

APROBACIÓN DE CRÉDITOS BANCARIOS UTILIZANDO INTELIGENCIA ARTIFICIAL

Israel Cueva Hidalgo

Mayo, 2010

I. INTRODUCCIÓN

Hoy en día son muchas las instituciones que conceden créditos a sus clientes; pero el problema al que se enfrentan estas instituciones no es la falta de créditos solicitados, sino la certeza necesaria para conceder dichos créditos a sus clientes; enfocándonos más en este tema, con este trabajo se pretende analizar la aprobación de créditos bancarios basándonos en un ambiente local, es decir, en la ciudad de Loja – Ecuador, más específicamente en el Banco Nacional de Fomento (BNF).¹ La aprobación o no de créditos bancarios es un tema delicado, y es una tarea que requiere de un análisis previo y de una interpretación correcta de los resultados. Basándonos en datos históricos almacenados previamente en una base de datos sobre los clientes que han solicitado créditos y utilizando el aprendizaje automático como medio para esto, se pueden encontrar patrones comunes que nos ayuden a tomar la mejor decisión a la hora de aprobar o rechazar dichos créditos.

II. DESCRIPCIÓN DEL PROBLEMA

El Banco Nacional de Fomento (BNF), brinda a sus clientes múltiples opciones para la concesión de créditos bancarios, créditos tales

como: créditos de transporte, pesqueros, microcréditos, créditos agrícolas, etc. (1)

Pero el problema de la aprobación o no de créditos es un problema global de todas las instituciones que los otorgan, la decisión final para aprobar estos créditos es de suma importancia, el impacto de un mal análisis de la aprobación de un crédito puede ocasionar grandes problemas a dichas instituciones; es por esto que la revisión para poder otorgar un préstamo es algo arriesgado y requiere de un tiempo considerable para su correcto análisis.

III. JUSTIFICACIÓN

La aprobación o no de créditos bancarios es un tema delicado, y es una tarea que requiere de un análisis previo y de una interpretación correcta de los resultados. Basándonos en datos históricos almacenados previamente en una base de datos sobre los clientes que han solicitado créditos y utilizando el aprendizaje automático como medio para esto, se pueden encontrar patrones comunes que nos ayuden a tomar la mejor decisión a la hora de aprobar o rechazar dichos créditos.

El área del aprendizaje automático que se estudiará en el presente trabajo, corresponde a la clasificación supervisada (CS); para el análisis de los datos se utilizarán los árboles de clasificación (también llamados árboles de decisión), los cuales forman parte de las técnicas de clasificación supervisada; se ha

¹ Banco Nacional de Fomento: <http://www.bnf.fin.ec/>

seleccionado los árboles de clasificación por la sencillez del modelo, su accesibilidad a ser representados gráficamente, la explicación que aporta a la clasificación y por su rapidez a la hora de clasificar nuevos patrones. (2)

IV. MARCO TEÓRICO

Clasificación supervisada (CS)

La CS parte de un conjunto de objetos descritos por un vector de características y la clase a la que pertenecen cada uno de ellos; a este conjunto de objetos del que se conoce la clase a la que pertenecen cada uno de ellos se le denomina 'conjunto de entrenamiento'; así basándose en este conjunto de entrenamiento, la clasificación supervisada construye un 'modelo' que se utilizará para clasificar objetos nuevos de los cuales no se sepa su clase. (2)

Árboles de clasificación

"Un clasificador es una partición del espacio de clasificación X en M subconjuntos disjuntos A_1, A_2, \dots, A_M , siendo X la unión de todos ellos y para todo x perteneciente a A_m la clase predicha es C_m ." (2)

Algoritmo j48

El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizados. Se trata de un refinamiento del modelo generado con OneR². Supone una mejora moderada en las prestaciones, y podrá conseguir una probabilidad de acierto

