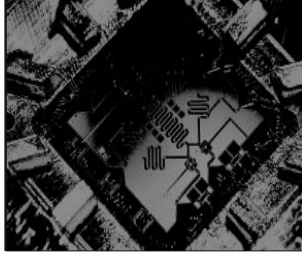


Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos

Analysis and evaluation of risk levels on loan aproval using data mining techniques

Martha L. Tello*
Hermes J. Eslava**
Lucy B. Tobías***

Fecha de envío: septiembre de 2011
Fecha de recepción: octubre de 2011
Fecha de aceptación: agosto de 2012

Resumen

En este artículo se presenta la aplicación de la minería de datos en el sector financiero, para evaluar el nivel de riesgo en el otorgamiento de créditos. Se tomó una muestra de datos de 1000 registros, correspondientes a una cartera comercial de una entidad bancaria. Se utilizó la metodología *Knowledge Discovery in Databases* (KDD)
y se desarrolló un software que permitió discretizar los datos, para poder utilizarlos como entradas en la herramienta de minería de datos WEKA. Se comparan los resultados obtenidos al aplicar las técnicas de minería de datos, árboles de clasificación ID3 y J48. Finalmente se obtiene como resultado las características que deben tener los clientes para recibir un crédito bancario.

Palabras clave
Minería de datos, KDD, árboles de decisión, reglas de decisión, ID3 y J48.

Abstract

This article describes a Data-Mining-based application intended for the financial sector. Such an application evaluates the level of risk associated to financial loans on approval. A sample of 1000 data records from a commercial bank were analyzed and further processed. *Knowledge Discovery in Databases* (KDD) methodology was implemented and a software tool that allows discrete-data conversion was developed so that the samples could be used as input data to the data mining tool called WEKA. Results were compared to assess the performance when applying data mining techniques and classification trees

* Ingeniera de Sistemas de la Universidad de Ibagué, especialista en Teleinformática de la Universidad Distrital Francisco José de Caldas (Colombia), estudiante de Maestría en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas (Colombia), integrante del grupo de investigación en telecomunicaciones Teletecno. Docente de la Universidad Distrital Francisco José de Caldas (Colombia). mtelloc@udistrital.edu.co

** Licenciado en Electrónica de la Universidad Pedagógica Nacional (Colombia), especialista en Teleinformática de la Universidad Distrital Francisco José de Caldas (Colombia), especialista en Instrumentación Electrónica de la Universidad Santo Tomas (Colombia), M.Sc. en Ingeniería de Telecomunicaciones de la Universidad Nacional de Colombia, Ph.D.(c) en Ingeniería de Sistemas y Computación de la Universidad Nacional (Colombia), director del grupo de investigación en telecomunicaciones Teletecno. Docente de la Universidad Distrital Francisco José de Caldas (Colombia). hjeslavab@udistrital.edu.co

*** Ingeniera de Sistemas de la Fundación Universitaria San Martín (Colombia), estudiante de Maestría en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas (Colombia), miembro del grupo Thomas Greg & Sons Ltda. lucy.tobias@reval.com.co.

ID3 and J48. Finally, the application yields the characteristics that customers should exhibit to be granted financial loans.

Key words
Data mining, KDD, decision trees, decision rules, ID3 and J48.

Introducción

El crédito bancario constituye una fuente primordial de financiamiento para el desarrollo de la economía. Todos los sujetos de crédito implican en menor o mayor medida un nivel de riesgo, dicha probabilidad está dada por la incertidumbre acerca de los factores y variables que pueden afectar en el futuro a los clientes y vuelven peligrosa la inversión bancaria. Cada cliente muestra sus características y factores propios que inciden en la existencia del riesgo crediticio.

En la actividad bancaria siempre los conceptos de riesgo y crédito son inseparables si se tiene en cuenta que entre las actividades bancarias la concesión de créditos es la más importante, se comprenderá entonces que la gestión del riesgo de crédito continua siendo la de mayor relevancia; el crédito ideal sería aquel que dé una seguridad total o un riesgo nulo, pero en la práctica esto es casi imposible no hay crédito sin riesgo. No obstante, sí es posible la disminución del riesgo determinando los factores que inciden en él y actuando sobre ellos para cada tipo de prestatario [1].

Un aspecto muy importante sobre el sistema de administración del riesgo de crédito es el seguimiento y control de procesos que tengan relación directa con este. Por lo tanto, se hace necesario el monitoreo de procesos tales como otorgamientos y comportamientos. Estos procesos sintetizan las diferentes etapas de la vida de una obligación, razón por la cual las variables contempladas en cada uno deben tener relación directa con el objeto mismo del crédito, así como su análisis y seguimiento.

