Class	Diagnóstico

La información en la tabla anterior fue obtenida desde la fuente donde se descargó el dataset. Sin embargo, los metadatos son insuficientes para poder iniciar el proceso de selección y ejecución de alguna de las técnicas consideradas.

Para este problema se considera la significancia de cada variable, el tipo de dato que maneja, las unidades de medida bajo las cuales fue evaluada, el tipo de distribución que posee y posibles transformaciones.

Significado de cada variable

- Clump Thickness: medida referente al espesor de la masa mamaria que se estudia.
- Uniformity of Cell Size/Shape: las células cancerígenas tienden a variar en tamaño y forma, por lo cual evaluar estas características es importante.
- Marginal adhesión: las células normales tienden a permanecer juntas, mientras las células cancerígenas pierden esta habilidad.
- Single Epithelial Cell Size: relacionada con la uniformidad antes mencionada. Las células que están significativamente amplias pueden ser malignas.
- Bare Nuclei: característica que indica cuando el núcleo de las células no está rodeado por el citoplasma.
- Bland chromatin: describe una textura uniforme de los núcleos vistos en células benignas. En células cancerígenas esta textura tiende a ser más burda.

- Normal nucleoli: nucleoli son pequeñas estructuras vistas en los núcleos. En células normales, estas estructuras usualmente son muy pequeñas. En células cancerígenas la estructura nucleoli se vuelve más prominente.

Tipos de datos:

De las 10 variables que se describen, 9 son variables cuantitativas con un rango de valores entre 1 y 10. La variable *Class*, a diferencia de las otras es dicotómica, dado que solo toma los valores de 2(benigno) ó 4(maligno).

Unidades de medida:

Para este dataset se desconocen las unidades de medidas empleadas para la evaluación de cada variable. Algunas podrían resultar evidentes como los tamaños, sin embargo para otras como las variables Bare Nuceli y Mitoses se desconoce la razón por las cuales se escogieron medidas cuantitativas, ni la significancia de cada valor.

Estadísticas básicas:

Las medidas estadísticas básicas permiten realizar las primeras inferencias sobre las variables, dado que se pueden determinar los mínimos, máximos, promedios, moda, además las gráficas permiten determinar frecuencias, valores atípicos entre otros. Para la obtención de estas medidas y gráficas se empleó el software STATGRAPHICS, el cual facilita la obtención de estas medidas y gráficas.

A continuación, se describe para cada variable, sus estadísticas básicas.

Tabla 4: Estadísticas básicas de las variables de estudio de cáncer de mama

	Media	Moda	Desviación Estándar	Mínimo	Máximo	Valores faltantes
ClumpThickness	4,417	1	2,818	1	10	0
Uniformity of CellSize	3,138	1	3,053	1	10	0
Uniformity of CellShape	3,211	1	2,973	1	10	0
Marginal Adhesion	2,809	1	2,857	1	10	0
Single EpithelialCellSize	3,2178	2	2,2154	1	10	0
BareNuclei	3,548	1	3,645	1	10	16
BlandChromatin	3,4384	2	2,4401	1	10	0
Normal Nucleoli	2,870	1	3,055	1	10	0
Mitoses	1,5903	1	1,7162	1	10	0

Un acercamiento al tipo de distribución de las variables es determinado a través del histograma de frecuencias correspondiente a cada una de ellas. Para la obtención de los histogramas se emplea STATGRAPHIC los cuales se ven en el anexo 1. Como resultado del análisis de estas gráficas se tiene que las variables poseen las siguientes distribuciones:

Tabla 5: Tipos de distribución de las variables de estudio de cáncer de mama

Atributo	Tipo de distribución
ClumpThickness	No normal
Uniformity of CellSize	No normal
Uniformity of CellShape	No normal
Marginal Adhesion	No normal
Single EpithelialCellSize	No normal
BareNuclei	No normal
BlandChromatin	No normal
Normal Nucleoli	No normal
Mitoses	No normal

- **Caracterización de las técnicas**

En el capítulo I de esta investigación se observa la descripción detallada de las reglas de asociación, la regresión lineal múltiple y las cópulas, que son las técnicas escogidas para modelar asociaciones para este estudio. Sin embargo, a continuación se retoman las características más relevantes a tener en cuenta sobre las mismas.

Regresión lineal:

Los modelos de regresión lineal son empleados teniendo en cuenta varios supuestos, entre los que están:

- La media del error es cero $E(\varepsilon i)=0$
- La varianza del error es constante $Var(\varepsilon i) = \delta^2$
- Los errores no están correlacionados $E(\varepsilon_i \varepsilon_j)\neq 0$ y
- El supuesto de normalidad, el cual se emplea para fines de probar hipótesis y establecer intervalos de confianza.

Además de los supuestos considerados por la regresión existen otros aspectos a considerar, características propias de los modelos de regresión que influyen al momento de considerar su aplicación para un trabajo o investigación. En (Montgomery, 2002) se indican algunas consideraciones sobre el uso de la regresión, dado que en muchas ocasiones es mal empleada. Entre las consideraciones se encuentran las siguientes:

- Los modelos de regresión tienen por objeto servir de ecuaciones de interpolación dentro del intervalo de las variables regresoras que se usan para ajustarlos. Se debe tener cuidado al extrapolar muy por fuera de ese intervalo.
- El modelo de regresión lineal múltiple permite modelar el comportamiento de una variable específica (variable dependiente) con relación a las otras (variables independientes) y evaluando únicamente el comportamiento en caso de que sólo una de las variables independientes cambie.
- El hecho de que un análisis de regresión haya indicado que existe una fuerte relación entre dos variables, no implica que éstas tengan relación alguna en el sentido causal. El análisis de regresión sólo puede examinar los asuntos de regresión, no puede manejar el asunto de la necesidad. De manera que los modelos de regresión lineal no deben ser empleados para tratar de establecer relaciones causa-efecto.
- Resulta bastante importante considerar la distribución de los valores, dado que valores extremos en los datos tienen influencia en la pendiente de la recta de regresión, en casos de Regresión Lineal Simple y en casos de Regresión Lineal Múltiple también afectan el modelo. En estos casos es importante tomar acciones correctivas.
- En problemas con datos de series de tiempo, la suposición de errores no correlacionados muchas veces no es adecuada, dado que $E(\varepsilon_i \varepsilon_{j+1}) \neq 0$, es decir que estos términos de error están correlacionados.

Además de estas características propias del modelo de regresión, existe otra particularidad que se desprende de la linealidad de este modelo. (Weaver, 2004) presenta un trabajo comparativo entre modelos lineales y no lineales. Entre sus conclusiones se resalta que:

- Los modelos no lineales pueden ser más precisos respecto a los modelos lineales, a expensas del nivel de interpretación del modelo, puesto que los modelos no lineales suelen ser matemáticamente más complejos.
- Los modelos lineales se caracterizan por su simplicidad y facilidad de interpretación, quedando de esta manera a consideración del analista y dependiendo de las necesidades del problema escoger entre un modelo lineal y no lineal.

Reglas de asociación:

Entre las características más relevantes se destacan:

- La falta de posibilidad para expresar reglas con atributos numéricos, dados que estás reglas no pueden ser descubiertas por los métodos existentes (Qodmanan et al, 2011). Bajo estas circunstancias será necesario discritizar.
- Facilidad de interpretación de resultados. Las Reglas de Asociación son fáciles de interpretar dada la naturaleza sencilla de su estructura, facilitando el trabajo de los analistas a la hora de evaluar los resultados y generar conocimiento.
- Posibilidad de trabajar con datos difusos. Algunos investigaciones han avanzado en el desarrollo de algoritmos que permitan trabajar con datos que por su naturaleza son difusos o porque han sido procesados para proveerlos con cierto grado de imprecisión.
- Trabajar con Reglas de Asociación implica un proceso de selección para determinar el algoritmo que será empleado para generar las Reglas.

Cópulas:

Esta técnica posee las siguientes características:

- No posee restricciones para trabajar con cualquier tipo de distribución marginal, es decir con cualquier tipo de distribución de las variables bajo estudio.
- Permite el análisis de funciones de distribución conjunta a través de sus funciones de distribución marginal, facilitando la labor de hallar la probabilidad de que ciertos eventos ocurran simultáneamente.
- Dentro del proceso de trabajo con cópulas es necesario realizar actividades, con el propósito de seleccionar el tipo de cópula que más se adapta al problema que se aborda.
- Modelo con alta complejidad estadística y matemática, aplicado principalmente dentro de procesos de análisis multivariante.
- Insensible a transformaciones (Schweizer & Wolff, 1981, citado en Frees & Valdes, 1997). Schweizer & Wolff demostraron que las variables transformadas de *g1(x1)* y g2(x2) de x1 y x2 respectivamente tiene la misma Cópula que x1 y x2. De esta forma, la manera en la que x1 y x2 se mueven juntas es capturada por la función Cópula independiente de la escala en la cual cada variable es medida.

- **Mapeo de la información**

Una vez conocidos los objetivos y las restricciones del problema, cuando se ha establecido la metadata de las variables y se han especificado las características destacadas de las técnicas, solo resta mapear esta información para obtener los primeros resultados sobre cual técnica puede ser aplicada para obtener modelos de asociación a partir de los cuales es posible determinar la relación que puede existir entre las nueve variables de estudio del tumor y el diagnóstico del mismo.

Como no existen restricciones ni especificaciones sobre la precisión o la visualización o facilidad de interpretación, estas características no son consideradas inicialmente, sin indicar esto que sean descartadas para el proceso de selección.

Teniendo en cuenta esto, el punto de partida para el mapeo es la caracterización de los datos existentes. Estos serán comparados inicialmente con los supuestos y especificaciones bajo los cuales deben trabajar las técnicas.

Regresión lineal múltiple

En la caracterización realizada al dataset **Breast Cancer Wisconsin (Diagnostic)** se observa que la variable dependiente de este estudio es de tipo dicotómica dado que el resultado del mismo es el diagnóstico del tipo de tumor, ya sea maligno o benigno, de tal manera que el modelo de regresión a usar debe ser el logístico y siendo este un caso particular del modelo regresión lineal múltiple se considera esta técnica como herramienta para abordar modelos de asociación sobre este estudio.

Reglas de asociación

La primera restricción que se observa cuando se describe la técnica es la necesidad de trabajar con datos categóricos y en vista de que todas las variables de estudio para este dataset son numéricas, se podría considerar la imposibilidad de emplear esta técnica para modelar las asociaciones en el problema planteado. Sin embargo, existe la posibilidad de discretizar las variables y obtener la información de manera categórica y poder emplear esta técnica.

Ninguna otra restricción existente sobre el problema o alguna característica propia de la técnica es impedimento para emplear esta técnica, incluso la facilidad que ofrece para la interpretación de los resultados resulta ser atractiva para abordar el problema planteado.

Respecto a la necesidad de seleccionar un algoritmo entre los que existen para obtener las reglas, esta solo representa una actividad investigativa que permita determinar el más conveniente, mas no genera mayor desgaste que permita reconsiderar el uso de la técnica.

Cópulas

Es necesario considerar que el planteamiento del problema es bastante superficial, no existe un usuario final o cliente que defina restricciones de eficiencia, sencillez o precisión sobre el problema. Se recuerda que una de las características definidas para las cópulas es su alta complejidad matemática y estadística que dificulta su aplicación y entendimiento, que bajo alguna restricción de nivel de interpretación podría descartarla para ser empleada en este caso de estudio.

Sin embargo el problema en cuestión plantea el estudio del comportamiento de una variable con base al comportamiento individual de otras, situación que descarta el uso de cópulas como herramienta para abordar este problema. En el capítulo I de esta investigación se especifica que estas abordan problemas donde es necesario analizar el comportamiento conjunto de dos o más variables de salida. Además se destaca que esta técnica aporta en términos de probabilidades y no en términos de dependencia.

De esta manera se puede concluir que las reglas de asociación y la regresión lineal múltiple pueden ser empleadas para abordar el problema planteado en el caso de estudio 1, mientras que la las cópulas se descartan para su uso.

4.2 Etapa de aplicación

Reglas de asociación

En esta etapa se ejecutarán las reglas de asociación para obtener los modelos de asociación que permitan determinar las relaciones existentes y relevantes entre las variables analizadas y el diagnóstico del tipo de tumor.

A continuación se describen las actividades que son de consideración dentro de esta etapa.

- **Transformaciones**

Discretización

Como se describió en la etapa de análisis, las reglas de asociación sólo pueden ser obtenidas a través de variables cualitativas, por tal razón y en vista de que las variables para el conjunto de datos de este problema son de tipo cuantitativa, es necesario realizar un proceso de discretización sobre estas, con el propósito de establecer categorías y con base a estas modelar reglas de asociación.

Para discretizar las distintas variables se emplea el software *Weka* que facilita el desarrollo de esta actividad. La muestra de las transformaciones se observa en el anexo dos. A continuación el resultado del proceso de discretizar:

Tabla 6: Resultados de discretizar las variables de estudio de cáncer de mama

Variable	Categorías	Rango	Cantidad
ClumpThickness	**Delgada**	[1,4]	383
	Mediana	(4,7]	187
	Gruesa	(7,10]	129
Uniformity of cellsize	**Pequeña**	[1,4]	521
	Mediana	(4,7]	76
	Grande	(7,10]	102
Uniformity of cellshape	**Clase 1**	[1,4]	512
	Clase 2	(4,7]	94
	Clase 3	(7,10]	93
Adhesion marginal	**Débil**	[1 - 5.5]	579
	Fuerte	(5.5 - 10]	120
Single EpithelialCellSize	**Pequeña**	[1,4]	553
	Mediana	(4,7]	92
	Grande	(7,10]	54
Barenuclei	**Ausente**	[1,4]	495
	Medianamente presente	(4,7]	42
	Presente	(7,10]	162
Blandchromatin	**Uniforme**	[1,4]	523
	Medianamente uniforme	(4,7]	117
	Burda	(7,10]	59
Normal nucleoli	**Pequeña**	[1,4]	541
	Mediana	(4,7]	57
	Grande	(7,10]	101
Uniformity of cellsize	**Mitosis 1**	[1,4]	659
	Mitosis 2	(4,7]	18
	Mitosis 3	(7,10]	22

- **Selección de algoritmos**

Para las reglas de asociación se han desarrollados diferentes algoritmos, de tal manera que para esta es necesario realizar un proceso que como resultado indique cual de estos algoritmos es más conveniente emplear.

Algunos software evitan esta tarea al tener implementados sólo el algoritmo Apriori o cualquier otro, mientras otros como *WEKA* ofrecen la posibilidad de escoger más de dos, de tal manera que para la obtención de reglas de asociación es necesario escoger un algoritmo de los disponibles.

Para seleccionar el algoritmo a través del cual se generan reglas de asociación es necesario conocer las características propias de cada uno y de acuerdo a estas determinar cuál emplear.

Tertius: algoritmo de programación lógica inductiva que trabaja bajo el concepto de "*Confirmatory Induction"*. Este algoritmo realiza una búsqueda óptima que intenta encontrar las k-reglas que son confirmadas por una función de confirmación. En otras palabras identifica las reglas para las cuales se obtiene el mayor valor una vez se evalúa la función de confirmación sobre cada una de ellas.

Como se observa, para la ejecución de este algoritmo es necesario definir una función de confirmación, cuyo objetivo es medir cuantas reglas son confirmadas, siendo esta una tarea adicional dentro del proceso de obtención de reglas de asociación empleando este algoritmo.

Predictiveapriori: el trabajo de (Garcia et al, 2011) cita a (Scheffer, 2005) e indica que este algoritmo mejora el Apriori al no requerir las especificaciones de las medidas de soporte y confianza que suelen definirse para identificar el conjunto de reglas que son realmente útiles. Este algoritmo logra un apropiado balance entre confianza y soporte para maximizar la precisión de la asociación.

GeneralizedSequentialPatterns: en se indica que los algoritmos de "sequential pattern" son importantes métodos de minería de datos para determinar comportamientos, relacionados con el tiempo, en bases de datos secuenciales, es decir sobre conjuntos de datos que presenten características temporales relevantes o donde el problema planteado involucra análisis en intervalos de tiempos.

Apriori: algoritmo que realiza búsquedas sobre grandes conjuntos de ítems. Este genera *k*-conjunto de ítems candidatos a partir de (*k-1*) conjuntos de ítems frecuentes encontrados previamente, evaluando posteriormente el soporte mínimo definido de los conjuntos candidatos para formar *k*-conjunto de ítems frecuentes. Este algoritmo tiene como desventaja la necesidad de recorrer varias veces los datos, afectando su eficiencia, por lo que se le considera como un acercamiento muy costoso, sobre todo en grandes volúmenes de información.

Analizando las características propias de cada algoritmo, se decide descartar el *Tertius* dado que sería necesario definir la función de confirmación, que implica un trabajo adicional que en este punto se considera innecesario dado que actualmente se tienen las medidas de soporte y confianza para evaluar una regla. El algoritmo *PredictiveApriori* no se considera porque este maneja las restricciones de soporte y confianza como sistemas de caja negra y se desea tener mayor control sobre estas medidas. El *GeneralizedSequentialPattern* no es considerado para su aplicación dado que el problema y conjunto de datos que se maneja no posee características temporales de ningún tipo. Por último se tiene el algoritmo *Apriori* el cual permite manejar las medidas de soporte y confianza para determinar las reglas más destacadas con mucha facilidad y entendimiento, es apto para el tipo de datos que se maneja y en vista de la cantidad de datos que serán procesados la desventaja asociada a la

eficiencia no es relevante en este caso, por tal motivo se decide emplear el algoritmo Apriori para la obtención de reglas de asociación al caso de estudio actual.