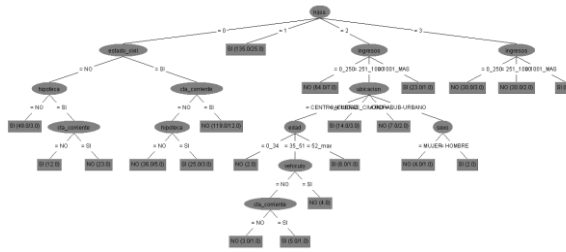
ligeramente superior al del anterior clasificador. (3)

El parámetro más importante que deberemos tener en cuenta es el factor de confianza para la poda "confidence level", que influye en el tamaño y capacidad de predicción del árbol construido. Para cada operación de poda, define la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. A probabilidad menor, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto es del 25%. Según baje este valor, se permiten más operaciones de poda. (3)

El algoritmo J48 se basa en la utilización del criterio ratio de ganancia (gain ratio). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido. (4)

Para tareas de clasificación, el algoritmo J48 resulta muy simple y potente, este algoritmo representa a su vez una evolución del algoritmo ID3. El procedimiento para generar el árbol consiste en seleccionar un atributo como raíz, y crear una rama con cada uno de los valores posibles de dicho atributo; con cada rama resultante se realiza el mismo proceso. En cada nodo se debe seleccionar un atributo para seguir dividiendo, y para ello se selecciona aquel que mejor separe los ejemplos de acuerdo a la clase (5). Se ha seleccionado los árboles de clasificación por la sencillez del modelo, su accesibilidad a ser representados gráficamente, la explicación que aporta a la clasificación y por su rapidez a la hora de clasificar nuevos patrones. (2)

² Es un clasificador de los más sencillos y rápidos. Sus resultados pueden ser muy buenos en comparación con algoritmos mucho más complejos. Selecciona el atributo que mejor "explica" la clase de salida. (3)



Árbol J48.

Cross-Validation

Evaluación con validación cruzada. Se dividirán las instancias en tantas carpetas como indica el parámetro “Folds”, y en cada evaluación se toman las instancias de cada carpeta como datos de test, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados serán el promedio de todas las ejecuciones. (3)

V. SELECCIÓN DE ATRIBUTOS

Los atributos a tener en cuenta para la aprobación del crédito bancario son los siguientes:

- Edad: número de años de la persona quien solicita el préstamo.
- Sexo: puede ser masculino o femenino
- Ubicación: donde se ubica el cliente (centro de la ciudad, afueras, zona rural)
- Ingresos del cliente
- Estado civil: si es soltero o casado
- Hijos: número de hijos que tiene.
- Si posee vehículo
- Si tiene una cuenta de ahorros activa
- Si tiene una cuenta corriente activa
- Si el cliente posee una hipoteca
- Si el préstamo fue aprobado o no.

Los datos almacenados de los clientes se encuentran en una base de datos histórica, la cual posee un registro de 600 personas las cuales han solicitado créditos bancarios, cada persona que se encuentra registrada en esta base de datos posee todos los atributos antes mencionados. Estos atributos fueron obtenidos en base a la entrevista realizada a un representante de la entidad bancaria.

VI. TRABAJOS RELACIONADOS

- Trabajos relacionados con el algoritmo J48

Estudio de técnicas de aprendizaje automático para detectar el fraude hipotecario

Para este proyecto se utilizó un conjunto de datos proporcionado por IndyMac Bank (IMB) para entrenar el modelo que podría aplicarse en general a los datos de la hipoteca y préstamos a los clientes. IndyMac Bank, con sede en California – USA es la segunda entidad privada en cuanto a la concesión de créditos en dicho país.

Se analizaron un total de 43.273 muestras de datos de préstamos. Los datos de cada préstamo contenían atributos como la fecha de aplicación, estado del préstamo, monto del préstamo, etc. Para este proyecto se analizaron distintos algoritmos como redes bayesianas, máquinas de vector soporte (SVM), algoritmo K-NN, y el algoritmo J48. Se utilizaron enfoques tanto supervisados como no supervisados. En este trabajo se concluye que con el aporte de expertos y la selección inteligente de los atributos, es posible predecir con gran precisión un posible

fraude; los algoritmos seleccionados al final fueron el algoritmo IBK (una versión alterna del algoritmo K-NN) y el algoritmo J48. Este proyecto se llevo a cabo con la herramienta WEKA. (6)