Para desarrollar la aplicación se consideraron 10 variables, distribuidas entre cuantitativas y cualitativas, y 1000 registros de muestra, correspondientes a una cartera comercial. El estudio comienza identificando las variables que estarán directamente implicadas y la clase de referencia que dará sentido a la información de acuerdo con el conjunto de datos con el que se cuenta, el cual muestra una relación directa entre las edades, la capacidad de endeudamiento de una persona y su comportamiento de pagos; estas variables se convierten en una base fundamental para determinar el perfil de los clientes y sectores que solicitan los servicios del crédito bancario.

Luego de tener identificadas las variables que serán estudiadas, se continúa con el proceso de preparación de los datos, dentro del cual se van a realizar tareas de limpieza, integración, transformación (en caso de ser necesaria) y reducción de la información suministrada con el fin del hacer el conjunto de datos consistente.

Luego de este proceso, se prepara la información para a partir de ella desarrollar las tareas de minería, y se utilizan una serie de primitivas existentes con el fin de llevar a cabo un descubrimiento del conocimiento fácil, eficiente y fructífero. Este descubrimiento debe llevarnos a resolver la pregunta dentro de la cual se enmarca todo este proceso de minería de datos: ¿qué características debe tener un cliente para recibir un crédito?

El procesamiento de los datos se efectuó utilizando la técnica KDD y se empleó como apoyo tecnológico la herramienta de minería de datos WEKA, por medio de la cual se llevó a cabo el preprocesamiento de los datos mediante la técnica de selección de atributos; posterior a esto, se utilizó como proceso de clasificación los árboles de decisión con los algoritmos de minería ID3 y J48.

El artículo se encuentra organizado en cinco secciones: la primera muestra un breve recuento del dominio de la aplicación, la segunda sección presenta el preprocesamiento de la información, la sección tres describe el proceso de minería de datos, en la cuarta sección se muestran los resultados y en la quinta sección se presentan las conclusiones a las que se puede llegar de acuerdo con el estudio realizado y los resultados obtenidos.

1. Determinación del conjunto de datos objeto

El conjunto de datos que se va analizar proviene de la información real de una entidad financiera. Dicha información está conformada por datos básicos de los clientes y por una clasificación o categorización propia de la actividad de la empresa que los provee.

El conjunto de datos que será estudiado tiene los siguientes atributos:

- No obligación.
- Saldo a cápita.
- Edad mora: acumulado de días que lleva en mora durante el crédito.
- Modalidad.
- Comportamiento de pago: mora actual.
- Endeudamiento con el sector.
- Perfil del cliente.
- Reporte en las centrales.
- Actividad económica del cliente.
- Capacidad de pago del cliente.

Las características de los atributos antes mencionados se describen brevemente a continuación:

- Conjunto de obligaciones en diferentes rangos de días en mora.
- Modalidades de crédito que dividen los datos en cuatro grandes grupos para ser analizados: consumo, comercial, hipotecario y microcrédito.
- Conjunto de los perfiles de comportamiento de los clientes de acuerdo con su comportamiento en el pago de las obligaciones con la entidad financiera.
- Conjunto de las diferentes actividades económicas del sector.

Este conjunto de datos debe ser sometido a etapas de limpieza y preprocesamiento con el fin de lograr la reducción y normalización de la información que se va a analizar.

2. Limpieza de datos y preprocesamiento

Las grandes cantidades de información que contienen las bases de datos requieren una eficiente representación, no solo que reduzcan la dimensionalidad sino también que preserven la información relevante para una clasificación eficiente.

En el conjunto de datos que se estudiará se encuentran datos continuos y categóricos, por lo tanto es necesario convertir los continuos a categóricos. Se obtiene el número de segmentos y rangos en que pueden agruparse

los segmentos de tiempo, conservando la información relevante [2].

Para el proceso de limpieza y preprocesamiento de datos se tomó el conjunto de datos inicial y se identificaron los atributos que debían ser categorizados, así como los atributos a los que les faltaba información, para posteriormente realizar un proceso de relleno de datos faltantes y discretización. Estos procesos se llevan a cabo tal como se describe a continuación.

2.1 Relleno de datos faltantes

Dentro del conjunto de datos de estudio se encontró que el atributo comportamiento de pago tenía filas que no contaban con un valor específico, en el momento de evaluar la dependencia del resto de atributos del conjunto de datos se encontró que no había ninguna relación que determinara el comportamiento o valor que debía tener este atributo, por lo tanto, para hacer confiable el proceso de minería de esta información dicho atributo se completó con el valor "sin evaluar".

2.2 Discretización de información

Para el proceso de discretización se hizo un análisis preliminar de los atributos y se estableció cuáles son discretos y cuáles son continuos. Luego se tomaron los datos continuos y se categorizaron los valores de cada atributo en tres grupos, que se especifican de acuerdo con el intervalo que se obtiene si se aplica la siguiente fórmula:

$$Intervalo = \frac{atributos\ max - atributos\ min}{3} \quad (1)$$

Los atributos a los cuales se les aplicó el proceso de discretización fueron los siguientes:

Edad de mora del cliente: en este atributo se muestra la sumatoria de la cantidad de días de atraso que el cliente ha tenido durante la existencia de la deuda.