- **Selección de modelos internos:**

Dado que la técnica de reglas de asociación posee únicamente reglas de la forma $X \rightarrow Y$, esta actividad no se realiza para este caso de estudio.

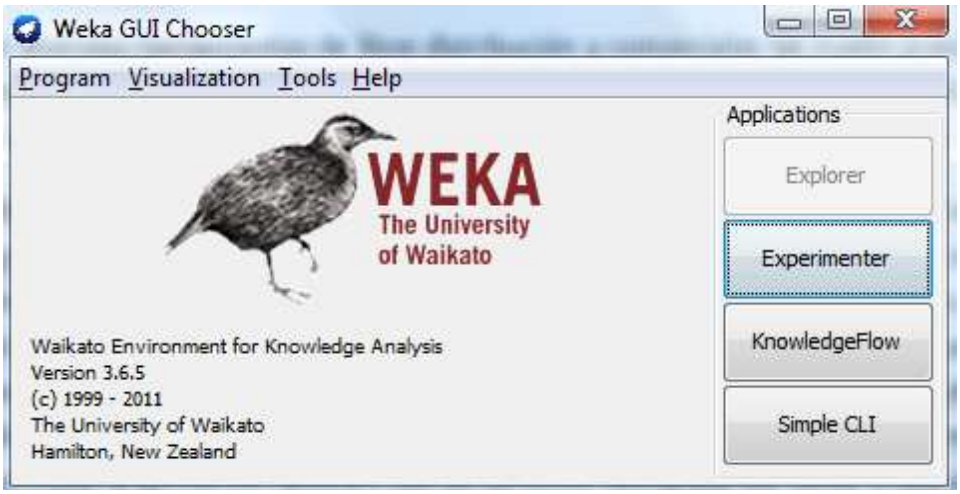
- **Datos de entrenamiento:**

Para obtener reglas de asociación no es necesario definir un conjunto de entrenamiento y otro de validación, dado que una regla no se valida con otra, se recuerda que estas se evalúan con las restricciones de soporte y confianza.

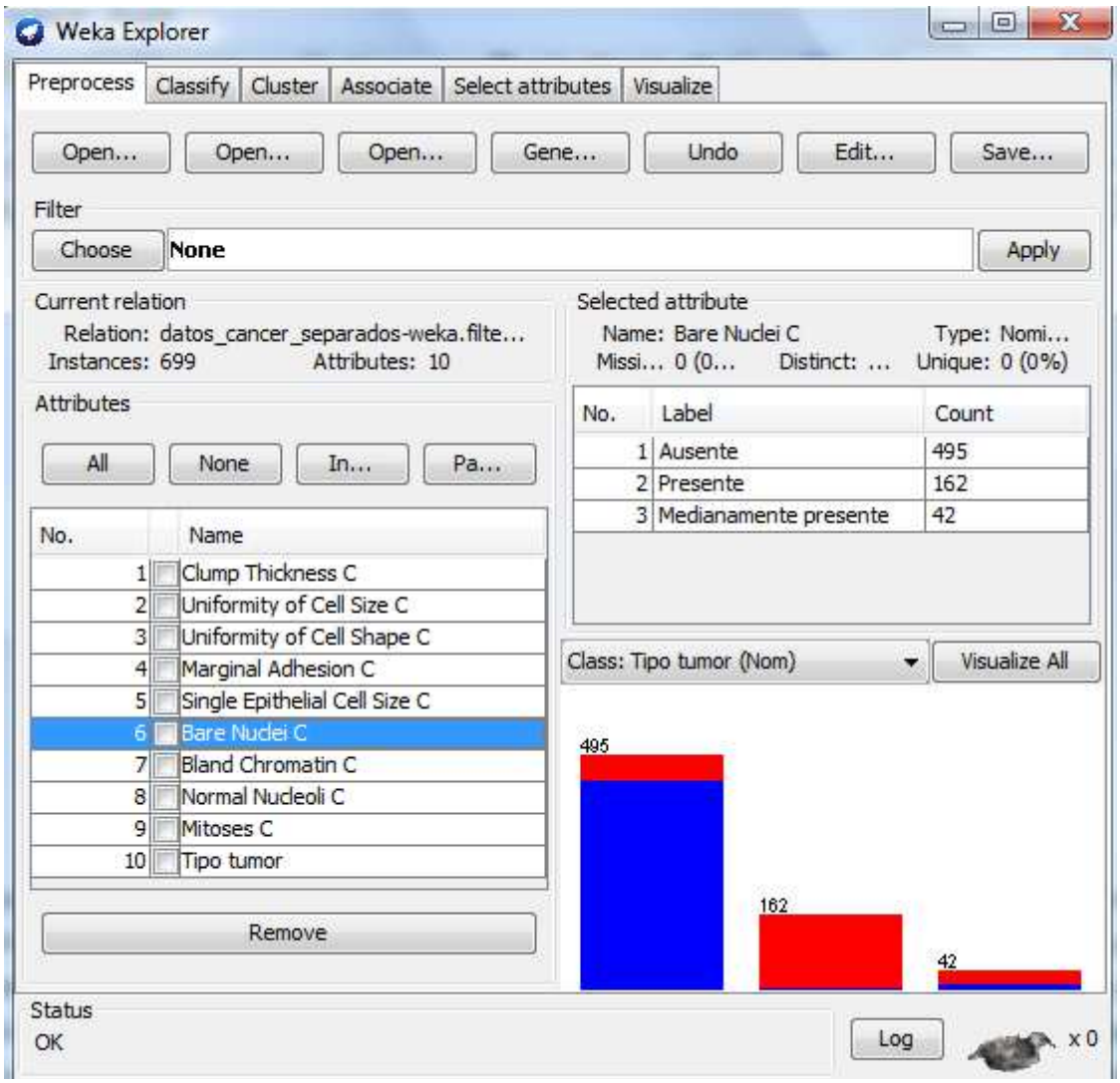
- **Selección de herramientas de aplicación**

Existen distintas herramientas de libre distribución y comerciales las cuales pueden ser empleadas para obtener reglas de asociación. Para esta investigación se analizan los software *Weka*, *Orange* y un paquete desarrollado para *R* (software estadístico).

Weka: este es un software de libre distribución que es ampliamente usado en investigaciones concernientes a procesos de minería de datos.

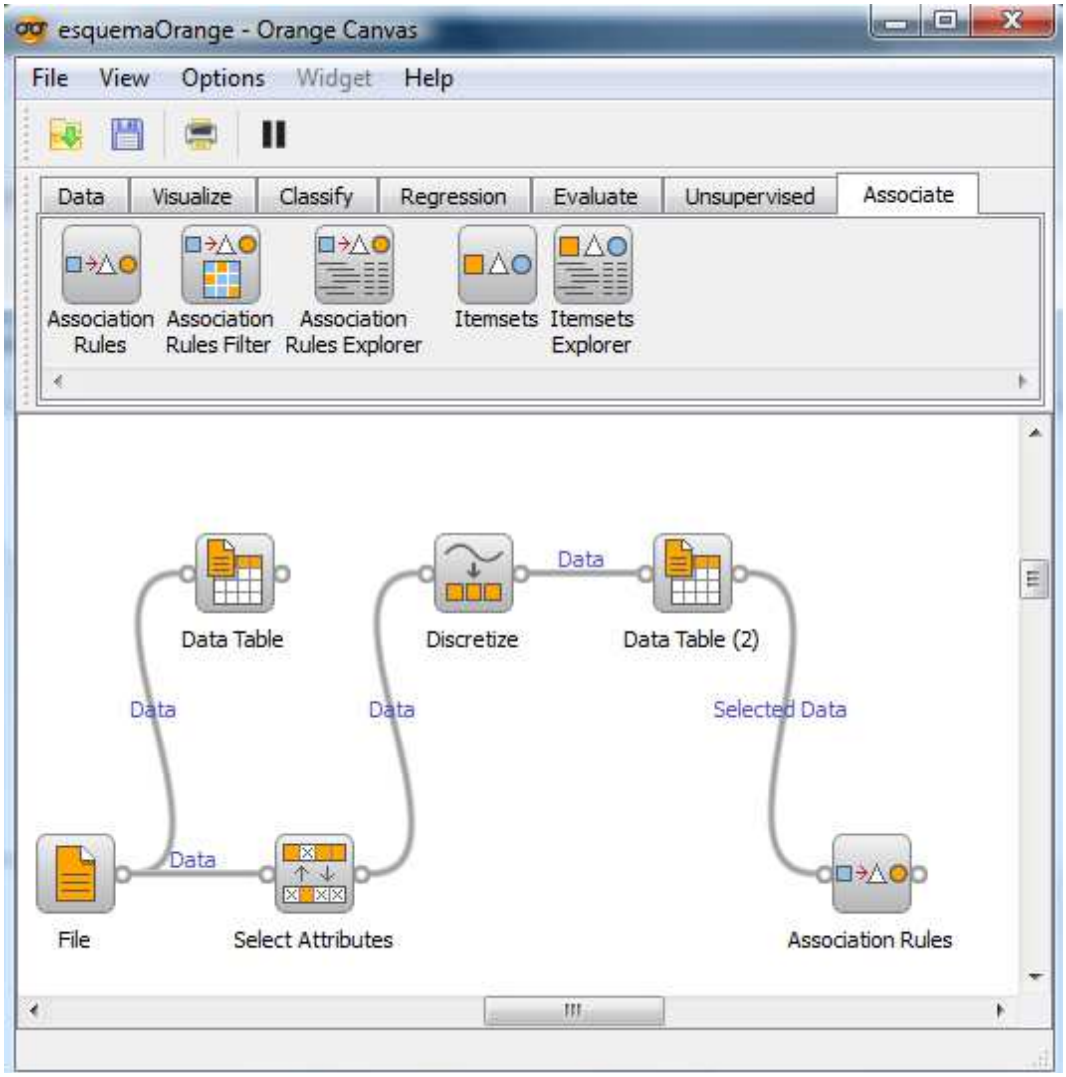


Permite obtener información básica de las variables de trabajo en un entorno sencillo e intuitivo, posee módulos para obtener modelos de asociación, clasificación y agrupamiento, dispone de técnicas para realizar transformaciones sobre los datos, identifica valores faltantes, tipos de variables y muestra histogramas según la variable seleccionada. Además de estas ventajas la aplicación también cuenta con un conjunto destacado de filtros u operaciones a ser aplicadas sobre los datos.



Respecto a las reglas de asociación, este ofrece un conjunto de algoritmos para obtener las reglas y la facilidad para la definición de los parámetros asociados a cada uno de los algoritmos.

Orange: software libre para realizar operaciones de minería de datos a través de un entorno gráfico y de programación en Python. Su entorno gráfico es práctico e intuitivo y permite manejar esta actividad como un flujo de actividades, un modelo secuencial de tareas que facilita la identificación de las actividades realizadas, como se muestra a continuación.

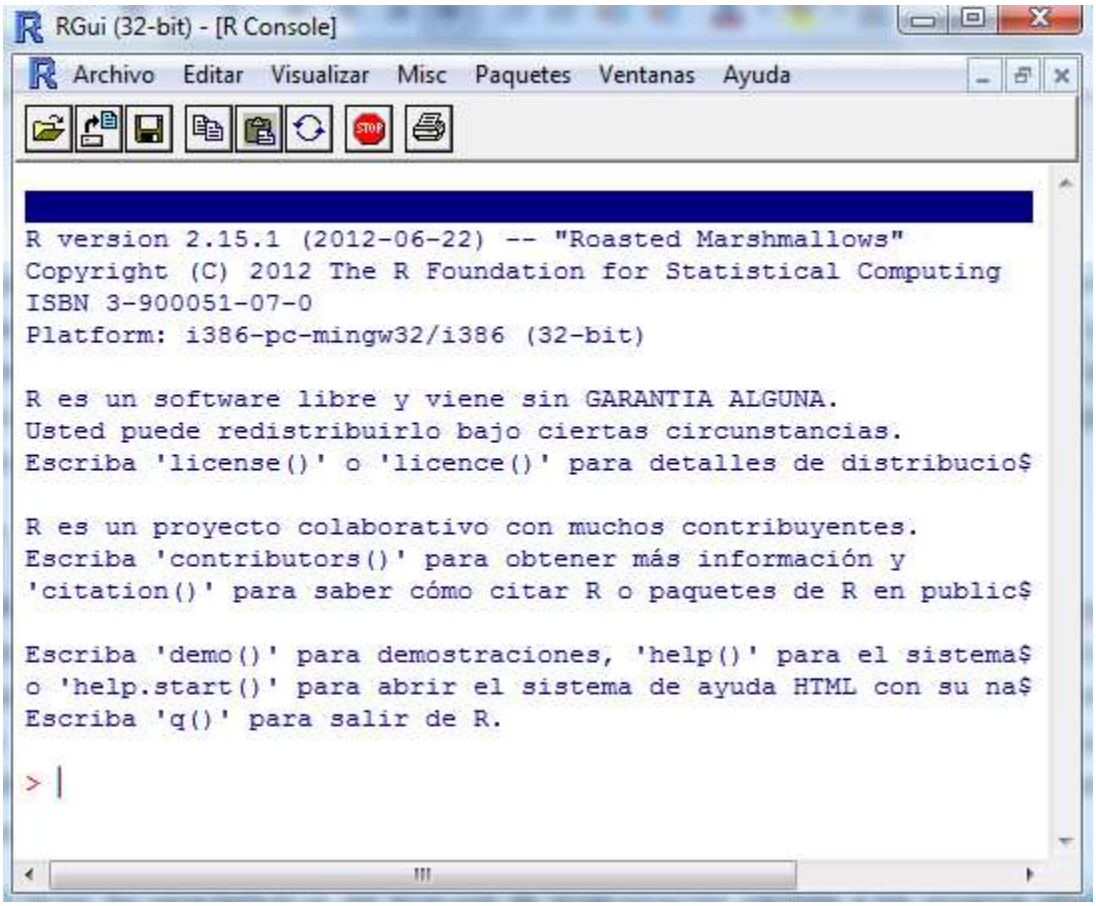


Dentro de las ventajas de este software se encuentra la opción de desarrollar la tarea a través de un diseño que permite manejar e identificar con mayor claridad el conjunto de actividades que son necesarias para obtener el modelo, permite realizar de manera sencilla y eficaz las tareas de transformaciones de datos, tiene la posibilidad para obtener modelos de asociación, regresión, clasificación y visualización, además de tener un módulo para evaluar los modelos obtenidos. Como desventajas se tiene que dentro del entorno gráfico existen algunos procesos que presentan deficiencias, como la discretización que aunque es muy gráfica y sencilla, no permite definir los nombre de las categorías que se identifican, este software no presenta la disposición de un conjunto de algoritmos que permitan obtener reglas de asociación, finalmente, la opción disponible de ejecutar procesos a través de programación Phyton es una ventaja sobre otros software, sin embargo el lenguaje empleado requeriría de un tiempo de aprendizaje y considerando que para algunos analista sólo resultaría útil en este caso, sería una ventaja poco considerable, dado que otras aplicaciones manejan todos los procesos necesario bajo un entorno gráfico suficientemente bueno.

R-Project:

Más que una aplicación R es un lenguaje y entorno para estadística computacional, este es un proyecto de libre distribución que provee una amplia variedad de técnicas estadísticas, además cuenta con un manejo efectivo de los datos, una larga, coherente e integrada colecciones de herramientas intermedias para análisis de datos y por último posee las características propias de un lenguaje de programación como condicionales, ciclos, parámetros de entradas y salidas, entre otros.

Al tener las características del lenguaje de programación permite a los usuarios añadir funcionalidad a través de la definición de nuevas funciones, convirtiéndolo en un recurso con grandes posibilidades para ser empleado en análisis de datos.

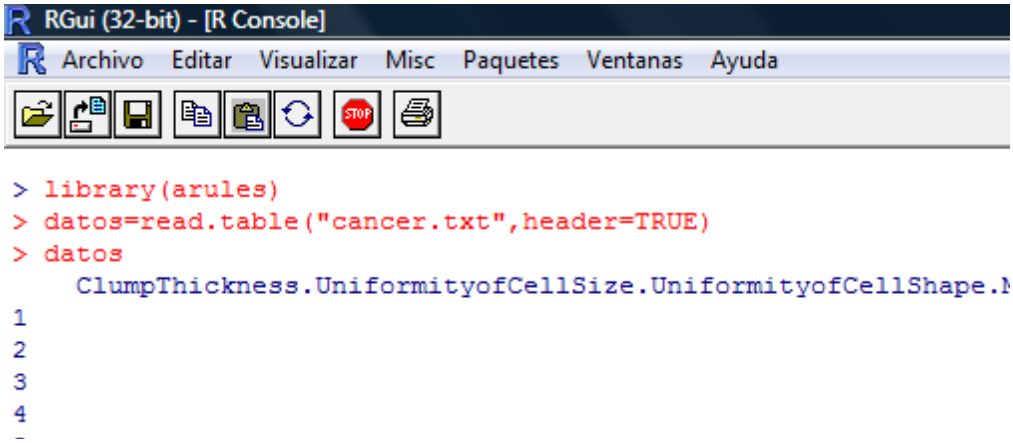


Este software tiene como ventaja la versatilidad para incorporar cualquier característica que se desee, su amplia gama de técnicas estadísticas disponible, su confiabilidad respaldada por la comunidad académica y científica. Como desventaja se puede considerar que este no cuenta con una opción propia para obtener reglas de asociación, por lo que es necesario cargar el paquete ***arules.***

Este paquete ofrece la posibilidad de obtener las reglas empleando el algoritmo Apriori o Eclat. Adicionalmente la ejecución del proceso de identificación de reglas no es

intuitiva y fácil, dado que es necesario ejecutar comandos que permiten cargar las librerías, para cargar los datos y para obtener las reglas, como se observa en la gráfica.