- **Trabajos relacionados con la utilización de otros algoritmos**

Modelo de inteligencia artificial Neuro-basado para decisiones sobre préstamos

Proyecto desarrollado por la Universidad de Al-Zaytoonah de Jordania. En este proyecto se desarrollo un modelo que identifica la red neuronal artificial como instrumento propicio para evaluar las solicitudes de crédito para apoyar las decisiones de préstamos en los bancos comerciales de Jordania. Este estudio es uno de los primeros realizados en la región de Jordania que se ocupa de una manera sistemática de la cuestión de utilizar redes neuronales artificiales en las evaluaciones de solicitudes de préstamos. El objetivo de usar redes neuronales artificiales en este proyecto es el de la simplificación del trabajo realizado para un préstamo oficial, para poder controlarlo y para lograr mayor eficiencia y productividad. Se analizaron en total 140 solicitudes de préstamos bancarios, de las cuales, 94 casos se utilizaron para el entrenamiento y 46 se utilizaron en los ensayos (test). Las variables tomadas en cuenta en este proyecto fueron: edad, tipo de cuenta, ingresos, nacionalidad, residencia, tipo de compañía, garante, experiencia laboral, seguro social y tamaño del préstamo. (7)

Decisión de créditos con árboles de decisión impulsados

Proyecto propuesto por la Universidad Técnica de Lisboa en el año del 2008. El objetivo del modelo es de clasificar a los

solicitantes de crédito en dos clases: el crédito bueno y el crédito malo (la clase a la cual se le debe negar el crédito debido a la alta probabilidad de impago); la clasificación depende de las características sociodemográficas del prestatario (tales como edad, escolaridad, ocupación e ingresos), los resultados de amortización de los préstamos anteriores y el tipo de préstamo. Estos modelos también son aplicables a las pequeñas empresas, puesto que pueden ser considerados como extensiones de un cliente individual. Este trabajo propone un modelo de calificación de crédito de los préstamos de consumo basados en árboles de decisión impulsada, una poderosa técnica de aprendizaje en el que se desarrolla un conjunto de árboles de decisión para formar un clasificador dado por una mayoría de votos ponderados de las clasificaciones predicho por los árboles individuales. (8)

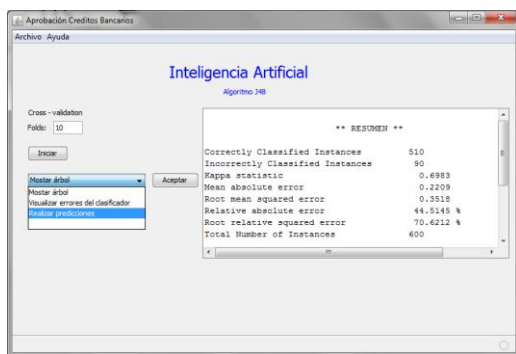
Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación cart

Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART. En este trabajo investigativo se utilizaron datos históricos de clientes para un producto crediticio, del cual ya se conocían los clientes morosos y no morosos; se buscaron patrones de comportamiento de morosidad y de no morosidad para incrementar la precisión al momento de evaluar el crédito para un nuevo cliente. Para ello, se realizó un estudio en el período comprendido entre Agosto del 2001 y Junio del 2002 de todas las solicitudes aprobadas del producto crediticio. Se tomaron las variables almacenadas en la base de datos de la entidad bancaria, que fueron recogidas a través de los formularios de solicitud del crédito, y se seleccionaron para participar

aquellas variables que tenían mayor poder predictivo. (9)

VII. Desarrollo del sistema

El sistema propuesto fue desarrollado en el lenguaje de programación JAVA; para poder utilizar el algoritmo J48 y poder llamar a los archivos con extensión *.arff se trabajó con el API de weka versión 3.6. El sistema consiste en analizar una base de datos de entrenamiento, esta base de datos contiene datos históricos de los clientes que han solicitado créditos al banco. También se contará con una base de datos en la cual se ingresarán los datos del nuevo cliente a ser analizados, estos datos se compararán con el resultado arrojado por el algoritmo a fin de determinar si al nuevo cliente se le concederá o no un crédito bancario. Vale recalcar que el sistema no necesariamente arroja una decisión final en la cual nos podamos basar para aceptar o rechazar un crédito; sino que en base a patrones encontrados y en estadísticas históricas se ofrezca a los gestores de información un valor añadido para reducir la incertidumbre de los resultados de decisiones para poder mejorar el servicio de calidad.