Para este atributo, aplicando la fórmula 1, la distribución de los grupos queda como se muestra en la tabla 1.

Capacidad de pago del cliente: este atributo muestra la capacidad de pago que tiene el cliente, determinada en porcentajes por millón. Para este atributo después de aplicar la fórmula 1, los grupos quedan conformados como lo muestra la tabla 2.

2.3 Reducción de los datos

En el proceso de reducción de los datos debe identificarse el tipo de información que estos

Tabla 1. Distribución en grupos de la edad de mora del cliente

Grupo	Rango de Valores
G1	25 – 136.67
G2	136.671 – 248.33
G3	248.331 – 360

Fuente: elaboración propia.

Tabla 2. Distribución en grupos de la capacidad de pago del cliente

Grupo	Rango de Valores
G1	0.3 – 0.87
G2	0.871 – 1.43
G3	1.431 – 2

Fuente: elaboración propia.

Tabla 3. Atributos del conjunto de datos

Variables	Descripción
No Obligación	Identificador asignado a la obligación por parte de la entidad.
Saldo a Capital	Saldo pendiente a la fecha
Edad Mora	Acumulado de días que lleva en mora durante el crédito
Modalidad	Tipo de Crédito que se otorga
Comportamiento de Pago	Mora actual
Endeudamiento con el Sector	Muestra el endeudamiento que tiene el cliente
Perfil del Cliente	Perfil que se le otorga al cliente de acuerdo con su comportamiento de pago y nivel de endeudamiento
Reporte en las Centrales	Indica si el cliente está o no reportado en las centrales de Riesgo
Actividad Económica del Cliente	Indica la actividad económica que desarrolla el cliente
Capacidad de pago del Cliente	Muestra la capacidad de pago por millón que tiene el cliente

Fuente: elaboración propia.

transmiten, dicha información puede ser de tres tipos:

1 Redundante: información repetitiva o predecible.

2 Irrelevante: información que no aporta al proceso de descubrimiento de la información.

3 Básica: la relevante, la que se constituye como parte importante en un proceso de predicción o descubrimiento de información [3].

De acuerdo con los tres tipos de información definidos antes y a partir de los datos que son objeto de estudio en este artículo, en la tabla 3 hay una descripción de todos los atributos que provee el conjunto de datos, y en la tabla 4 se muestran los atributos que van a ser removidos del conjunto de datos, especificando la razón que lleva a realizar esta acción intuitivamente.

Tabla 4. Atributos que serán removidos

Variables	Justificación
Saldo a Capital	No es relevante para determinar el perfil del cliente

Fuente: elaboración propia.

2.4 Filtros de atributos

Después de realizar la categorización de los datos y de eliminar un dato que no presenta relevancia para el proceso de descubrimiento de la información, se ingresan en la herramienta de minería de datos WEKA el conjunto de datos modificado y discretizado, compuesto por nueve columnas y mil registros.

WEKA permite realizar manipulaciones sobre los datos aplicando filtros. Se pueden aplicar en dos niveles, atributos e instancias. De los

filtros implementados en la sección de supervisados, se ha decidido aplicar sobre los datos el filtro de selección de atributos, el cual permite encontrar aquellos atributos que tienen más peso a la hora de determinar si los datos son de una clase u otra, el resultado de estos filtros servirá de ayuda para aplicar posteriormente las técnicas de minería de datos [4].

El resultado que se obtuvo fue un nuevo conjunto de datos conformado por siete campos que contienen la información relevante para el proceso de descubrimiento de la información. En la figura 1 se muestra la estructura de la información cargada inicialmente y en la figura 2 los atributos que resultan luego de aplicar el proceso de selección de atributos.

Figura 1. Estructura de la información cargada inicialmente

No.	Name
1	No obligacion
2	Modalidad
3	Comportamiento de Pago
4	Endeudamiento con el Sector
5	Perfil del Cliente
6	Reporte en las Centrales
7	Actividad Economica del Cliente
8	Edad Mora-C
9	Capacidad de pago del Cliente-C

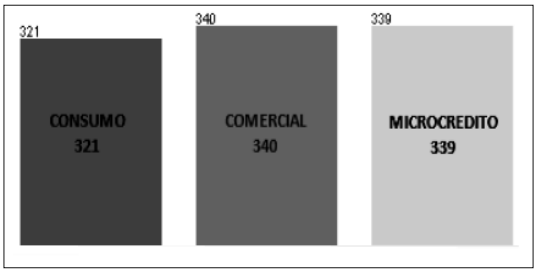
Fuente: elaboración propia.

Figura 2. Estructura de la información luego de aplicar el proceso de selección de atributos

No.	Name
1	Modalidad
2	Comportamiento de Pago
3	Perfil del Cliente
4	Reporte en las Centrales
5	Actividad Economica del Cliente
6	Edad Mora-C
7	Capacidad de pago del Cliente-C

Fuente: elaboración propia.