Básicamente la principal desventaja es la necesidad de ejecutar demasiados comandos para que las tareas sean ejecutadas.



Una vez identificadas las ventajas y desventajas de las aplicaciones estudiadas se concluye que *Weka* es la mejor opción para generar reglas de asociación, dado que ofrece la mayor cantidad de algoritmos para generar reglas, facilita las tareas de transformaciones sobre los datos más que las otras aplicaciones, todas las actividades son en un entorno gráfico intuitivo y básico, ofrece opciones de configuración de parámetros proporcionando mayor versatilidad para la generación de reglas.

- **Obtención de reglas**

El proceso de obtención de reglas se realizó en tres experimentos los cuales se caracterizan por trabajar con distintas variables, cada uno de ellos considerando una confianza de 0.9 y un soporte mínimo de 0.1.

Se consideran tres casos con el propósito de conocer cómo cambian las reglas de acuerdo a la cantidad de variables que se incluyen.

Primer experimento:

Para este experimento se tiene la siguiente información obtenida desde la aplicación

=== Run information ===

Relation: datos_cancer_separados-weka.filters.unsupervised.attribute.Remove-R1-11-weka.filters.unsupervised.attribute.Remove-R4-9

Instances: 699

Attributes: 4

Clump Thickness C

Uniformity of Cell Size C

Uniformity of Cell Shape C

Tipo tumor

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

Apriori

=======

Minimum support: 0.1 (70 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 16

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 2

En esta información se observan las cuatro variables de análisis, los valores de las medidas de confianza y soporte. Bajo esta configuración se obtiene un conjunto de cuarenta reglas de las cuales las más destacadas para nuestro estudio son:

9. uniformity of cell shape c=clase 3 93 ==> tipo tumor=maligno 92 conf:(0.99)

12. uniformity of cell size c=grande uniformity of cell shape c=clase 3 74 ==> tipo tumor=maligno 73 conf:(0.99)

14. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 355 ==> Tipo tumor=benigno 350 conf:(0.99)

15. Uniformity of Cell Size C=Grande 102 ==> Tipo tumor=maligno 100 conf:(0.98)

19. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena 360 ==> Tipo tumor=benigno 351 conf:(0.98)

20. Clump Thickness C=Delgada Uniformity of Cell Shape C=Clase 1 359 ==> Tipo tumor=benigno 350 conf:(0.97)

23. Clump Thickness C=Gruesa 129 ==> Tipo tumor=maligno 125 conf:(0.97)

25. Uniformity of Cell Size C=Mediana 76 ==> Tipo tumor=maligno 73 conf:(0.96)

34. Clump Thickness C=Delgada 383 ==> Tipo tumor=benigno 352 conf:(0.92)

36. Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 490 ==> Tipo tumor=benigno 449 conf:(0.92)

1. Uniformity of Cell Shape C=Clase 2 94 ==> Tipo tumor=maligno 86 conf:(0.91)

Las primeras variables seleccionadas se caracterizan por analizar el tamaño y forma del tumor, permitiendo realizar las siguientes conclusiones de acuerdo a los resultados arrojados:

- Si el tamaño del grupo de células del tumor es "*Grande"* y la forma del tumor pertenece a la "*Clase c"*, entonces el tumor puede ser maligno. (reglas 9, 12 y 15).

- Si el espesor de la masa es *"Delgada"* y la forma del tumor pertenece a la "*Clase 1*" entonces el tumor puede ser benigno. (reglas 34,14).

- Si el espesor de la masa es *"Gruesa"* puede que el tumor sea maligno. (reglas 23).

- Si el tamaño del grupo de células es más que "*Pequeño*" es decir "*Mediano*" o *"Grande"* y la forma del mismo grupo pertenece a la "*Clase 2*" o la *"Clase 3"* el tumor puede ser maligno.(reglas 37,25,15).

Segundo experimento

Este experimento cuenta con las siguientes características

=== Run information ===

Relation: datos_cancer_separados-weka.filters.unsupervised.attribute.Remove-R1-11-weka.filters.unsupervised.attribute.Remove-R5,8-9

Instances: 699

Attributes: 7

Clump Thickness C

Uniformity of Cell Size C

Uniformity of Cell Shape C

Marginal Adhesion C

Bare Nuclei C

Bland Chromatin C

Tipo tumor

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

Apriori

=======

Minimum support: 0.1 (70 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 44

Size of set of large itemsets L(3): 56

Size of set of large itemsets L(4): 55

Size of set of large itemsets L(5): 36

Size of set of large itemsets L(6): 13

Size of set of large itemsets L(7): 2

Para este experimento se obtiene un conjunto de 1371 reglas de las cuales se destacan las siguientes

20. Clump Thickness C=Delgada Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente 340 ==> Tipo tumor=benigno 340 conf:(1)

22. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente 340 ==> Tipo tumor=benigno 340 conf:(1)

25. Clump Thickness C=Delgada Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 336 ==> Tipo tumor=benigno 336 conf:(1)

28. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 336 ==> Tipo tumor=benigno 336 conf:(1)

31. Clump Thickness C=Delgada Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 335 ==> Tipo tumor=benigno 335 conf:(1)

33. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 335 ==> Tipo tumor=benigno 335 conf:(1)

47. Marginal Adhesion C=Fuerte Bare Nuclei C=Presente 87 ==> Tipo tumor=maligno 87 conf:(1)

48. Bare Nuclei C=Presente Bland Chromatin C=Medianamente uniforme 79 ==> Tipo tumor=maligno 79 conf:(1)

49. Clump Thickness C=Gruesa Bare Nuclei C=Presente 78 ==> Tipo tumor=maligno 78 conf:(1)

91. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 338 ==> Tipo tumor=benigno 337 conf:(1)

93. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 337 ==> Tipo tumor=benigno 336 conf:(1)

132. Clump Thickness C=Delgada Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bland Chromatin C=Uniforme 343 ==> Tipo tumor=benigno 341 conf:(0.99)

133. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bland Chromatin C=Uniforme 343 ==> Tipo tumor=benigno 341 conf:(0.99)

168. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil 352 ==> Tipo tumor=benigno 349 conf:(0.99)

224. Uniformity of Cell Shape C=Clase 3 93 ==> Tipo tumor=maligno 92 conf:(0.99)

233. Clump Thickness C=Delgada Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil 353 ==> Tipo tumor=benigno 349 conf:(0.99)

Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Marginal Adhesion C=Debil Bland Chromatin C=Uniforme 346 ==> Tipo tumor=benigno 342 conf:(0.99)

259. Uniformity of Cell Size C=Grande Uniformity of Cell Shape C=Clase 3 74 ==> Tipo tumor=maligno 73 conf:(0.99)

810. Uniformity of Cell Size C=Mediana 76 ==> Tipo tumor=maligno 73 conf:(0.96)

992. Clump Thickness C=Gruesa Marginal Adhesion C=Debil 79 ==> Tipo tumor=maligno 75 conf:(0.95)

1168. Marginal Adhesion C=Debil Bare Nuclei C=Presente 75 ==> Tipo tumor=maligno 70 conf:(0.93)

1288. Uniformity of Cell Shape C=Clase 2 94 ==> Tipo tumor=maligno 86 conf:(0.91)

De este conjunto de reglas se puede concluir que:

- Si el espesor de la masa es *"Delgada"*, la forma del tumor pertenece a la "*Clase 1*", la adhesión marginal entre las células es "*Debil*", la característica del *Bare Nuclei* está *"Ausente",* el tamaño del grupo de células es *"Pequeña"* y la textura del núcleo es *"Uniforme"* entonces el tumor puede ser benigno. (Regla 33).

- Si la adhesión marginal entre las células es *"Fuerte"* y la característica de *Bare Nuclei* está *"Presente"* entonces el tumor puede ser maligno. (Regla 47).

- Si la característica de *Bare Nuclei* está *"Presente"* y la textura del núcleo es "*medianamente uniforme*" entonces el tumor puede ser maligno. (Regla 48).

- Si la característica de *Bare Nuclei* está *"Presente"* y el espesor de la masa mamaria es "*Gruesa*" entonces el tumor puede ser maligno. (Regla 49).

- Si el tamaño del grupo de células es "*Grande*" y la forma de este grupo de células pertenece al grupo *"Clase 3"* entonces es posible que el tumor sea maligno. (Regla 259).

- Si la adhesión marginal de las células es "*Debil*" y el espesor de la masa mamaria es *"Grande"* entonces el tumor puede ser maligno. (Regla 992).

Tercer experimento:

Este experimento cuenta con todas las variables del conjunto de datos original, además de las siguientes características:

=== Run information ===

Relation: datos_cancer_separados-weka.filters.unsupervised.attribute.Remove-R1-11

Instances: 699

Attributes: 10

Clump Thickness C

Uniformity of Cell Size C

Uniformity of Cell Shape C

Marginal Adhesion C

Single Epithelial Cell Size C

Bare Nuclei C

Bland Chromatin C

Normal Nucleoli C

Mitoses C

Tipo tumor

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

Apriori

=======

Minimum support: 0.1 (70 instances)

Minimum metric <confidence>: 0.99

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 23

Size of set of large itemsets L(2): 80

Size of set of large itemsets L(3): 166

Size of set of large itemsets L(4): 294

Size of set of large itemsets L(5): 378

Size of set of large itemsets L(6): 336

Size of set of large itemsets L(7): 204

Size of set of large itemsets L(8): 81

Size of set of large itemsets L(9): 19

Size of set of large itemsets L(10): 2

Bajo estas condiciones se obtienen 6236 reglas de asociación de las cuales las siguientes son las más destacadas para el problema que aborda este estudio.

729. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente 340 ==> Tipo tumor=benigno 340 conf:(1)

883. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Single Epithelial Cell Size C=Pequena Bare Nuclei C=Ausente 337 ==> Tipo tumor=benigno 337 conf:(1)

Line 951: 901. Clump Thickness C=Delgada Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 336 ==> Tipo tumor=benigno 336 conf:(1)

922. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Marginal Adhesion C=Debil Single Epithelial Cell Size C=Pequena Bare Nuclei C=Ausente 336 ==> Tipo tumor=benigno 336 conf:(1)

975. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Single Epithelial Cell Size C=Pequena Bare Nuclei C=Ausente 336 ==> Tipo tumor=benigno 336 conf:(1)

1289. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Single Epithelial Cell Size C=Pequena Bare Nuclei C=Ausente Bland Chromatin C=Uniforme 333 ==> Tipo tumor=benigno 333 conf:(1)

1657. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Bare Nuclei C=Ausente Bland Chromatin C=Uniforme Normal Nucleoli C=Pequena Mitoses C=Mitosis 1 332 ==> Tipo tumor=benigno 332 conf:(1)

1903. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Single Epithelial Cell Size C=Pequena Bare Nuclei C=Ausente Bland Chromatin C=Uniforme Normal Nucleoli C=Pequena Mitoses C=Mitosis 1 329 ==> Tipo tumor=benigno 329 conf:(1)

2900. Marginal Adhesion C=Fuerte Bare Nuclei C=Presente 87 ==> Tipo tumor=maligno 87 conf:(1)

2901. Bare Nuclei C=Presente Bland Chromatin C=Medianamente uniforme 79 ==> Tipo tumor=maligno 79 conf:(1)

2902. Clump Thickness C=Gruesa Bare Nuclei C=Presente 78 ==> Tipo tumor=maligno 78 conf:(1)

3343. Clump Thickness C=Delgada Uniformity of Cell Size C=Pequena Uniformity of Cell Shape C=Clase 1 Marginal Adhesion C=Debil Single Epithelial Cell Size C=Pequena Bland Chromatin C=Uniforme Normal Nucleoli C=Pequena Mitoses C=Mitosis 1 336 ==> Tipo tumor=benigno 335 conf:(1)

Considerando que este experimento involucra todas las variables que hacen parte del conjunto de datos original, las reglas obtenidas contienen la mayor cantidad de información sobre las relaciones existentes entre estas y el diagnóstico del tipo de tumor.

Las conclusiones que se obtienes son:

- Si el espesor de la masa mamaria es "*Delgada*", el tamaño del grupo de células es *"Pequena",* la forma de estas células es de *"Clase 1",*el tamaño de las células epiteliales es *"Pequena",* la textura del núcleo es *"Uniforme"* y la característica de *Bare Nuclei* no está presente entonces el tipo de tumor puede ser benigno.(Regla 1289)

- Si la adhesión marginal es *"Fuerte"* y la característica de *Bare Nuclei* está presente entonces el tumor puede ser maligno.(Regla 2900).

- Si la característica de *Bare Nuclei* está presente y la textura del núcleo es *"Medianamente uniforme"* entonces el tumor puede ser maligno.(Regla 2901).

- Si la característica de *Bare Nuclei* está presente y el espesor de la masa mamaria es *"Gruesa"* entonces el tumor puede ser maligno.

Todas las conclusiones hasta ahora realizadas están sujetas a las interpretaciones realizadas sobre las medidas de evaluación bajo las cuales se hicieron las mediciones de las diferentes variables y de las cuales no se manejó información. Es por eso quizás que se observa en las reglas y conclusiones que la adhesión marginal débil es característica de los tumores benignos, hecho que contrasta con la definición realizada sobre esta variable, que indica que las células cancerígenas pierden la habilidad de permanecer agrupadas.

Regresión logística

A continuación se describe el proceso de obtención de un modelo de regresión logística que permita modelar las asociaciones existentes entre las variables del caso de estudio 1.

- **Transformaciones**

La transformación más destacada es la que garantiza el reemplazo de los valores faltantes que se encuentran presentes en la variable *Bare Nuclei,* para que la técnica trabaje correctamente.

Esta transformación se realizó previamente para el proceso de discretización que se observa en el anexo 2.

Adicional a esta transformación, también es necesario convertir los valores reales del diagnóstico (2: benigno - 4: maligno) a los típicos valores lógicos que manejan las aplicaciones de software (0: benigno – 1: benigno).

- **Selección de algoritmos**

Se desconoce de la existencia de un conjunto de algoritmos de los cuales se pueda seleccionar para obtener modelos de regresión logística, por tal razón esta actividad no se desarrolla para esta técnica.

- **Selección de modelos internos**

Esta actividad no tiene trascendencia para esta técnica dado que se conoce de antemano que se trabajará bajo el siguiente modelo de regresión logística múltiple:

$$ln\frac{P}{1-P} = \propto + \beta 1X1 + \beta 2X2 + \cdots + \beta kXk \quad [3.1]$$ (Silva & Barroso, 2004)

- **Datos de entrenamiento**

La técnica de regresión logística no necesita de datos de entrenamiento, esta no necesita ser entrenada para luego ser evaluada.

- **Selección de herramientas de aplicación**

Diferentes herramientas estadísticas permiten obtener un modelo de regresión logística, sin embargo muchas de ellas son de tipo comercial como ***StatGraphic*** o ***MiniTab***, otras en cambio poseen versiones de libre distribución para entornos académicos y que son bien conocidas y con gran respaldo en el entorno de trabajo como ***MatLab*** y el software ***R***.

Estas dos últimas herramientas a diferencia de las comerciales no ofrecen un entorno de trabajo amigable, por el contrario para su uso se requiere de habilidades de

programación, manejo de consola y de sentencias que representan funciones para poder ejecutar las tareas que se desean. Para este caso MatLab ofrece un menor esfuerzo para obtener el modelo de regresión logística y gran confiabilidad sobre los resultados obtenidos, por lo cual se decide emplear esta herramienta para obtener dicho modelo.

- **Obtención del modelo de regresión logística**

Para la obtención del modelo de regresión logística que permite determinar las asociaciones existentes para el caso de estudio 1, se emplea la función GLMFIT de MatLab que posee la siguiente sintaxis:

b = glmfit(X,y,distr)

b = glmfit(X,y,distr,param1,val1,param2,val2,...)

[b,dev] = glmfit(...)

[b,dev,stats] = glmfit(...)

Para este trabajo se emplea la tercera sentencia, donde los parámetros de la función están definidos así:

X: matriz de tamaño *n*x*p* donde *n* es la cantidad de registros que posee el dataset (699 para el dataset Breast Cancer Wisconsin) y *p* es la cantidad de variables predictoras (9 para el mismo dataset). Esta matriz contiene todas las mediciones para las distintas variables de interés tomadas sobre 699 individuos.

y: vector de la forma *n*x*1* que contiene los resultados observados, donde *n* es la cantidad de registros que posee el dataset (699 para el dataset Breast Cancer Wisconsin). En este caso contiene los valores dicotómicos que indican si el tumor analizado es maligno o benigno.

b: vector de la forma *n*x*1,* donde *n* es la cantidad de coeficientes retornados por la función, más un coeficiente adicional. Nueve de estos coeficientes se asocian a las nueve variables predictoras y el otro se refiere al valor α de la ecuación 3.1.

Stats: contiene información estadística sobre los resultados del modelo de regresión obtenido, que permite determinar la fiabilidad de modelo.