Interfaz de la aplicación.

RESULTADOS:

Tras la ejecución de la base de datos en WEKA, se pudo obtener la siguiente clasificación de las instancias:

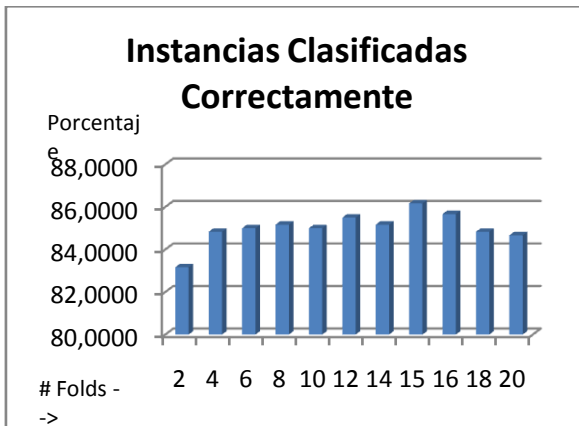
	Etiqueta	Cantidad	Porcentaje
Edad	0 -34	195	32,50
	35 -51	214	35,67
	52 - más	195	32,50
sexo	Hombres	300	50,00
	Mujeres	300	50,00
ubicación	Centro ciudad	269	44,83
	Afuras ciudad	173	28,83
	Rural	96	16,00
	Sub urbano	62	10,33
Ingresos	0 -250	285	47,50
	251 -1000	235	39,17
	1000 - más	80	13,33
Estad civil	Soltero (NO)	204	34,00
	Casado (SI)	394	65,67
Hijos	0	263	43,83
	1	135	22,50
	2	134	22,33
	3	68	11,33
Vehículo	No	304	50,67
	Si	296	49,33
Cta corriente	No	186	31,00
	Si	414	69,00
Cta ahorro	No	145	24,17
	Si	455	75,83
Hipoteca	No	391	65,17
	Si	209	34,83
Aprobado	No	326	54,33
	Si	274	45,67

Como resultado de aplicar el árbol de clasificación J48 en la base de datos de muestra se pudo observar que: De un total de 600 instancias analizadas, 510 fueron clasificadas correctamente, es decir un 85% del total, mientras que 90 instancias fueron clasificadas erróneamente con un total de un 15%. El resultado arrojada por la aplicación muestra la siguiente tabla resumen:

Correctly Classified Instances	510	85	%
Incorrectly Classified Instances	90	15	%
Kappa statistic	0.6983		
Mean absolute error	0.2209		
Root mean squared error	0.3518		
Relative absolute error	44.5145 %		
Root relative squared error	70.6212 %		
Total Number of Instances	600		

Resumen tras haber ejecutado el programa

La figura siguiente muestra como varía el porcentaje de instancias clasificadas correctamente cuando vamos cambiando la variable 'Folds', el valor por defecto que se asigna a Folds es el valor de 10, pero como podemos observar al establecer un valor de 15 el porcentaje de clasificación aumenta.



Lo que se pretende a continuación es aumentar el porcentaje de instancias clasificadas correctamente y reducir las instancias clasificadas incorrectamente.