Figura 3. Distribución de la información para el atributo modalidad



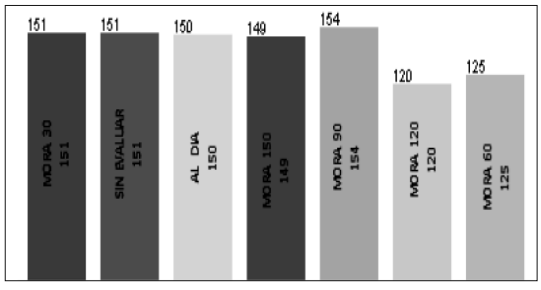
Fuente: elaboración propia.

Luego de haber realizado la reducción del conjunto de datos puede entrarse a analizar la distribución de la información en algunos atributos, en las siguientes figuras se muestra gráficamente dicha distribución con datos concretos:

En la figura 3 se muestra la distribución de la información de acuerdo con el atributo modalidad, como se puede observar hay un equilibrio entre las categorías del atributo.

En la figura 4 se muestra la distribución del atributo comportamiento de pago, y la tendencia de acuerdo con cada categoría.

Figura 4. Distribución de la información para el atributo comportamiento de pago



Fuente: elaboración propia.

3. Definición de la técnica de minería de datos

Para el desarrollo de esta investigación se utiliza la técnica de árboles de decisión en la construcción de modelos a partir de los datos. Algunos de los modelos predictivos más empleados en el área del riesgo crediticio son las técnicas de árboles de decisión.

Los árboles de decisión (*Decision Trees*, DT) son una popular herramienta utilizada en análisis estadístico y minería de datos. Los DT son ideales para realizar clasificación y predicción, y por lo general los métodos basados en árboles representan reglas. Los árboles de decisión son muy útiles en la exploración de datos en los cuales se desea encontrar relaciones entre una cantidad enorme de datos. También los DT combinan la exploración y el modelamiento de datos.

Un árbol de decisión es una estructura que permite dividir un extenso conjunto de datos relacionados entre sí en conjuntos más pequeños de datos mediante la aplicación secuencial de sencillas reglas de decisión. Adicionalmente, los árboles de decisión poseen una estructura de árbol donde cada nodo representa una "prueba" o condición sobre el valor de un atributo, las ramas representan el resultado de la evaluación del atributo y las hojas (finales en el árbol) son las clases o variables dependientes [5].

Los arboles de decisión a diferencia de otras técnicas [6]:

- Facilitan la interpretación de los datos.
- Proporcionan un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
- Explican el comportamiento respecto a una determinada tarea de decisión.
- Reducen el número de variables independientes.
- Permiten establecer la selección del algoritmo de minería de datos.

Para clasificar los datos se ha utilizado la herramienta de minería de datos llamada WEKA y para medir la efectividad del algoritmo de clasificación se ha comparado la clase predicha con la clase real de las instancias. Existen diversos modos para llevar a cabo la evaluación, en este caso se empleó *use training set* que permite utilizar la misma muestra para entrenar y probar. Los resultados obtenidos son positivos, pero no corresponden con la realidad (está clasificando los mismos datos con los que se ha entrenado) [7].

Todos los algoritmos de clasificación tienen dos etapas, entrenamiento y test. La primera ajusta el algoritmo de clasificación con una parte del conjunto de datos (conjunto de entrenamiento). La segunda, evalúa dicho algoritmo en la etapa de test con el conjunto de datos de test; la división del conjunto de datos suele ser 70% para el entrenamiento y 30% para la evaluación [8].

El conjunto de entrenamiento se utiliza para generar el modelo (árbol, lista de reglas, etc.) y el conjunto de test para verificar si el comportamiento del modelo es correcto con ejemplos no vistos anteriormente [9].

Entre los algoritmos que proporciona WEKA, se analizaron los siguientes:

3.1 Algoritmo ID3

Uno de los algoritmos de inducción de árboles de clasificación más populares es el denominado ID3 introducido por Quinlan (1986). En este, el criterio escogido para seleccionar la

variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase. La terminología usada en este contexto para denominar a la cantidad de información mutua es la de ganancia en información (*information gain*).

Esto es debido a que:

$$I(X_i; C) = H(C) - H(C \mid X_i) \qquad (2)$$

Lo que viene a representar esta cantidad de información mutua entre X_i y C es la reducción en incertidumbre en C debida al conocimiento del valor de la variable X_i.

Matemáticamente se demuestra que este criterio de selección de variables utilizado por el algoritmo ID3 no es justo, ya que favorece la elección de variables con mayor número de valores. Además, el algoritmo ID3 efectúa una selección de variables previa (denominada *pre-running* en este contexto) que consiste en efectuar un test de independencia entre cada variable predictora X_i y la variable clase C, de manera que para la inducción del árbol de clasificación tan solo se van a considerar aquellas variables predictoras para las que se rechaza el test de hipótesis de independencia [10].