Una vez se ha aplicado la función se obtienen los coeficientes de la ecuación 3.1, de tal forma que el modelo de regresión queda de la forma:

$$\ln (P / 1- P) = -9.7145 + (0.5346)X1 + (0.0113)X2 + (0.3238)X3 + (0.2376)X4 + (0.0583)X5 + (0.4282)X6 + (0.4121)X7 + (0.1582)X8 + (0.5358)X9 \quad [3.2]$$

Donde,

X1: Clump Thickness

X2: Uniformity of Cell Size

X3: Uniformity of Cell Shape

X4: Marginal Adhesion

X5: Single Epithelial Cell Size

X6: Bare Nuclei

X7: Bland Chromatin

X8: Normal Nucleoli

X9: Mitoses

Una vez obtenido el modelo es necesario proceder a evaluarlo, corroborar que este representa adecuadamente el proceso que se estudia, para lo cual se procede evaluando la significancia de cada variable a través de la prueba de Wald, para luego realizar una prueba de bondad de ajuste que permita conocer la calidad del ajuste.

Prueba de Wald: su objetivo fundamental es constatar si un coeficiente β es significativamente diferente de cero, para esto se divide la estimación del coeficiente de interés (b) por su error estándar (se (b)). (Silva & Barroso, 2004)

$$Zwald = \frac{b}{se(b)} \quad [3.3]$$

Prueba de bondad de ajuste: esta prueba permite evaluar la calidad del ajuste de la regresión empleando los siguientes pasos (Silva & Barroso, 2004):

- Calcular $P_1, P_2,...,P_n$ a partir del modelo ajustado, es decir los 699 valores P calculados a partir de la ecuación 3.2.
- Ordenar los *n* valores de menor a mayor.
- Segregar los *n* valores en grupos considerando dos métodos:

Método A: dividir dicha secuencia ordenada en cuartiles, deciles u otra separación

Método B: Se forman grupos a partir de rangos de valores. El primer grupo con todas las observaciones cuyos valores están entre 0.1 y 0.2, así de manera consecutiva.

- Se determina la cantidad de registros sobre cada grupo, la frecuencia observada de valores donde Y=1, es decir, los casos donde se presenta un tumor maligno, además de la frecuencia esperada de cada grupo.

- Luego se comparan los valores de las frecuencias observadas contra los valores de las frecuencias esperadas, donde se espera que los valores sean parecidos para determinar que se tiene un buen ajuste de bondad.

4.3 Etapa de evaluación

Reglas de asociación

Para el caso de las reglas de asociación no existe una medida de evaluación posterior a la obtención de las reglas para determinar si son o no adecuadas. La confiabilidad de las reglas obtenidas se encuentra soportada por las medidas de soporte y confianza definidas justo antes de la ejecución de cada proceso. Una evaluación posterior podría darse a través de un experto que determine la coherencia y veracidad de las reglas y conclusiones.

Bajo estas consideraciones se tiene que las reglas de asociación obtenidas son confiables dado los datos bajo los cuales fueron ejecutadas.

Regresión logística

Como se mencionó en la fase de aplicación de esta técnica, para validar los resultados del modelo y poder realizar las correspondientes conclusiones es necesario realizar una prueba de significancia de las variables y una prueba de bondad de ajuste como se muestra a continuación:

- **Prueba de Wald**

Para esta prueba se plantea la hipótesis nula

$H0: \beta 1 = \beta 2 = \cdots = \beta n = 0;$ donde n=9 que corresponde a la cantidad de variables analizadas para el estudio. Esta sugiere que ninguna de las variables involucradas en el estudio tiene relevancia en el tipo de tumor que se obtiene.

y la hipótesis alternativa

$H1: \beta 1 \neq 0, \beta 2 \neq 0, \ldots, \beta n \neq 0$; donde n=9 corresponde a la cantidad de variables analizadas para el estudio, la cual indica que todas las variables son relevantes para determinar el tipo de tumor que se estudia, ya sea maligno o benigno.

Para rechazar o no la hipótesis nula se compara el estadístico Wald bajo un nivel de significancia del 95%.

A continuación se muestra la información de este estadístico.

Tabla 7: Valores del estadístico Wald para las variables de estudio de cáncer de mama

	Coeficiente (b)	**Error estándar (se)**	**Zwald**
ClumpThickness	0.5346	0.1349	3,96293551
Uniformity of CellSize	0.0113	0.1932	0,05848861
Uniformity of CellShape	0.3238	0.2139	1,51379149
Marginal Adhesion	0.2376	0.1167	2,03598972
Single EpithelialCellSize	0.0583	0.1526	0,38204456
BareNuclei	0.4282	0.0905	4,73149171
BlandChromatin	0.4121	0.1569	2,6265137
Normal Nucleoli	0.1582	0.1041	1,5196926
Mitoses	0.5358	0.3070	1,74527687

Dado que el nivel de significancia para este estudio se fijó en un 95%, se tiene que el valor crítico o valor de comparación es de 1.96. Teniendo en cuenta este valor se concluye que sólo para las variables *Uniformity of Cell Size, Uniformity of Cell Shape, Mitosis, Normal Nucleoli* y *Single Epithelial Cell Size,* no se rechaza la hipótesis nula, dando a entender que estas variables no tienen relevancia a la hora de identificar si el tipo de tumor de mama que se estudia es benigno o maligno, las otras cuatro variables son influyentes en el resultado del diagnóstico.

- **Prueba de bondad de ajuste**

Paso 1:

Para el cálculo de los valores *P* se considera la ecuación

$$P(Y = 1) = \frac{exp\ (x)}{1+exp\ (x)} \quad [3.4]$$

La cual se obtiene despejando el valor *P* de la ecuación [3.1] y *x* corresponde a la combinación que se observa en la ecuación [3.2].

Paso 2:

Los grupos de definen de la siguiente manera:

Primer grupo: cuyos valores se encuentren por debajo de 0.26

Segundo grupo: cuyos valores se encuentren en el rango de 0.26 y 0.50

Tercer grupo: cuyos valores se encuentren en el rango de 0.51 y 0.75

Cuarto grupo: cuyos valores se encuentren por encima de 0.75

Paso 3:

A continuación se ilustra la información de las frecuencias de registros asociadas a cada grupo.

Tabla 8: Frecuencias observadas y esperadas para los grupos definidos

	0.00 – 0.25	**0.26 – 0.50**	**0.51 - 0.75**	**0.76 – 1.00**
Cantidad de registros	445	11	8	235
Frecuencias observadas	2	8	4	227
Frecuencias esperadas	7.29	3.98	5,17	234.8031

La información que se observa en la tabla anterior indica que después de realizar los cálculos con base al modelo probabilístico de regresión logística [3.2] se tiene que:

- De los 699 registros analizados, 445 tienen una probabilidad entre 0.01 y 0.25 de ser malignos.

- Existen 11 casos analizados que tienen un rango probabilístico entre 0.26 y 0.50 de ser malignos.

- Con una probabilidad entre 0.51 y 0.75 se tienen 8 casos donde es posible que el tumor sea maligno.

- Existen 235 casos cuya probabilidad de que el tumor sea maligno se encuentra entre 0.76 y 1

- De los casos de estudios pertenecientes al primer grupo solo 2 se reportan como malignos.

- Para el segundo y tercer grupo se reportan 8 y 4 casos como malignos, respectivamente.

- Para el cuarto grupo se reportan 227 casos como malignos.

- El modelo de regresión estima que para el primer grupo se tienen 7.29 casos cuyo tipo de tumor es maligno.

- Se estima que se presentan 3.98 y 5.17 casos de tumores malignos dentro del segundo y tercer grupo respectivamente.

- El modelo estima que existen 234.80 casos de tumores malignos para el cuarto grupo.

Paso 4:

Se observa ahora que los valores observados para cada grupo y los estimados por el modelo no son muy distintos. Bajo estas observaciones se puede concluir que el modelo de regresión logístico obtenido ([3.2]) es bueno para modelar las asociaciones entre las variables que se estudian y el tipo de tumor, ya sea maligno o benigno.

Finalmente, dado que los dos modelos permiten modelar las asociaciones entre variables, es posible emplear los dos para nuestros objetivos, sin embargo, el modelo de regresión logística permite identificar la magnitud y el sentido de la relación a diferencia de las reglas de asociación. Se observa que el modelo matemático de la regresión ofrece mayor precisión y mayor información a través de las probabilidades obtenidas, por tal razón, la técnica seleccionada para el establecimiento de asociaciones concernientes al primer caso de estudio es la regresión logística.

El caso de estudio permite observar la forma paso a paso como se aborda el problema, considerando los aspectos más relevantes y destacados que son necesarios considerar para poder seleccionar y evaluar las técnicas de minería de datos enfocadas en el establecimiento de asociaciones.

5. Conclusiones

- La guía es un punto de partida para abordar problemas de selección y evaluación de técnicas de minería de datos enfocadas en el establecimiento de asociaciones, indicando determinadas actividades a ejecutar.
- La guía no es una herramienta completamente definida, se trata de un marco de trabajo que puede ser complementado con el análisis de otras técnicas para modelar asociaciones que no se consideraron en esta investigación.
- La guía muestra ser una herramienta útil, práctica, fácil de entender y aplicar que será de gran valor a mineros de datos, en cuanto al desarrollo de sus actividades cotidianas.
- La falta de información y metadatos del conjunto de datos de trabajo, dificulta el desarrollo de las actividades definidas en las etapas de la guía y por ende dificulta el proceso de selección y evaluación de técnicas de minería de datos.
- Algunos modelos de asociación tienen la característica de ser tanto descriptivos como predictivos, tal es el caso de modelos obtenidos a partir de la regresión lineal múltiple, mientras otros solo son descriptivos.
- La técnica de cópulas no define una relación entre variables sino el efecto u ocurrencia conjunta de eventos. Esta se enfoca en establecer funciones de probabilidad conjunta, mas no en determinar asociaciones entre variables.
- Las actividades de formalización de procesos resultan importantes para fomentar y consolidar el desarrollo de los mismos, por lo cual el desarrollo de guías como la propuesta en esta investigación, deben continuar, enfocadas en otras etapas del proceso de inteligencia de negocio y de minería de datos.
- Las herramientas de libre distribución para trabajar temas de minería de datos aún no se terminan de desarrollar, estas aun no proporcionan un conjunto completo de funcionalidades que permitan ejecutar todas las actividades satisfactoriamente. Es frecuente tener que emplear más de dos aplicaciones para ejecutar procesos que son complementarios.

Referencias Bibliográficas

Al-Harthy, M., Begg, S., Reidar, B. 2007. Copulas: A new technique to model dependence in petroleum decision making. Journal of Petroleum Science and Engineering, 57, 195 – 208.

Amon, I. 2010. Guía metodológica para la selección de técnicas de depuración de datos. Tesis de Maestría en Ingeniería de Sistemas. Universidad Nacional de Colombia, Medellín.

Baesens, B., Mues, C., Martens, D., Vanthienen, J. 2009. 50 years of data mining and OR: upcoming trends and challenges. Journal of the Operational Research Society, 60, s16 – s23.

Chen, W.s., Du, Y.K. 2009. Using neural networks and data mining techniques for the financial distress prediction model. Expert Systems with Applications, 36, 4075–4086.

Delene, D., Fuller, C., McCann, C., Ray, D. 2009. Analysis of healthcare coverage: A data mining approach. Expert Systems with Applications, 36, 995–1003.

Enke, D., Thawornwong, S. 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Systems with Applications, 29, 927-940.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine.

Frees, E., Valdez, E. 1997. Understanding Relationships using Copulas. North American Actuarial Journal, 2.

Garcia, E., Romero, C., Ventura, S., De Castro, C. 2011. A collaborative educational association rule mining tool. Internet and Higher Education, 14, 77-88.

Hung, S.H., Yen, D.C., Wang, H.Y. 2006. Applying data mining to telecom churn management. Expert Systems with Applications, 31, 515–524.

I-Cheng, Y., Che-hui, L. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36, 2473-2480.

Kim, Y. 2008. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. Expert Systems with Applications, 34, 1227–1234.

Kuo, R.J., Shih, C.W. 2007. Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan. Computers and Mathematics with Applications, 54, 1303–1318.

Kurt, I., Ture, M., Turhan Kurum, A. 2008. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications, 34, 366-374.

Lahiri, R. 2006. Comparisong of Data Mining and Statistical Techniques for Classification Model. Tesis de Maestría en Ciencias. Jadavpur University, India.

Liao, S.H., Chen, Y.N., Tseng, Y.Y. 2009. Mining demand chain knowledge of life insurance market for new product development. Expert Systems with Applications, 36, 9422–9437.

Meyfroidt, G., Güiza, F., Ramon, J., Bruynooghe, M. 2009. Machine learning techniques to examine large patient databases. Best Practice & Research Clinical Anaesthesiology, 23, 127-143.

Mitra, S., Pal, S.K., Mitra, P. 2002. Data Mining in Soft Computing Framework: A Survey. IEEE Transactions on Neural Networks, 13.

Montgomery, D., Peck, E., Vining G. 2002. Introducción al Análisis de Regresión Lineal. 3ra Edición. Programas Educativos S.A, Asturias, Mexico.

Moreno, M., Ramos, I., Garcia, J., Toro, M. 2008. An association rule mining method for estimating the impact of project management policies on software quality, development time and effort. . Expert Systems with Applications, 34, 522–529.

Nelsen, R. 1999. An Introduction to Copulas. Springer-Verlag, New York.

Ngai, E.W.T., Xiu, L., Chau, D.C.K. 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36, 2592-2602.

Ozisikyilmaz, B., Memik, G., Choudhary, A. 2008. Machine Learning Models to Predict Performance of Computer System Design Alternatives. En IEEE: 37th International Conference on Parallel Processing.

Qodmanan, H., Nasiri, M., Minaei-Bidgoli, B. 2011. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Systems with Applications, 38, 288-298.

Scott, P., Wilkins, E. 1999. Evaluating data mining procedures: techniques for generating artificial data sets. Information and Software Technology, 41, 579–587.

Serna, S. 2009. Comparación de Árboles de Regresión y Clasificación y regresión logística. Tesis de Maestría en Estadística. Universidad Nacional de Colombia, Medellín.

Silva, L., Barroso, I., 2004. Cuadernos de Estadística. Regresión Logística. La Muralla S.A, Madrid, España.

Spott, M., Nauck, D. 2006. Towards the automation of intelligent data analysis. Applied Soft Computing, 6, 348-356.

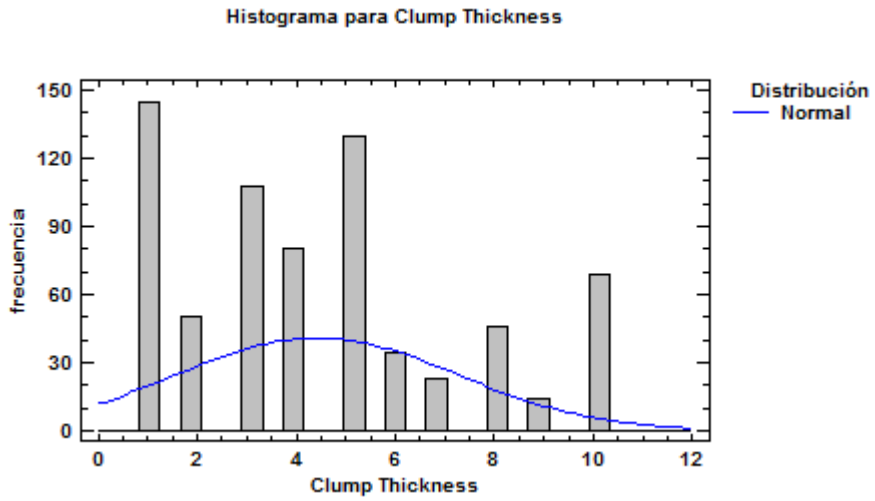
Viscarra, R., Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158, 46-54.

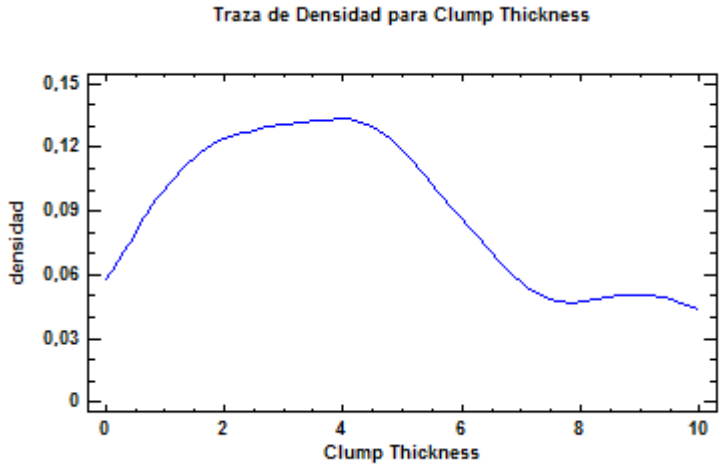
Wang, Y.F., Chuang, Y.L., Hsu, M.H., Keh, H.C. 2004. A personalized recommender system for the cosmetic business. Expert Systems with Applications, 26, 427–434.

Weaver, D. 2008. Applying data mining techniques to library design, lead generation and lead optimization. Current Opinion in Chemical Biology, 8,264–270.