3.2 Algoritmo J48 (C4.5)

El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizado. Se trata de un refinamiento del modelo generado con OneR [11].

El algoritmo C4.5 construye árboles de la decisión de un sistema de datos del entrenamiento de la misma forma que ID3, que usa el concepto de entropía de la información. Los datos del entrenamiento son un sistema S = s1,s2,... de muestras ya clasificadas. Cada muestra s_i = x1,x2,... está en un vector donde x1,x2,... representa las cualidades o las características de la muestra.

Los datos del entrenamiento se aumentan con un vector C = c1,c2,... donde c1,c2,... representa la clase a la que pertenece cada muestra.

C4.5 utiliza el hecho de que cada cualidad de los datos puede utilizarse para tomar una decisión que parta los datos en subconjuntos más pequeños. C4.5 examina la diferencia en entropía, eso resulta de elegir una cualidad para partir los datos. La cualidad con el aumento normalizado más alto de la información es la que está usada para tomar la decisión. El algoritmo entonces se repite en las sublistas más pequeñas [11].

3.3 Comparación entre el algoritmo C4.5 e ID3

El algoritmo C4.5 llevó a cabo un número de mejoras a ID3, algunas de estas son:

- Dirigiendo las cualidades continúas y discretas para manejar las cualidades continuas, C4.5 crea un umbral y después parte la lista en las que valor de la cualidad esté sobre el umbral y las que sean inferior o igual a él [12].
- Si se manejan datos de entrenamiento con valores faltantes C4.5 permite que los valores faltantes sean marcados como "?". Los valores que faltan simplemente no se utilizan en cálculos del aumento de la entropía.
- Manipulación de cualidades con valores diferentes.
- Árboles de poda después de la creación. C4.5 pasa a través del árbol una vez que se haya creado y procura quitar las ramas que no ayudan substituyéndolos por nodos de la hoja [11].

4. Interpretación de los resultados

Las consideraciones importantes para construir un buen modelo radican en la calidad de los datos escogidos y en la selección adecuada de las variables que influyen en los modelos. Todo esto depende también de las técnicas de minería empleadas en el preprocesamiento de los datos y de cómo afronte el modelo la información disponible.

Teniendo en cuenta las consideraciones antes mencionadas, se ha aplicado a un conjunto de datos que denominaremos de entrenamiento los algoritmos de árboles de decisión ID3 y J48; de los resultados obtenidos, se ha decidido mostrar un comparativo entre los elementos más relevantes de dichos algoritmos para evidenciar la precisión con que analizaron la información, y para decidir con cuál sería más adecuado trabajar en pro de tener un proceso de minería de datos confiable y con un nivel de precisión alto.

En las subsecciones siguientes agrupamos los resultados de los dos algoritmos para hacer las comparaciones pertinentes.

4.1 Comparación entre los resultados de los métodos ID3 y J48

Luego de aplicar los algoritmos ID3 y J48 al conjunto de datos de entrenamiento, se obtuvieron los resultados que se muestran en la tabla 5, la cual presenta de manera comparativa las instancias correctas y el valor del error absoluto, generadas por cada algoritmo.

4.2 Comparación de matrices de confusión

La matriz de confusión es una herramienta de visualización que se emplea en el aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases [12].

Tabla 5. Comparativa de algoritmos de clasificación

Algoritmo	Instancias Correctas	Error Absoluto
ID3	80%	0,0837
J48	51,3%	0,2366

Fuente: elaboración propia.

A continuación se muestran las matrices de confusión generadas por cada uno de los algoritmos, aplicados al mismo conjunto de datos.

En la figura 5 podemos observar que los valores de la diagonal son los aciertos y el resto los errores. Para el Algoritmo ID3 se observa que de los 214 usuarios con perfil A, 208 fueron bien clasificados y 6 presentaron errores.

Figura 5. Matriz de confusión weka.classifiers.trees.Id3

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 208   0   3   0   3 |   a = A
  21 166   2   1   1 |   b = B
  21   8 156   3   1 |   c = D
  25  22  13 142   2 |   d = C
  26  13  17  18 128 |   e = AA
```

Fuente: elaboración propia.

Figura 6. Matriz de confusión weka.classifiers.trees.J48

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 130  27  17  20  20 |   a = A
  40  92  16  18  25 |   b = B
  34  13 100  26  16 |   c = D
  36  21  26  96  25 |   d = C
  30  21  27  29  95 |   e = AA
```

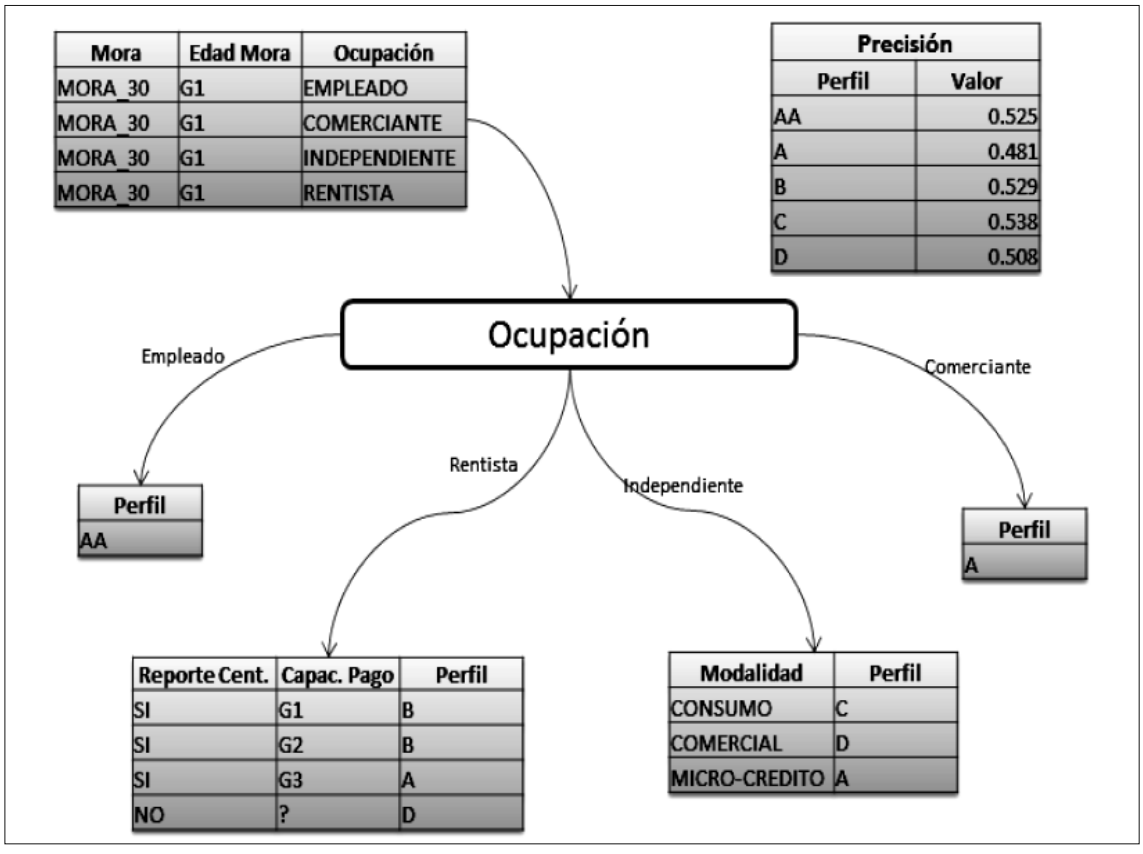
Fuente: elaboración propia.

Para el Algoritmo J48 en la figura 6 se observa que de los 214 usuarios con perfil A, 130 fueron bien clasificados y 84 presentaron errores.

4.3 Comparación entre árboles de decisión generados

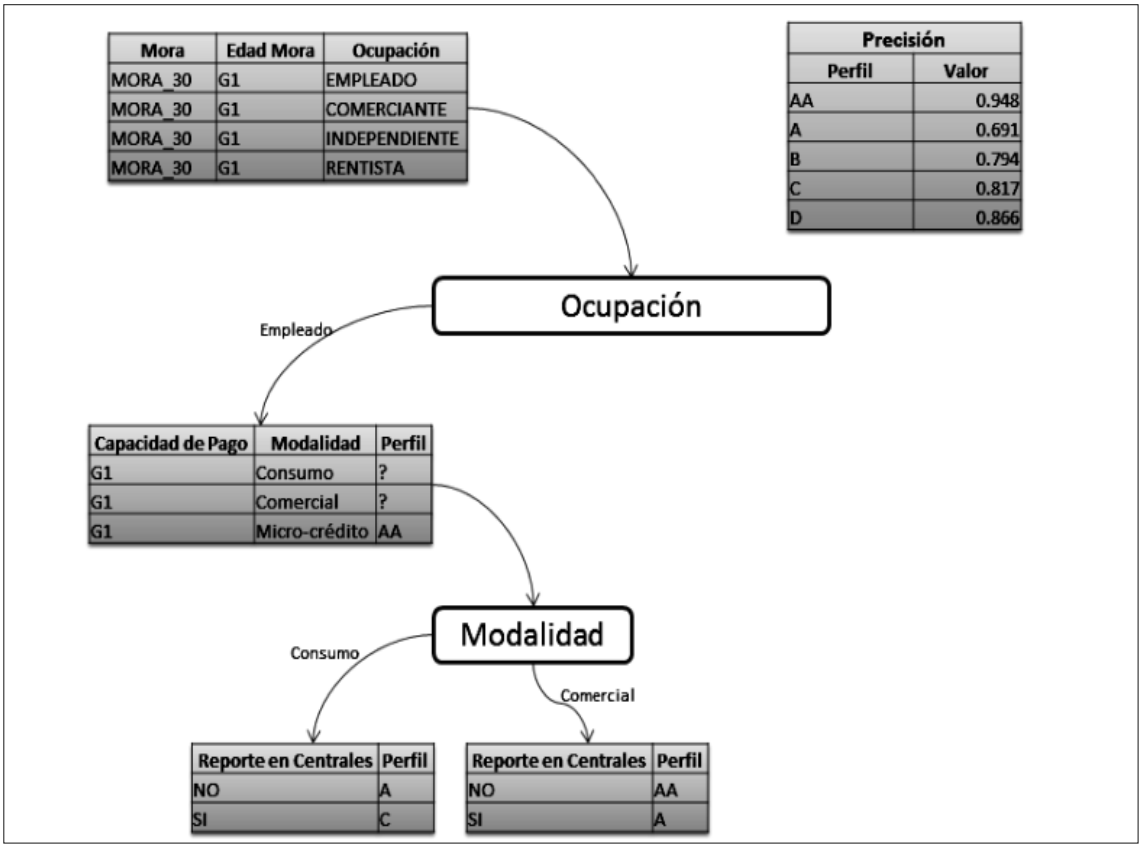
Para ilustrar cómo funcionan las reglas que se generan luego de ejecutar los algoritmos ID3 y J48 en WEKA, a continuación se muestra la ejemplificación de una regla generada por cada uno de los algoritmos antes mencionados. En la figura 7 se expone una regla generada con el algoritmo J48, y en la figura 8 se presenta la ejemplificación de una regla generada con el algoritmo ID3; como puede observarse, en la comparación entre estas dos imágenes el árbol generado para las reglas del algoritmo ID3 cuenta con mayor profundidad, teniendo en cuenta el porcentaje de instancias