Anexo 1: Histogramas de frecuencias para las variables de Breast Cancer Wisconsin (Diagnostic) Data Set

- Clump Thickness





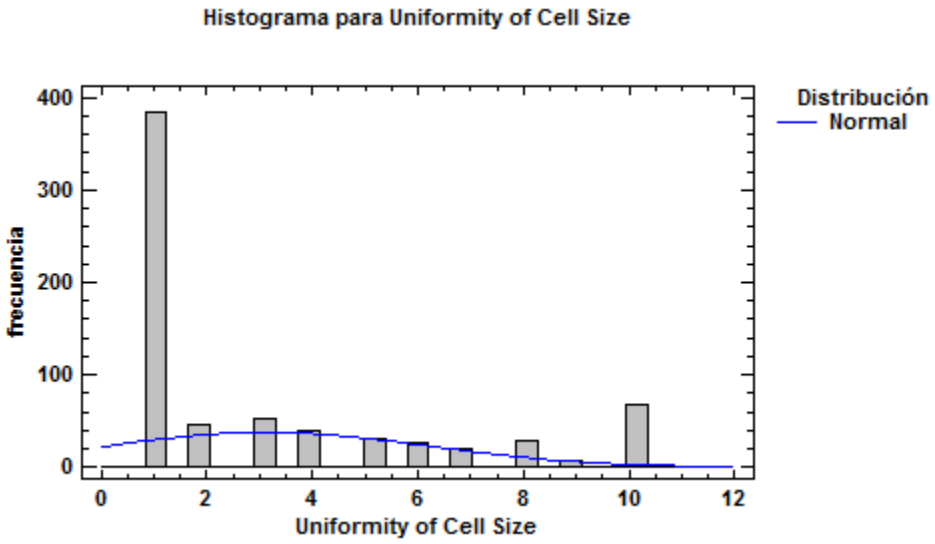
Pruebas de Bondad-de-Ajuste para Clump Thickness

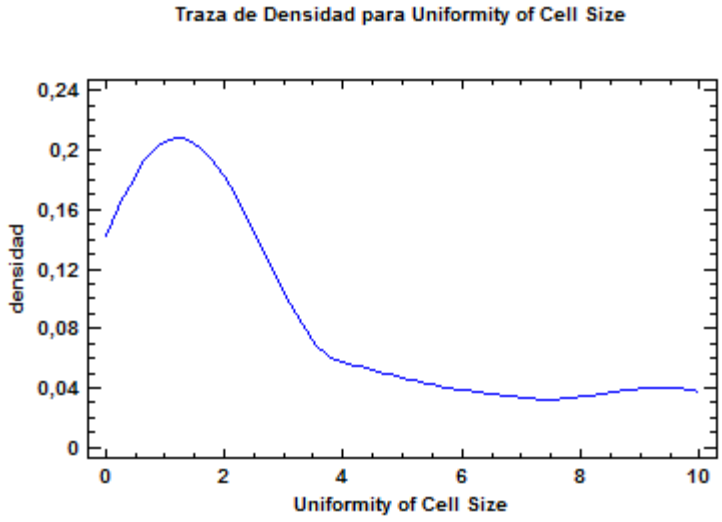
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,151991
DMENOS	0,112412
DN	0,151991
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Clump Thickness proviene de una distribución normal con 95% de confianza.

- Uniformity of Cell Size





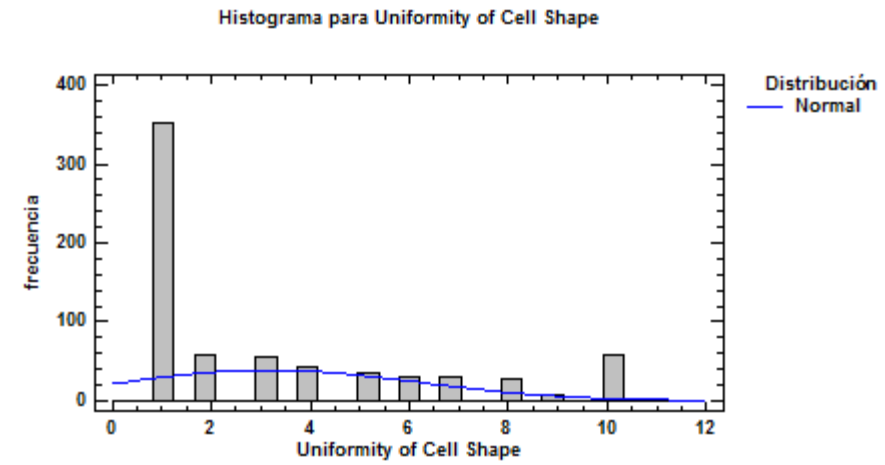
Pruebas de Bondad-de-Ajuste para Uniformity of Cell Size

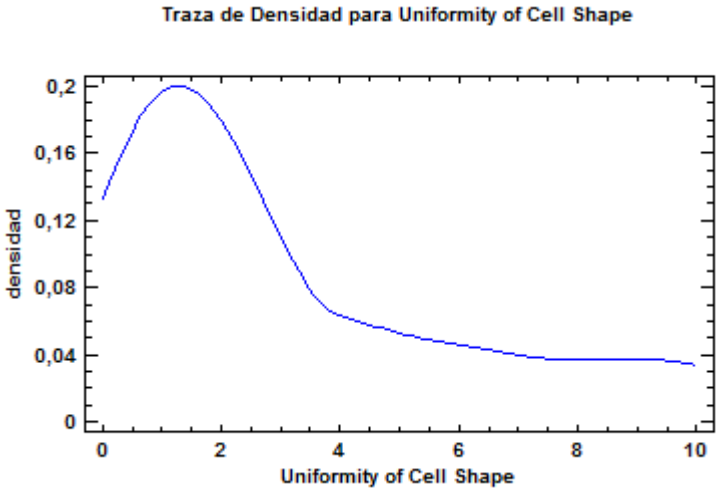
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,307236
DMENOS	0,24212
DN	0,307236
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Uniformity of Cell Size proviene de una distribución normal con 95% de confianza.

- Uniformity of Cell Shape





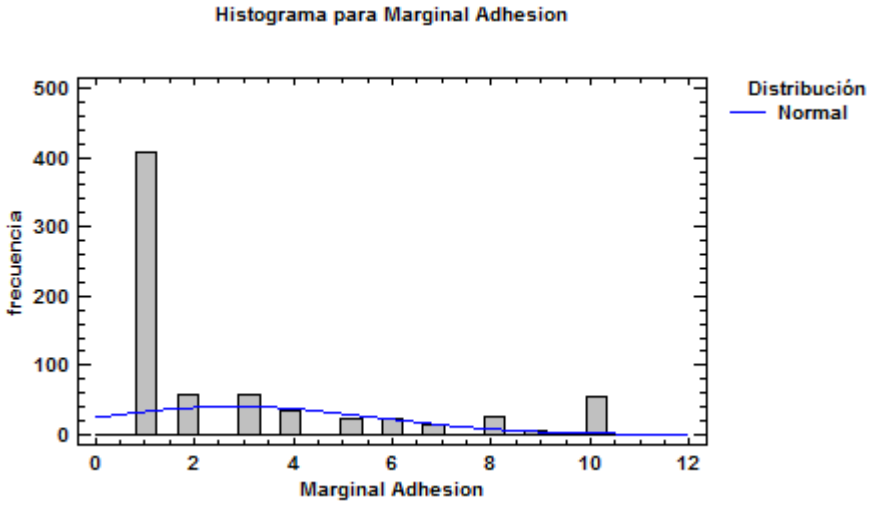
Pruebas de Bondad-de-Ajuste para Uniformity of Cell Shape

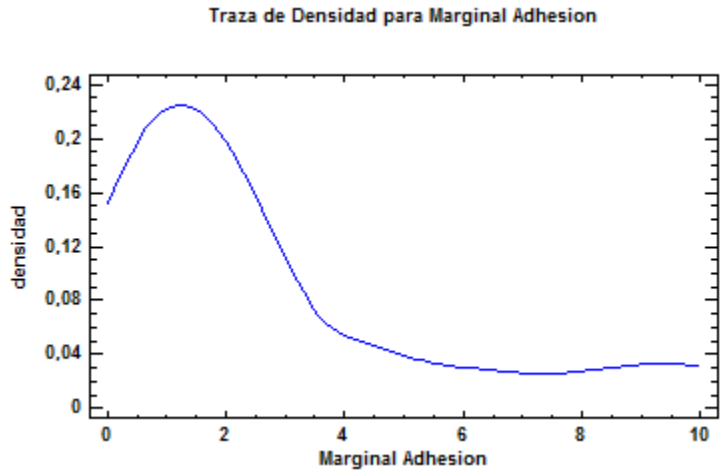
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,276197
DMENOS	0,22881
DN	0,276197
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Uniformity of Cell Shape proviene de una distribución normal con 95% de confianza.

- Marginal Adhesion





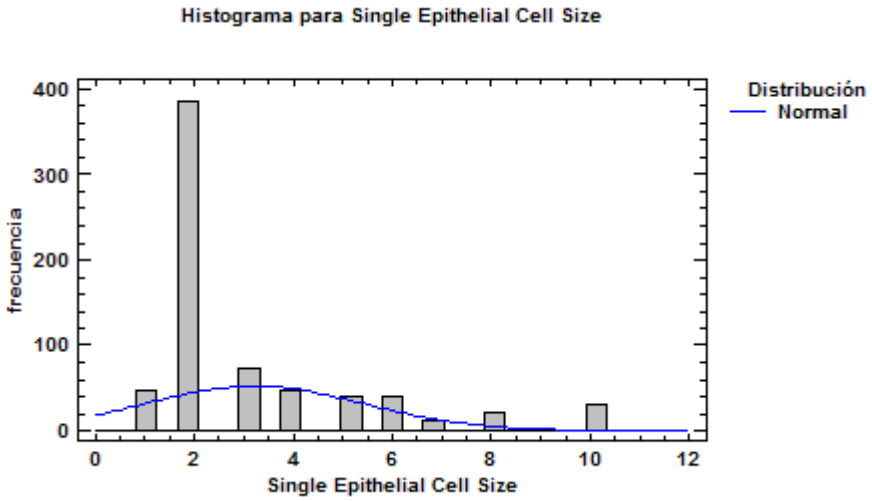
Pruebas de Bondad-de-Ajuste para Marginal Adhesion

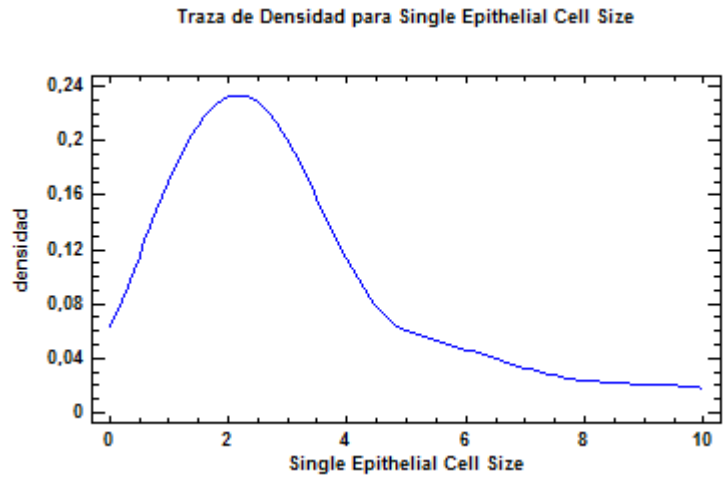
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,318828
DMENOS	0,263433
DN	0,318828
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Marginal Adhesion proviene de una distribución normal con 95% de confianza.

- Single Epithelial Cell Size





Pruebas de Bondad-de-Ajuste para Single Epithelial Cell Size

Prueba de Kolmogorov-Smirnov

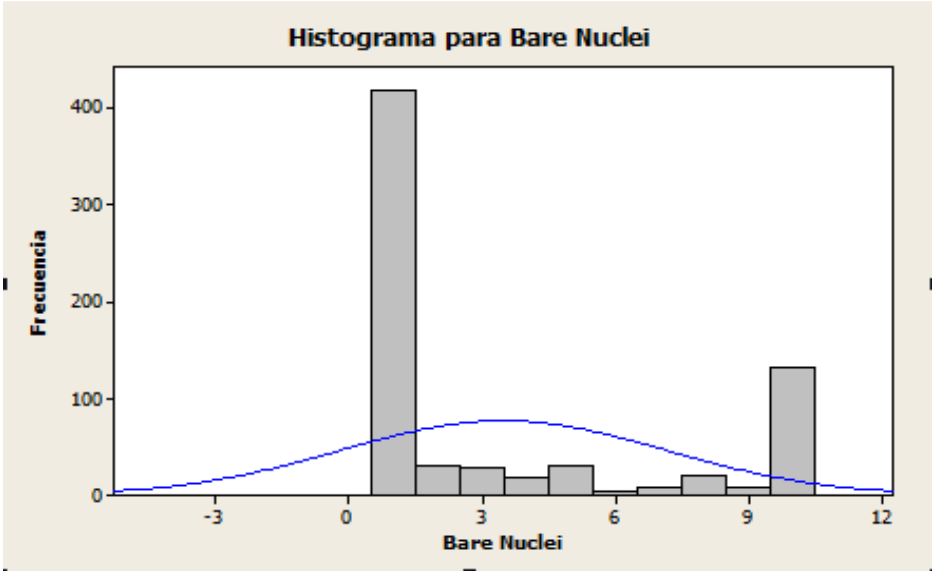
	Normal
DMAS	0,328013
DMENOS	0,224205
DN	0,328013
Valor-P	0,0

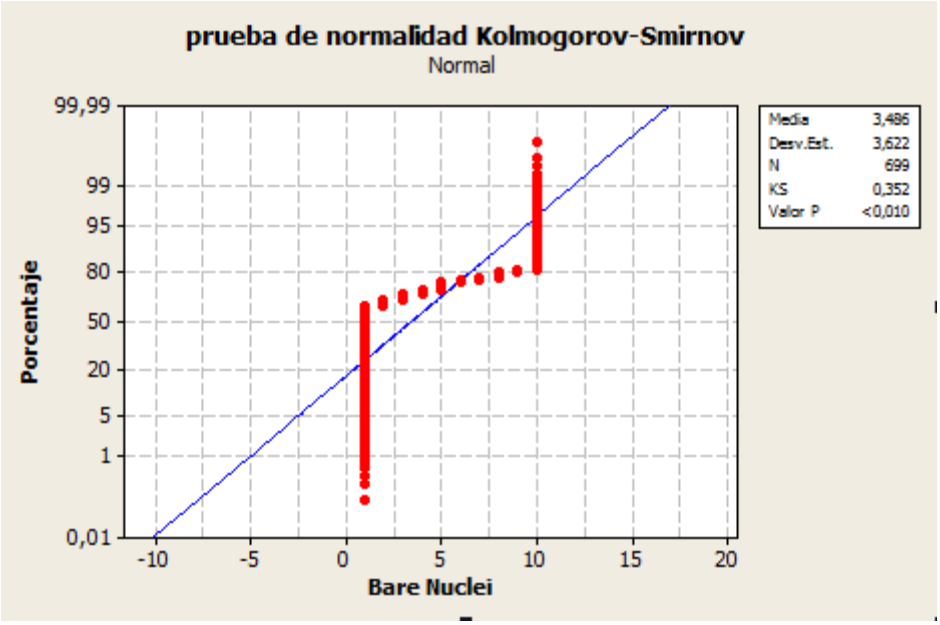
Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Single Epithelial Cell Size proviene de una distribución normal con 95% de confianza.

- Bare Nuclei

Esta variable tiene la particularidad de poseer valores faltantes que deben ser manejados antes de iniciar con cualquier tipo de cálculo o gestión de información asociada a la misma, como lo es la actividad donde se evalúa el tipo de distribución de los datos. Por tal razón, se consideran algunas de las técnicas existentes en minería de datos para manejar los valores faltantes (Amon, 2010) y se decide proceder por aquella que indica que es posible reemplazar el valor faltante por el estadístico Moda de esta variable, esto con el propósito de no desprestigiar la información de los distintos registros afectados.

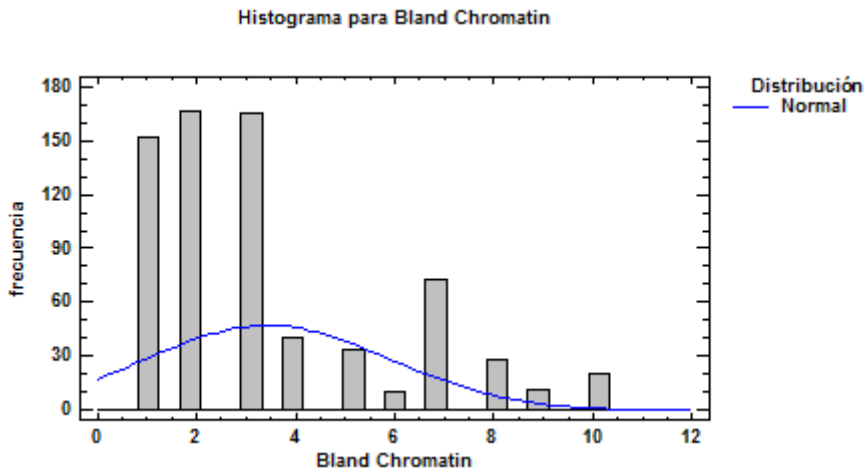
Una vez realizados estos cambios se puede proceder a identificar el tipo de distribución de la variable e cuestión.

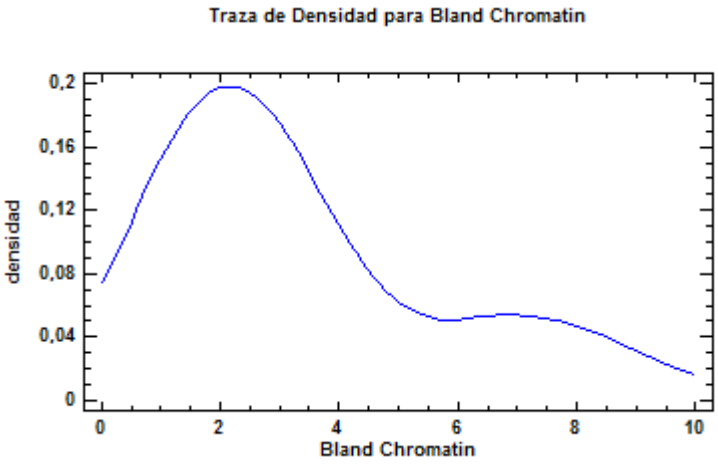




La prueba de normalidad de Kolmogorov-Smirnov devuelve un valor –p y debido a que este es menor a 0,05, se puede rechazar la idea de que Bare Nuclei proviene de una distribución normal con 95% de confianza.