Figura 7. Ejemplificación de una regla generada con el algoritmo J48 (C4.5)



Fuente: elaboración propia.

Figura 8. Ejemplificación de una regla generada con el algoritmo ID3



Fuente: elaboración propia.

correctas para cada algoritmo puede pensarse que mientras más profundidad tenga el árbol, se va a obtener mayor precisión en el proceso de minería de los datos.

5. Comparación de resultados entre datos de entrenamiento y datos de prueba

Anteriormente se han presentado los resultados que surgen de aplicar a los datos de entrenamiento los algoritmos J48 (C4.5) e ID3, en la herramienta de minería de datos WEKA. Los datos de entrenamiento constituyen el 70 % del conjunto de datos original, el 30 % restante son datos de prueba, a continuación se muestran los resultados obtenidos en una vista comparativa con el resultado de los datos de entrenamiento.

Se comienza por establecer una comparación en la precisión de los algoritmos, luego por efectuar la comparación de las matrices de confusión y, finalmente, se desarrolla una breve conclusión acerca de la comparación de estos resultados.

5.1 Comparación de precisión entre los algoritmos ID3 y J48

En la tabla 6 se puede ver que para el algoritmo ID3 la cantidad de instancias correctas

Tabla 6. Comparativo de precisión de los algoritmos con los datos de entrenamiento y pruebas

alg	Entrenamiento		Pruebas	
	Instancias Correctas	Error Absoluto	Instancias Correctas	Error Absoluto
ID3	80%	0.0837	80.40%	0.0829
J48	51.30%	0.2366	46.18%	0.2384

Fuente: elaboración propia.

y el error absoluto son muy parecidos para los datos de entrenamiento y de pruebas, y además se puede deducir que si el número de instancias correctas sube, entonces el error absoluto disminuye.

Para el algoritmo J48 la cantidad de instancias correctas disminuye, mientras que el error absoluto aumenta en proporciones similares.

5.2 Comparación de matrices de confusión

Como lo muestra la figura 9, se evidencia que ambas matrices tienen una estructura muy parecida, la clasificación de los datos mantiene las proporciones de distribución dentro de la matriz.

La figura 10 muestra las matrices de confusión para el algoritmo J48, al igual que con el algoritmo ID3, la distribución de la clasificación dentro de la matriz mantiene las proporciones entre los datos de entrenamiento y de prueba.

6. Trabajo futuro

En un futuro se pueden utilizar las reglas obtenidas para implementar un algoritmo predictivo que, basado en dichas reglas, determine si un cliente cumple con las condiciones necesarias para que le sea otorgado un crédito.

7. Conclusiones

Al realizar la comparación de los resultados obtenidos, proporcionando a los algoritmos seleccionados los datos de entrenamiento y

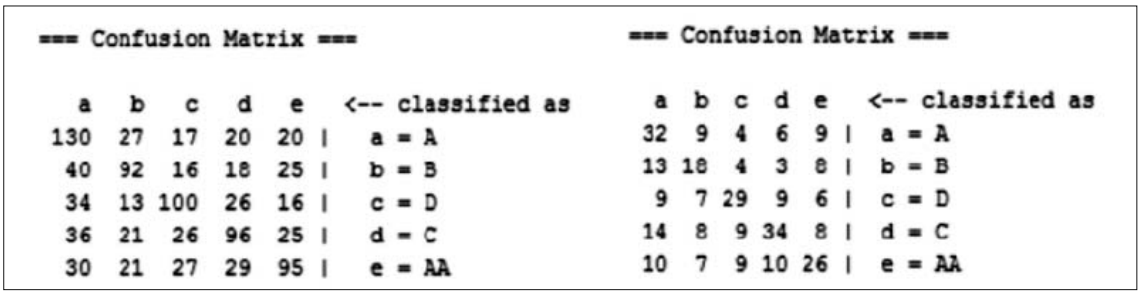
Figura 9. Matrices de confusión para datos de entrenamiento y de prueba algoritmo ID3

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 208   0   3   0   3 |   a = A
  21 166   2   1   1 |   b = B
  21   8 156   3   1 |   c = D
  25  22  13 142   2 |   d = C
  26  13  17  18 128 |   e = AA
```

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 60  0  0  0  0 |  a = A
  2 44  0  0  0 |  b = B
  7  4 46  2  1 |  c = D
  8  7  8 50  0 |  d = C
  8  4  3  5 42 |  e = AA
```

Fuente: elaboración propia.

Figura 10. Matrices de confusión para datos de entrenamiento y de prueba algoritmo J48



```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 130  27  17  20  20 |   a = A
  40  92  16  18  25 |   b = B
  34  13 100  26  16 |   c = D
  36  21  26  96  25 |   d = C
  30  21  27  29  95 |   e = AA
```

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  32   9   4   6   9 |   a = A
  13  18   4   3   8 |   b = B
   9   7  29   9   6 |   c = D
  14   8   9  34   8 |   d = C
  10   7   9  10  26 |   e = AA
```

Fuente: elaboración propia.

prueba, puede concluirse que el algoritmo ID3, al tener más profundidad en el árbol de decisión, provee mayor precisión al proceso de clasificación de la información de los clientes.

Los datos de entrenamiento proveídos a los algoritmos dan un alto nivel de efectividad al proceso de clasificación, esto se comprueba fácilmente al ejecutar dichos algoritmos con los datos de prueba y al notar que los resultados de precisión y matrices de confusión conservan las proporciones con respecto a los resultados obtenidos con el conjunto de datos de entrenamiento.

Referencias

[1] J. C. Mayo y N. O. Fonseca, "Fundamentación teórica sobre el proceso del crédito bancario a usufructuarios de tierras en Bandec las tunas", *Observatorio de la Economía Latinoamericana*, N.° 143, 2011.[En línea] disponible en http://www.eumed.net/cursecon/ecolat/cu/2011/

[2] D. A. García, "Algoritmo de discretizacion de series de tiempo basado en entropía y su aplicación en datos colposcopicos", tesis para obtener el grado de Maestro en Inteligencia Artificial. Universidad Veracruzana. México. Sep., 2007.

[3] "Compresión de Datos, compresión compresores de archivos, ficheros y carpetas. Formatos de compresión zip, arj, arc, gz, tar, 7z, sqx, rar" [Online]. Available http://www.compresion.es/compresion-de-datos/. [Accessed: 23-May-2011].

[4] M. G. Jiménez y A. Álvarez, "Análisis de datos en WEKA – pruebas de selectividad". [En línea] disponible en http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

[5] A. Y. Ramírez, "Técnicas de minería de datos aplicadas a la construcción de modelos de score crediticio: estado del arte".

[6] J. Han y M. Kamber, *Data mining: concepts and techniques*. Morgan Kaufmann. United States of America. 2006.

[7] "miweb - Concha Bielza". [Online] available http://www.dia.fi.upm.es/~concha/. [Accessed: 30-May-2011].

[8] E. J. Vázquez y D. G. Bertoli, "Sistema de localización en redes Wi-Fi con WEKA". [En línea] disponible en http://www.utim.edu.mx/~svalero/docs/e4.pdf

[9] "ISA - Ingeniería de Sistemas y Automática": [Online] available: http://isa.umh.es/. [Accessed: 31-May-2011].

[10] P. Larranaga, I. Inza, y A. Moujahid, "Tema 10: árboles de clasificación". [En línea] disponible en http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t10arboles.pdf

[11] M. A. Ayuso y M. Á. B. Mancha, "Minería de datos: intrusiones de Red". [En línea] disponible en http://www.it.uc3m.es/jvillena/irc/practicas/07-08/Intrusiones De Red.pdf

[12] C. L. Corso y S. L. Alfaro, "Alternativa de herramienta libre para la implementación de aprendizaje automático". [En línea] disponible en http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/Alternativa_de_herramienta_para_Mineria_Datos_CNEISI_2009.pdf