- Bland Chromatin





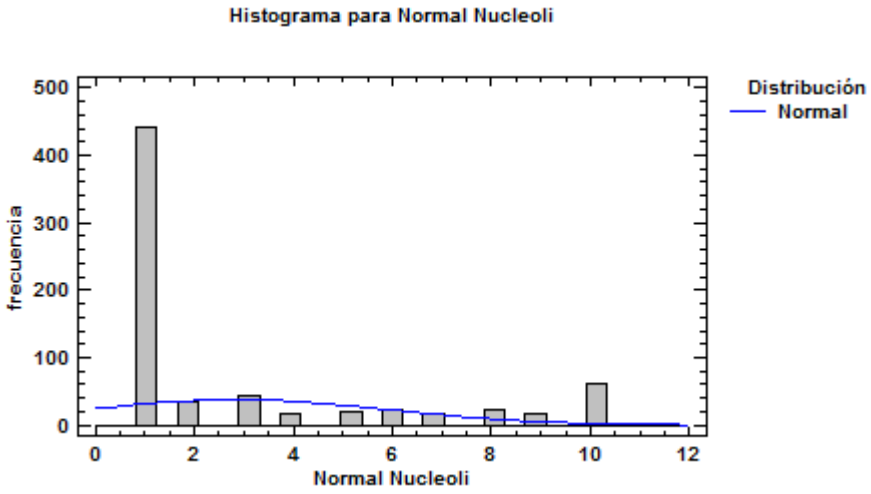
Pruebas de Bondad-de-Ajuste para Bland Chromatin

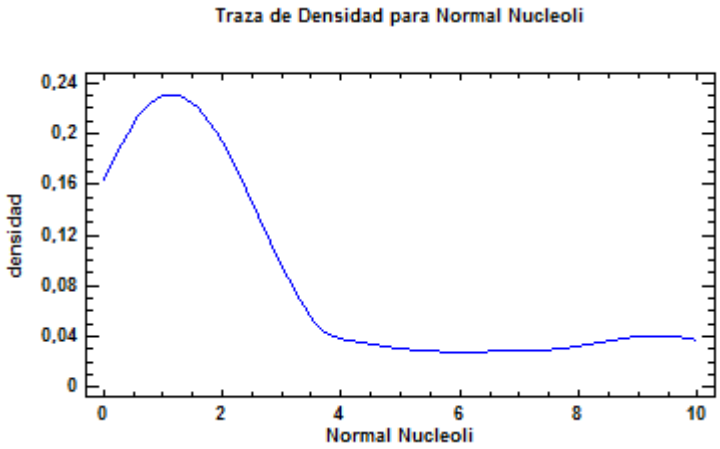
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,26223
DMENOS	0,158714
DN	0,26223
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Bland Chromatin proviene de una distribución normal con 95% de confianza.

- Normal Nucleoli





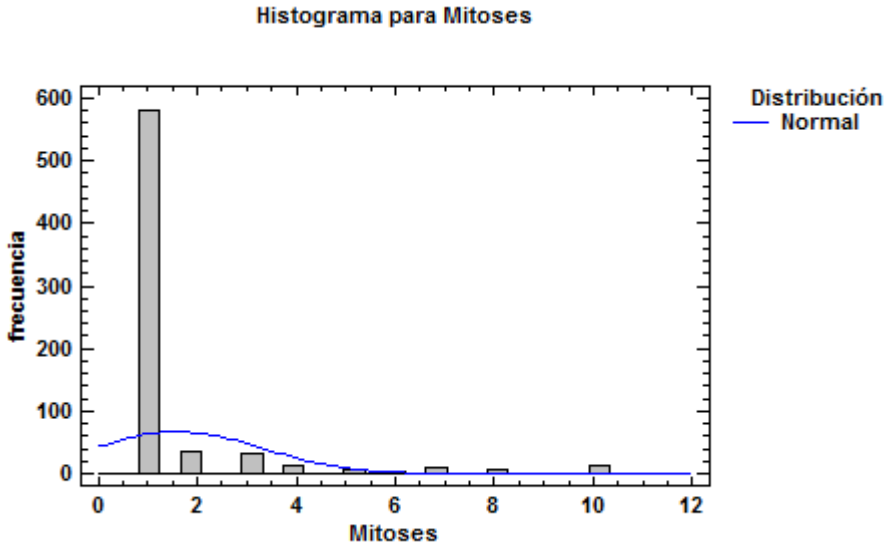
Pruebas de Bondad-de-Ajuste para Normal Nucleoli

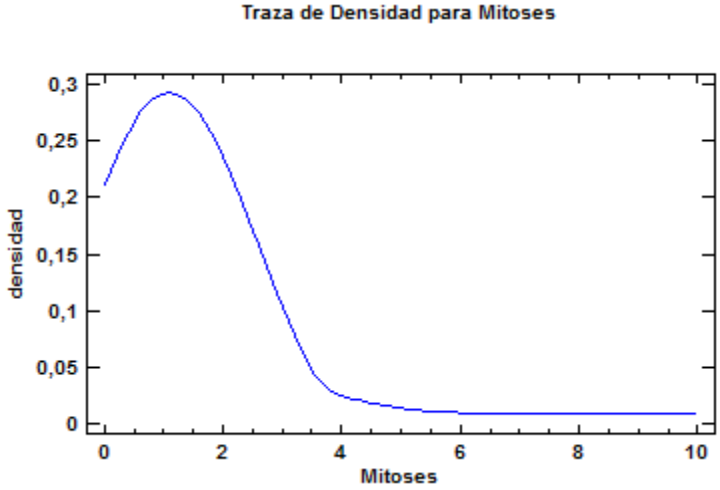
Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,363292
DMENOS	0,27047
DN	0,363292
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Normal Nucleoli proviene de una distribución normal con 95% de confianza.

- Mitoses





Pruebas de Bondad-de-Ajuste para Mitoses

Prueba de Kolmogorov-Smirnov

	Normal
DMAS	0,462779
DMENOS	0,365547
DN	0,462779
Valor-P	0,0

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que Mitoses proviene de una distribución normal con 95% de confianza.

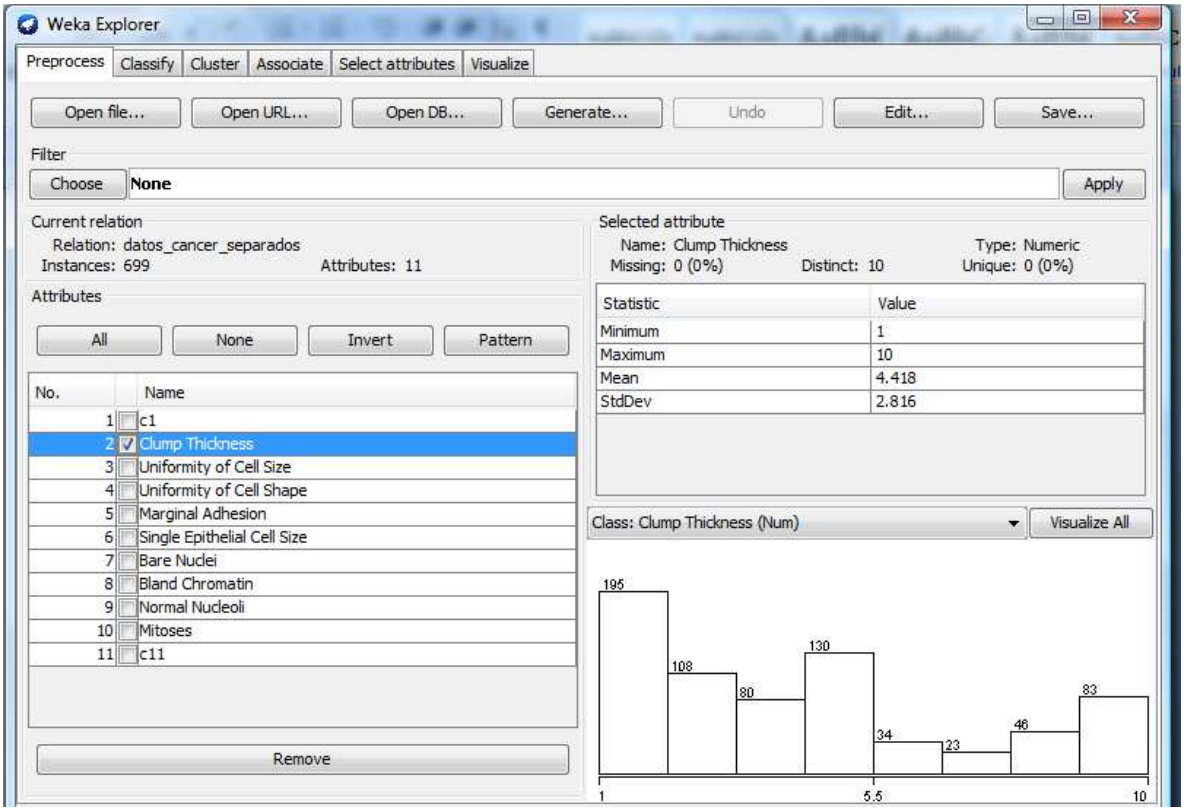
Anexo 2: Proceso de discretización sobre las variables de Breast Cancer Wisconsin (Diagnostic) Data Set.

La actividad de discretizar las variables del conjunto de datos mencionados tiene como objetivo agrupar los valores numéricos actuales de las variables en categorías, con el propósito de emplearlas para obtener reglas de asociación empleando el software WEKA.

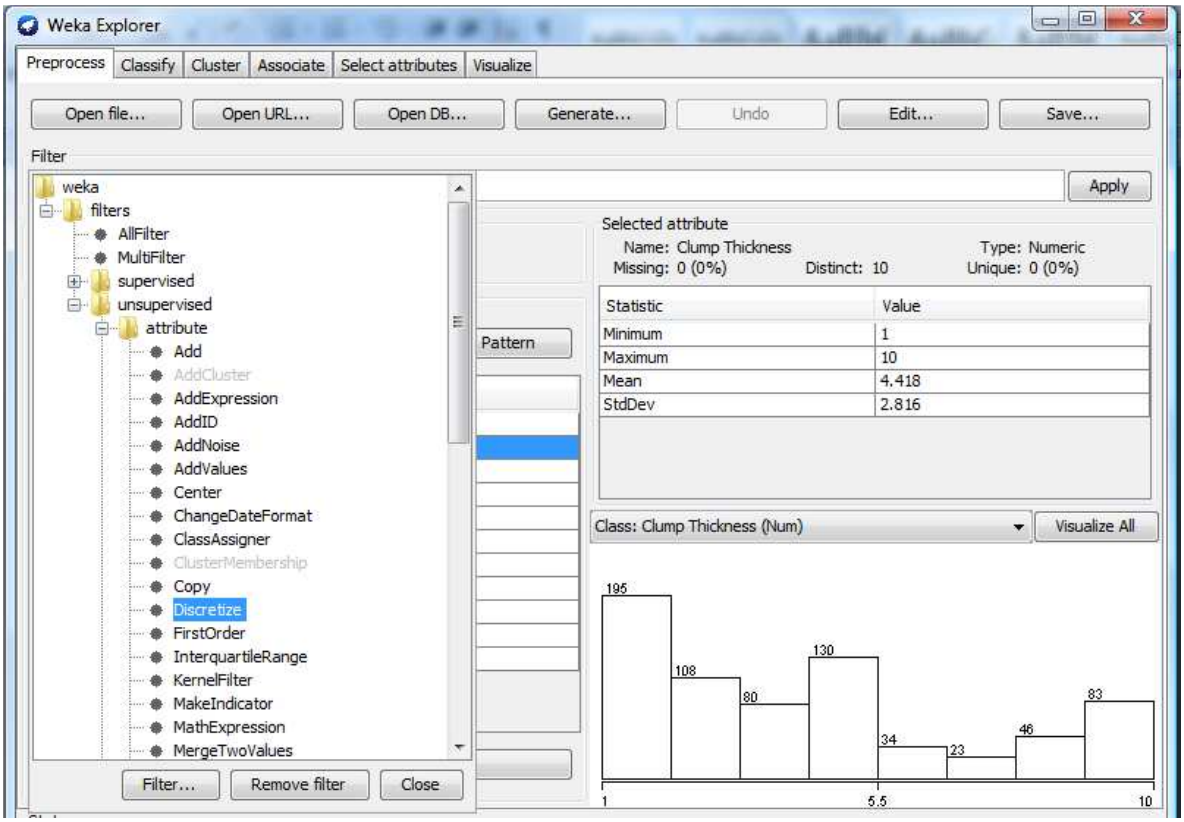
Este mismo software se emplea para ejecutar la actividad de discretizar, facilitando las opciones de configuración del algoritmo y ejecutándolo de manera muy sencilla. A continuación se ilustran los procesos sobre cada variable.

- Clump Thickness: las medidas del espesor de la masa mamaria se dividirán en los siguientes tres grupos: delgada, mediana y gruesa. El proceso lo realizamos de la siguiente manera:

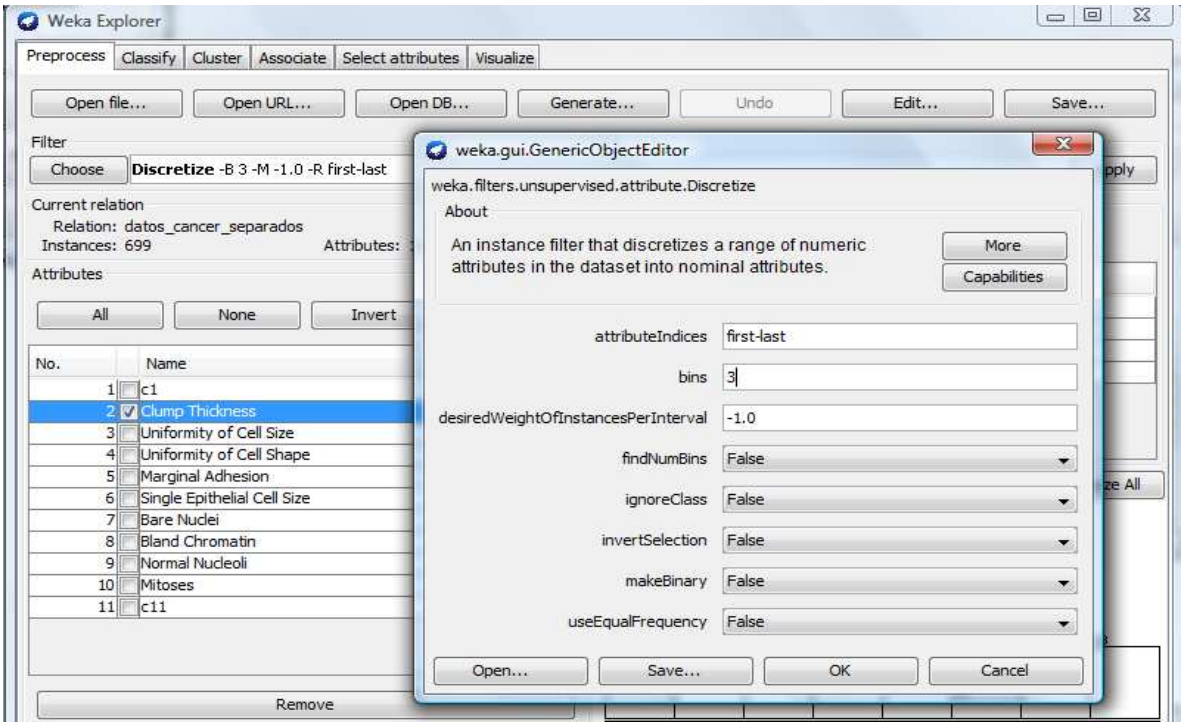
Una vez se ha ingresado al software y se han cargado los datos al mismo, como se muestra a continuación:



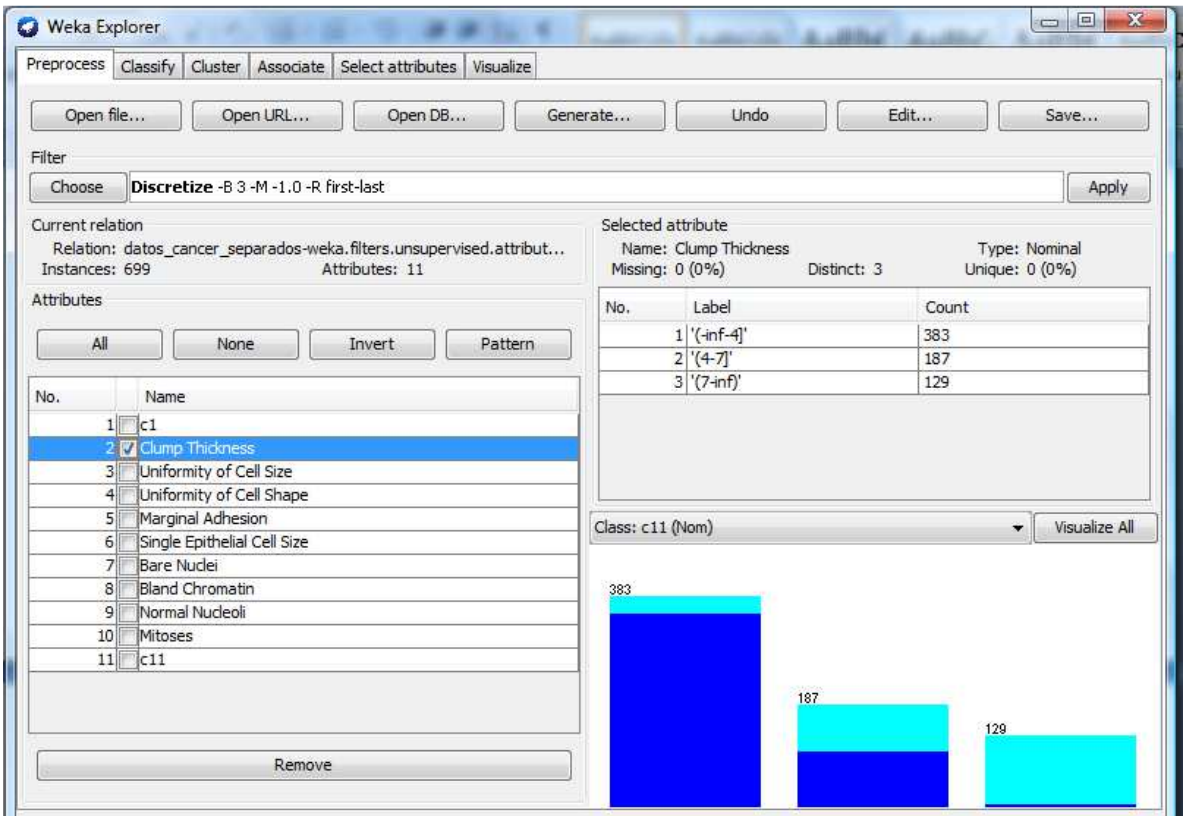
Lo siguiente es presionar el botón **Choose** para desplegar las opciones donde se puede seleccionar la opción de discretizar como se observa.



Cuando se ha seleccionado la actividad, lo siguiente es configurar los parámetros del algoritmo, donde se indica principalmente la cantidad de categorías que se desean obtener (bins).



Finalmente cuando se ha terminado la configuración solo queda ejecutar la tarea con el botón **Apply** obteniendo los siguientes resultados:



Los resultados del proceso para esta variable se resumen en la siguiente tabla:

Tabla 1: Resultados de discretizar variable Clump Thickness

Categorías	Rango	Cantidad
Delgada	[1,4]	383
Mediana	(4,7]	187
Gruesa	(7,10]	129

- Uniformity of Cell Size: Las categorías que se consideran para esta variable son: pequeña, mediana y grande obteniendo como resultado:

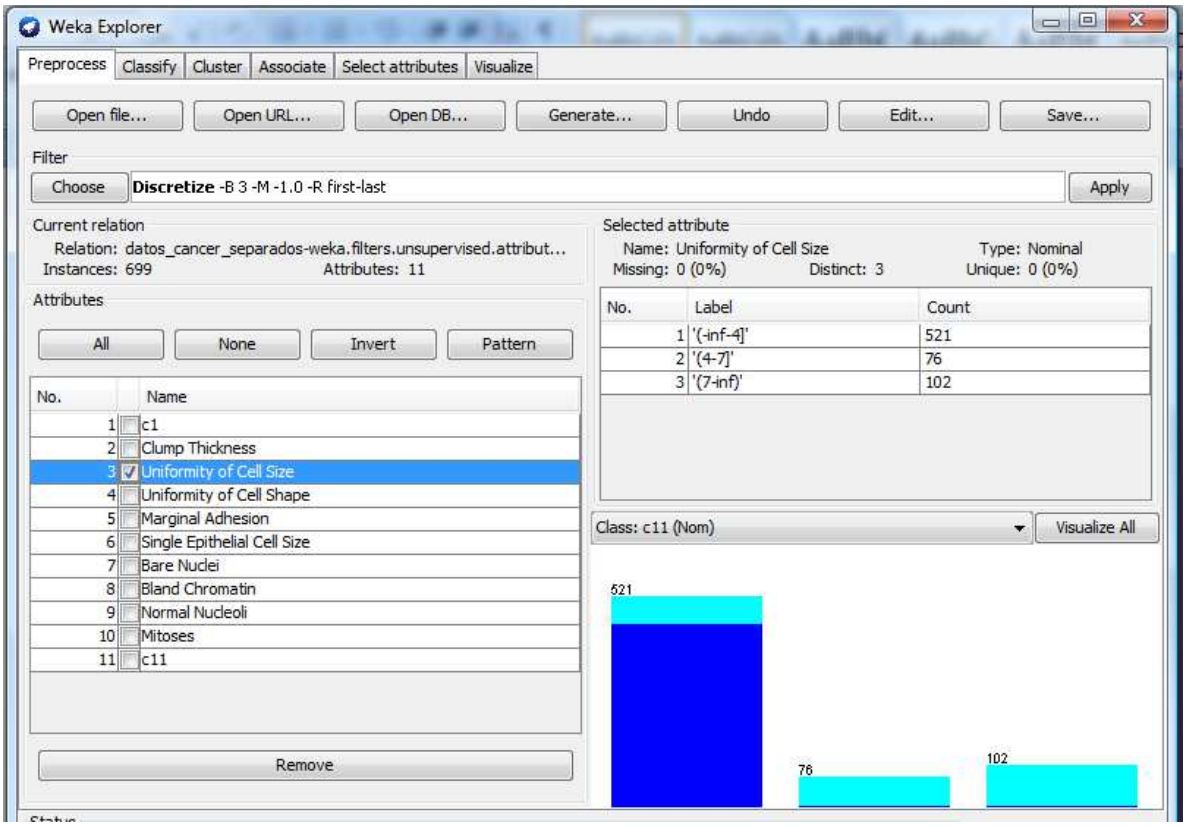


Tabla 2: Resultados de discretizar variable Uniformity of Cell Size

Categorías	Rango	Cantidad
Pequena	[1,4]	521
Mediana	(4,7]	76
Grande	(7,10]	102

- Uniformity of cell shape: dada la falta de información que se posee respecto a las variables y las razones por las que decidieron evaluarlas de forma numérica, no es claro como categorizarlas. Este es el caso de esta variable que hace referencia a la forma de las células, por tal razón se decide definir las siguientes tres categorías: clase1, clase2 y clase3.

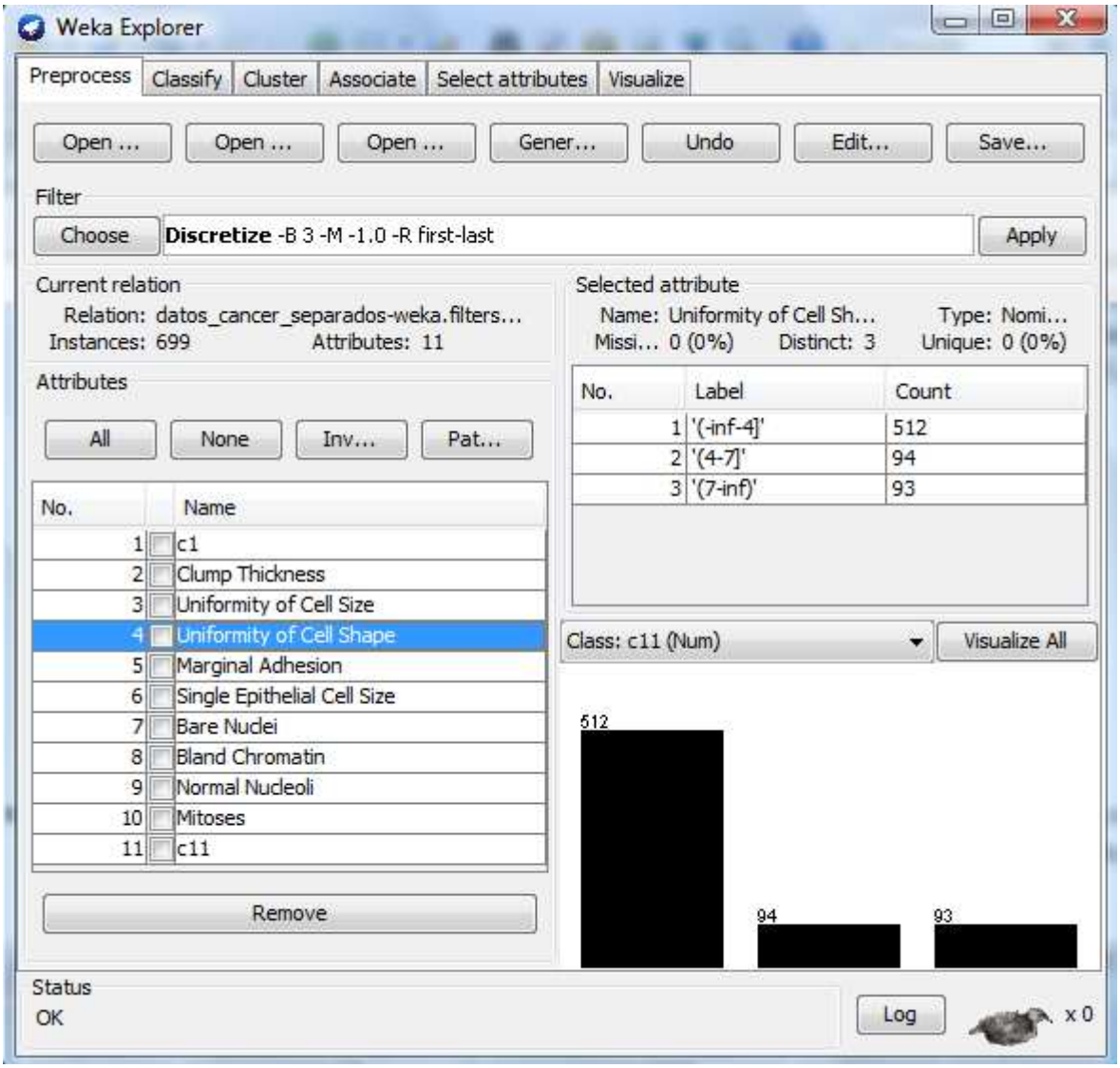


Tabla 3 Resultados de discretizar variable Uniformity of Cell Shape

Categorías	Rango	Cantidad
Clase 1	[1,4]	512
Clase 2	(4,7]	94
Clase 3	(7,10]	93

- Adhesion marginal: Los grupos en los cuales se pretende categorizar estas variables son débil y fuerte. Para este caso se cambia el parámetro que indica el número de categorías que se desean obtener.

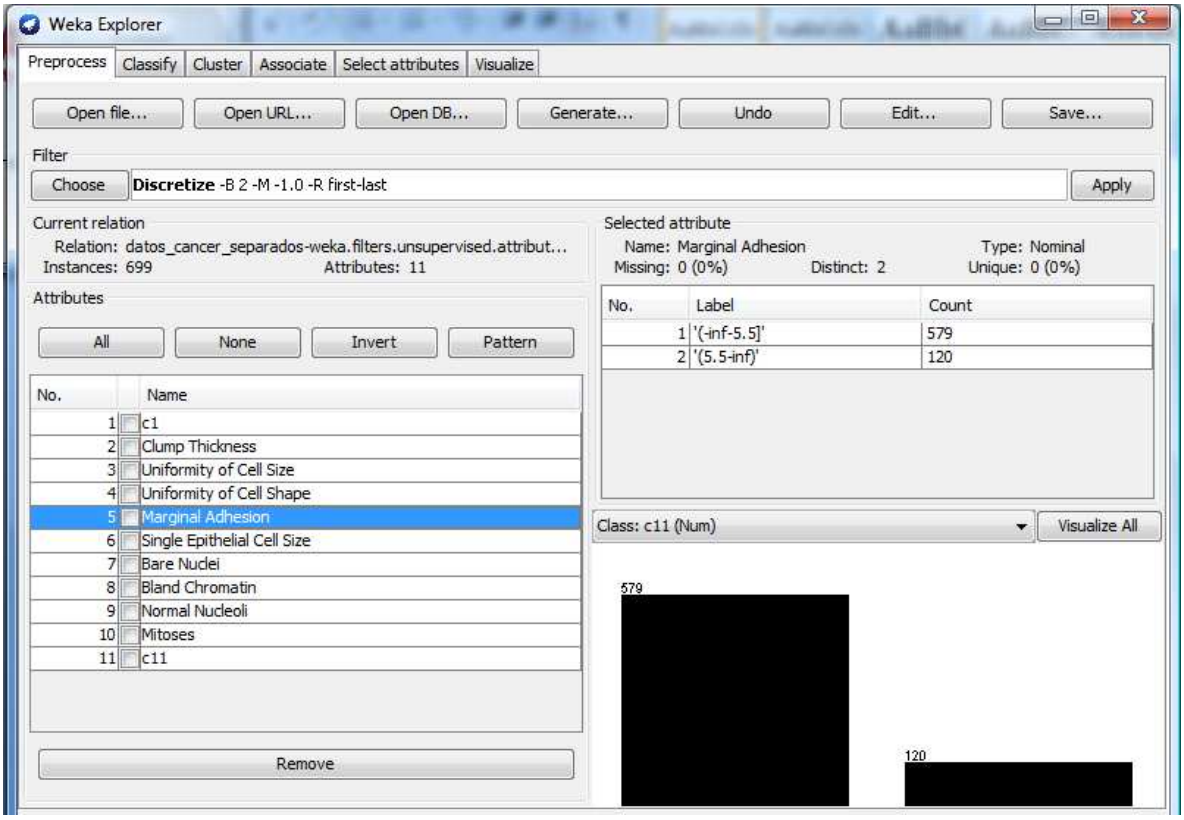


Tabla 4: Resultados de discretizar variable Adhesion marginal

Categorías	Rango	Cantidad
Débil	[1 - 5.5]	579
Fuerte	(5.5 - 10]	120

- Single Epithelial Cell Size: las categorías que se manejan para esta variable son: pequeña, mediana y grande. Los resultados que se obtienen son:

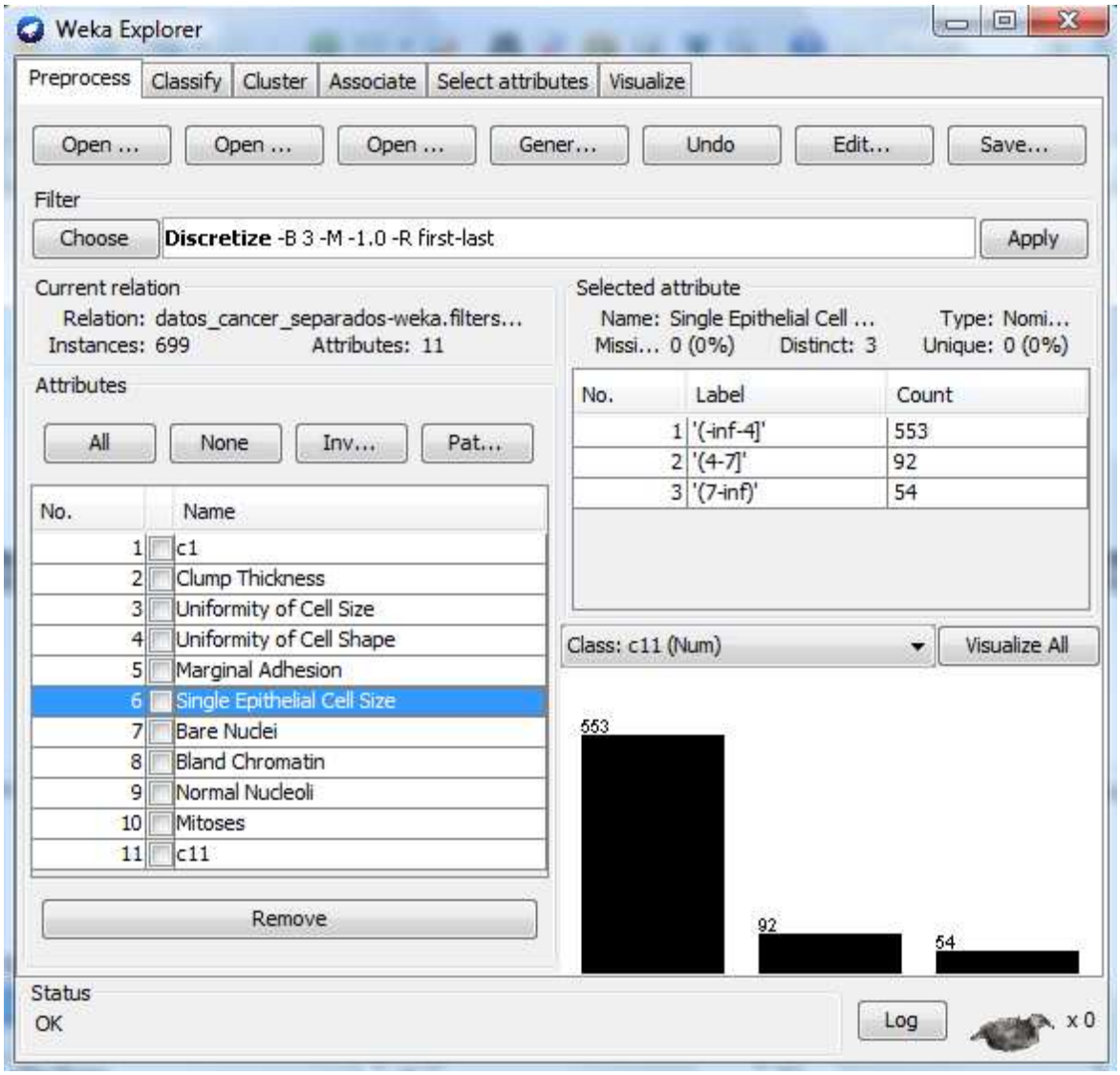


Tabla 5: Resultados de discretizar variable Single Epithelial Cell Size

Categorías	Rango	Cantidad
Pequena	[1,4]	553
Mediana	(4,7]	92
Grande	(7,10]	54

- Bare nuclei: Dado que no se conoce la significancia de los valores numéricos asignados a esta variables, se decide definir las siguientes tres categorías, bajo la suposición de que a valores más pequeños indica la omisión de esta característica y a valores altos indica la presencia de esta característica: ausente, medianamente presente, presente.

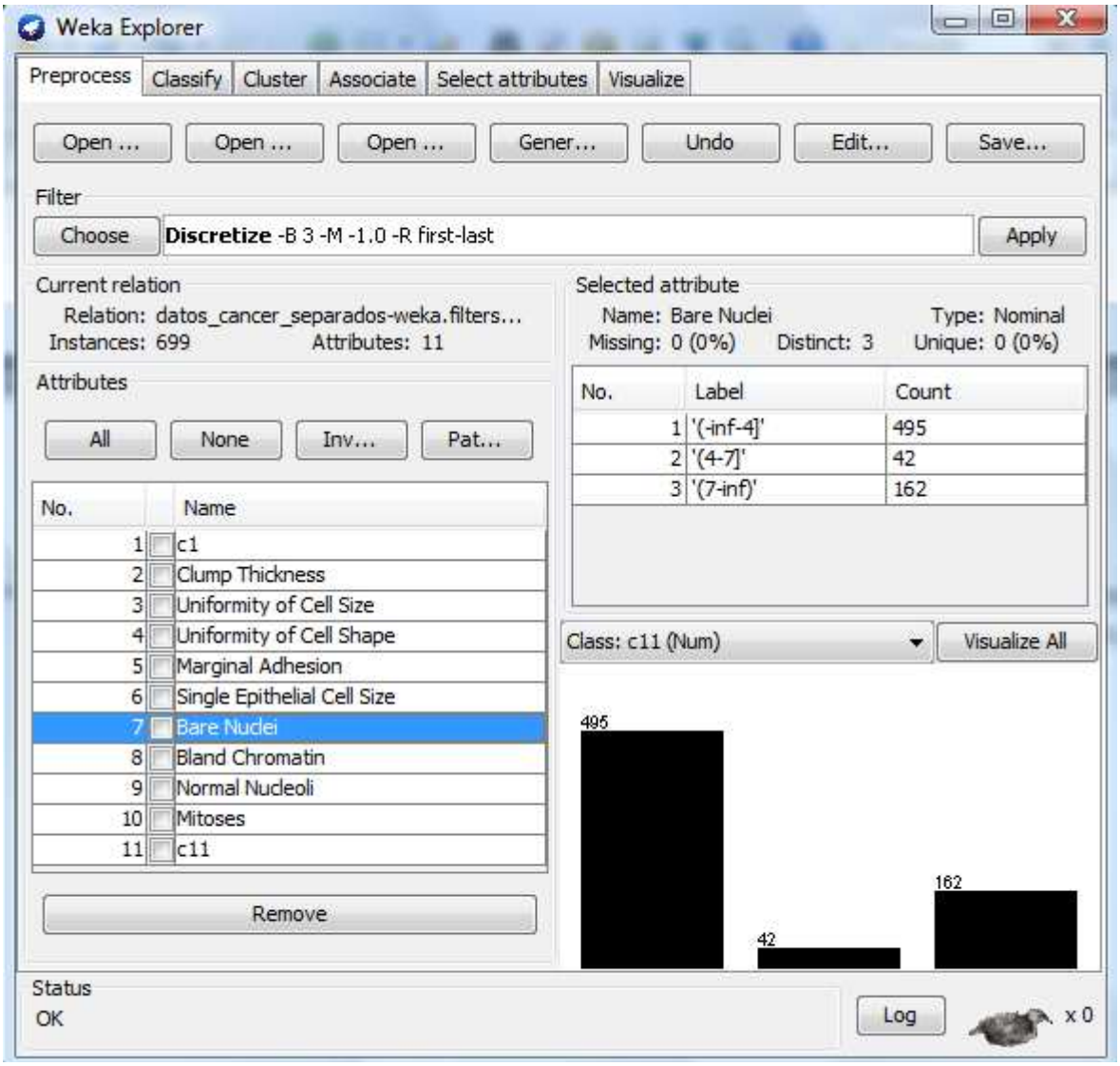


Tabla 6: Resultados de discretizar variable Bare Nuclei

Categorías	Rango	Cantidad
Ausente	[1,4]	495
Medianamente presente	(4,7]	42
Presente	(7,10]	162

- Bland chromatin: la categorización realizada para esta variable se realiza bajo la suposición de que a valores más pequeñas se tiene una textura uniforme o lisa y a valores más altos se tiene una textura burda. Dado este se define las siguientes categorías: uniforme, medianamente uniforme y burda.

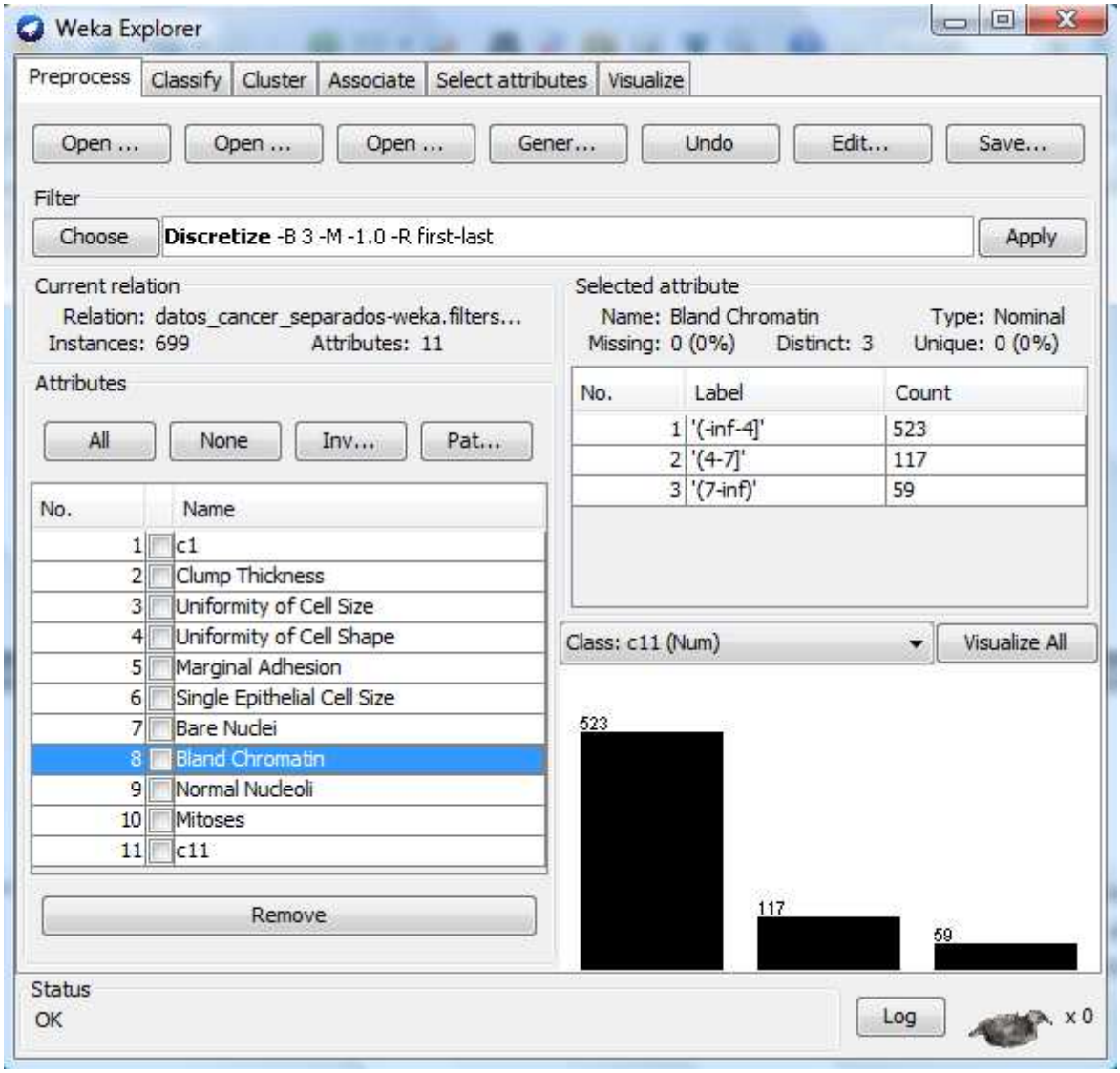


Tabla 7: Resultados de discretizar variable Bland chromatin

Categorías	Rango	Cantidad
Uniforme	[1,4]	523
Medianamente uniforme	(4,7]	117
Burda	(7,10]	59

- Normal nucleoli: para esta variable se definen las siguientes categorías: pequeña, mediana y grande, suponiendo que valores pequeños de las mediciones que se tienen hacen parte de la categoría “pequeña” y suponiendo que valores grandes hacen parte de la categoría “grande”.

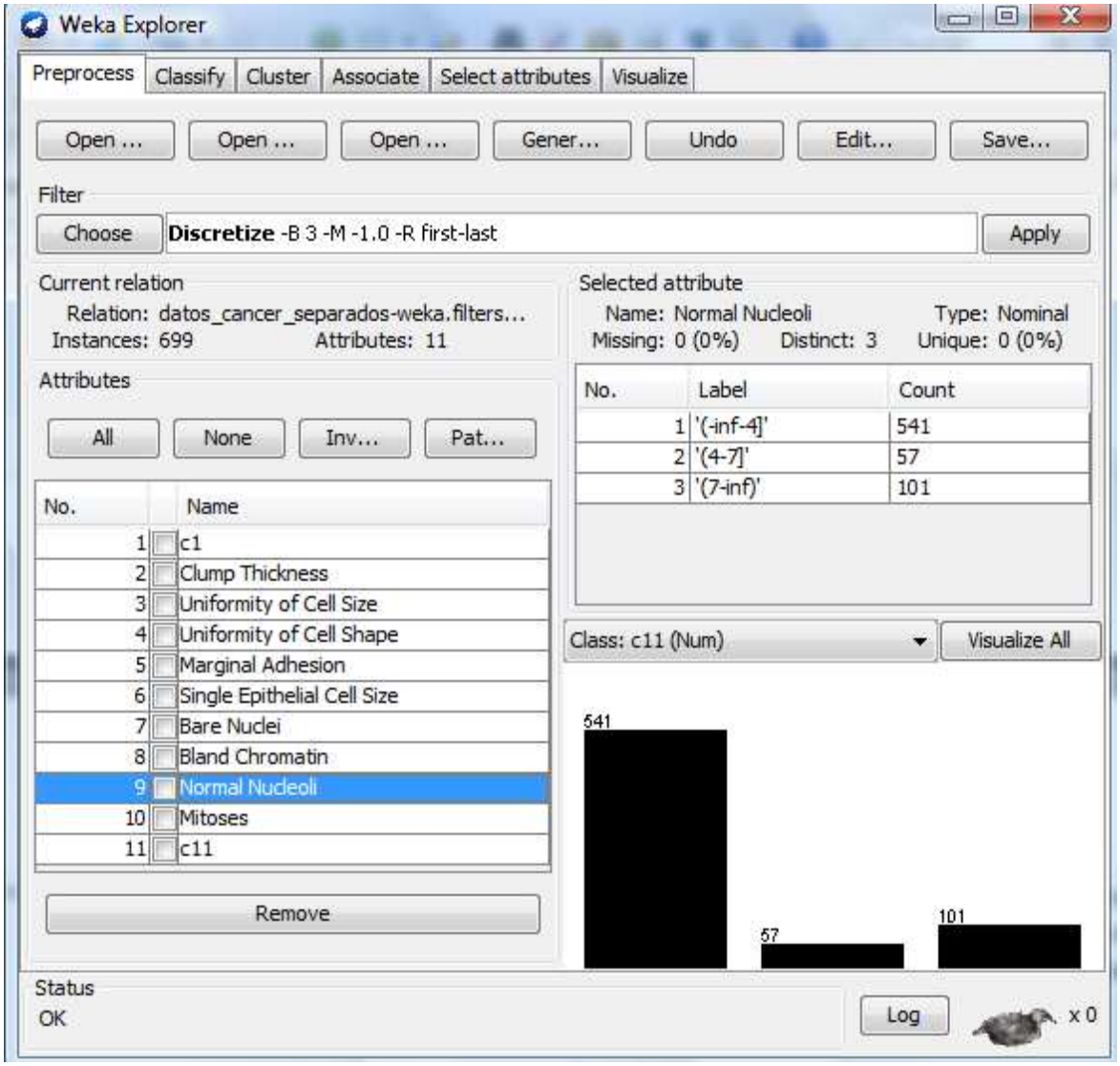


Tabla 8: Resultados de discretizar variable Normal nucleoli

Categorías	Rango	Cantidad
Pequeña	[1,4]	541
Mediana	(4,7]	57
Grande	(7,10]	101

- Mitoses: como se desconoce la significancia de las medidas que se tienen sobre esta variable, de defines las siguientes categorías: mitosis 1, mitosis 2 y mitosis 3.

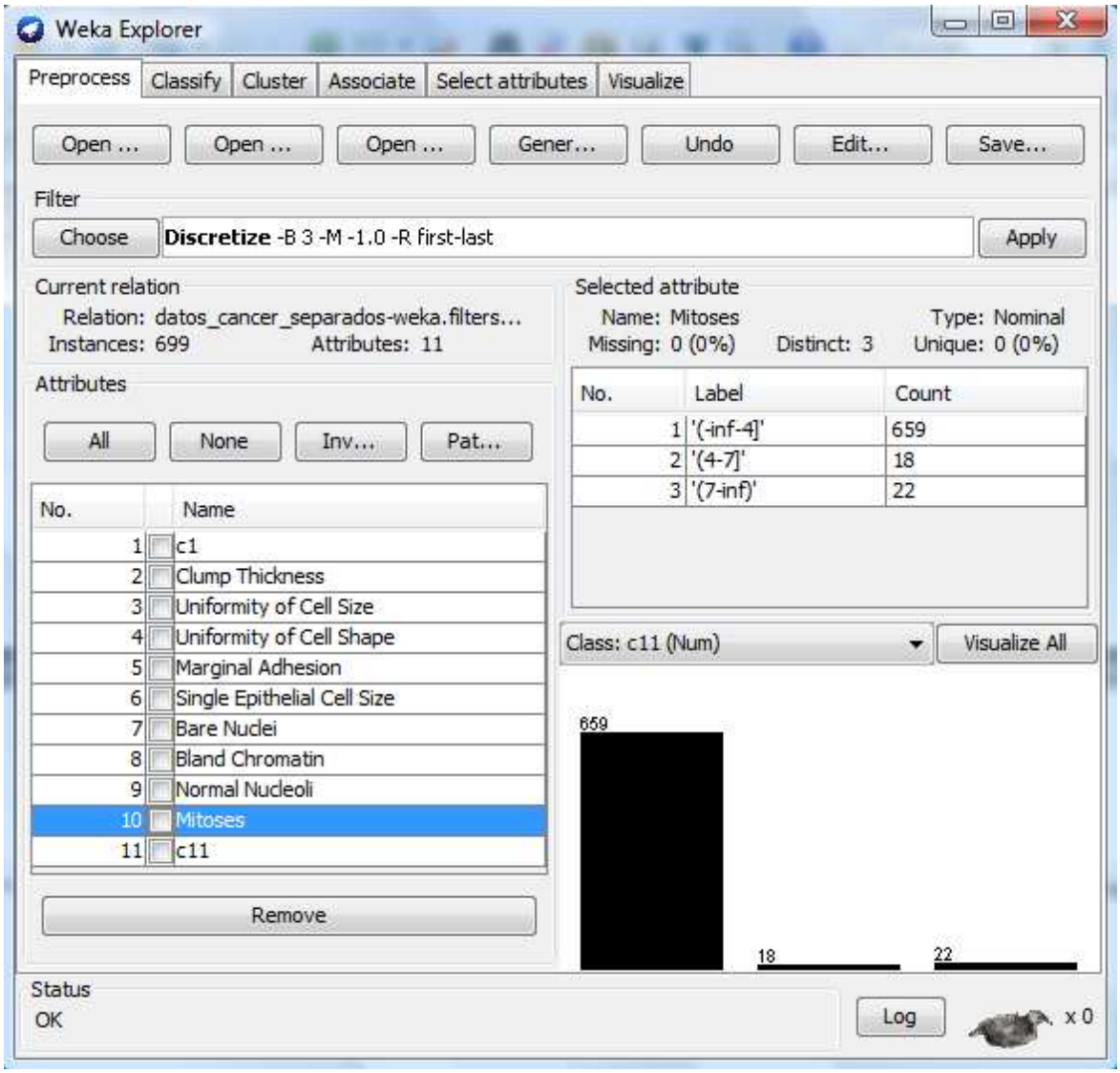


Tabla 9: Resultados de discretizar variable Mitoses

Categorías	Rango	Cantidad
Mitosis 1	[1,4]	659
Mitosis 2	(4,7]	18
Mitosis 3	(7,10]	22