En la herramienta WEKA hemos detectado los atributos que más influyen en la decisión para la aprobación de un crédito, mediante la selección de atributos se obtuvo la siguiente información:

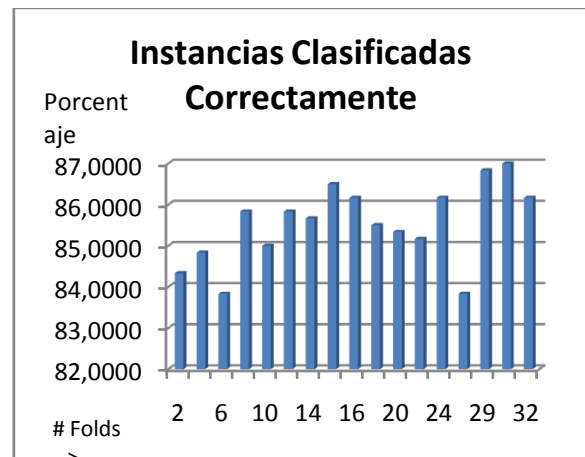
```
Attribute selection output
aprobado
Evaluation mode: 10-fold cross-validation

=== Attribute selection 10 fold cross-validation (stratified),

number of folds (%) attribute
0( 0 %) 1 edad
7( 70 %) 2 sexo
1( 10 %) 3 ubicacion
10(100 %) 4 ingresos
10(100 %) 5 estado_civil
10(100 %) 6 hijos
0( 0 %) 7 vehiculo
0( 0 %) 8 cta_corriente
0( 0 %) 9 cta_ahorros
0( 0 %) 10 hipoteca
```

Como se puede observar, los atributos que más influyen en esta decisión son el número

de hijos, los ingresos y el estado civil de la persona, esto como atributos principales ya que como secundarios están el sexo y la ubicación de la persona. Es sobre estos dos atributos secundarios sobre los que vamos a prestar mayor énfasis a continuación; se ha considerado que para la aprobación de un crédito no se tomará en cuenta el sexo del individuo ni la ubicación del mismo, ya que no por ser hombre o mujer o no vivir dentro de la ciudad tomaremos la decisión de aprobarlo o no. Después de quitar estos dos atributos con el fin de poder mejorar el número de instancias clasificadas correctamente se pudo observar que sí mejoró el porcentaje de instancias clasificadas correctamente, de las 510 (85%) que se obtuvieron inicialmente, ahora se obtuvieron 522 (87%) de instancias clasificadas correctamente, este es el mejor caso que se pudo encontrar realizando pruebas variando el número de Folds en cada corrida, esto se muestra en la siguiente figura:



Correctly Classified Instances	522	87	%
Incorrectly Classified Instances	78	13	%
Kappa statistic	0.7385		
Mean absolute error	0.2132		
Root mean squared error	0.3333		
Relative absolute error	42.964	%	
Root relative squared error	66.9068	%	
Total Number of Instances	600		

Pero el punto que se presenta aquí no es la eliminación de atributos, por lo que asignaremos mayor peso a los atributos con mayor relevancia a la hora de solicitar un crédito bancario.

```

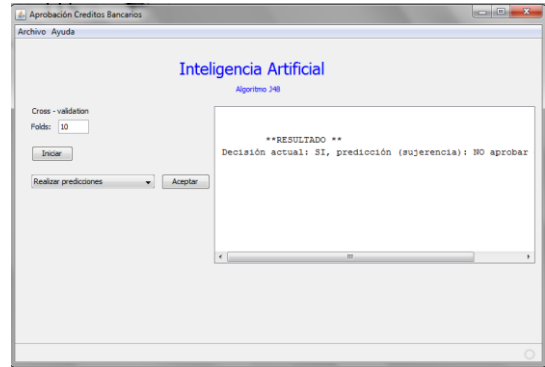
aprobado
Evaluation mode: 10-fold cross-validation

=== Attribute selection 10 fold cross-validation (stratified), seed:
number of folds (%) attribute
10(100 %) 1 edad
9( 90 %) 2 sexo
10(100 %) 3 ubicacion
10(100 %) 4 ingresos
0( 0 %) 5 estado_civil
0( 0 %) 6 hijos
2( 20 %) 7 vehiculo
0( 30 %) 8 cta_corriente
0( 40 %) 9 cta_ahorros
10(100 %) 10 hipoteca

```

Predicción del préstamo al cliente:

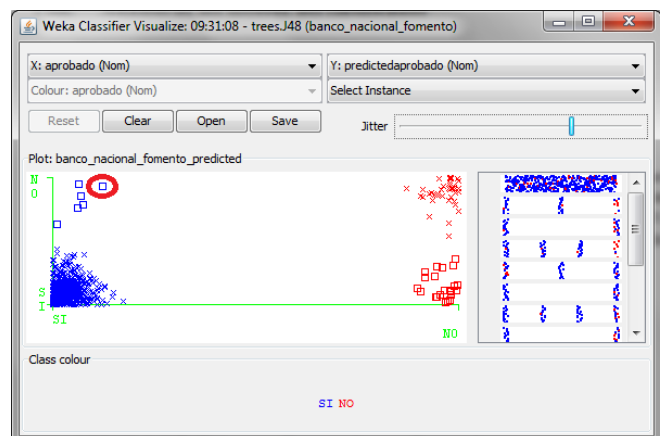
Para determinar si se concede o no el préstamo al cliente, el sistema analizará los 600 casos de la base de datos de entrenamiento, y con los resultados obtenidos se procederá a analizar la base de datos ‘test.arff’, donde se encuentra la información del nuevo caso; es decir, del nuevo cliente quien solicita el crédito. Este método es conocido como ‘supplied test set’, donde se proporciona un nuevo archivo de datos (en formato ARFF y con los mismos atributos) sobre el que se realizará la clasificación. Como resultado de esto el sistema responderá y mostrara la sugerencia para dicho cliente como se muestra en la siguiente figura:



Resultado de la predicción del crédito.

VIII. Análisis de los errores del clasificador

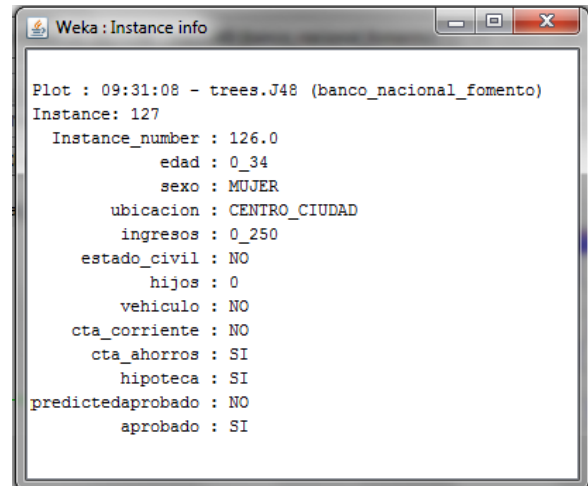
Ahora el algoritmo alcanza un total del 94.9838% de instancias clasificadas correctamente, pero que hay del otro 5.0162 % de las instancias que se clasificaron incorrectamente. WEKA nos ofrece la opción de analizar estos casos de error, después de ejecutar el algoritmo J48 podemos seleccionar la opción de ‘visualize classifier errors’ donde veremos las instancias que se clasificaron incorrectamente. Como se muestra en la figura siguiente, podemos añadir ‘ruido’ a dicho resultado, entonces la clasificación de aprobación vendrá dada como se muestra a continuación:



Visualizador de errores de WEKA.

En la intersección del eje X con el eje Y se sitúan las instancias de cuyos clientes si se les aprobó el crédito, mientras que en la parte superior derecha se muestran los clientes cuyos créditos no fueron aprobados; esta clasificación corresponde a las instancias clasificadas correctamente. Pero los errores que se dan, es decir las instancias clasificadas incorrectamente se muestran en la figura a manera de '□' (un cuadrado) como se puede observar. Los de la parte superior derecha corresponde a los falsos positivos, es decir aquellos créditos que fueron aprobados por el sistema cuando en realidad no debieron ser aprobados; por otro lado tenemos en la esquina inferior derecha los falsos negativos, es decir aquellos créditos que fueron rechazados por el sistema cuando en realidad sí debían ser concedidos.

Por ejemplo en la figura siguiente tenemos un falso positivo, que como ya se menciona es un crédito el cual fue aprobado por el sistema cuando en realidad tendría que reprobárselo; como se observa en la figura la mujer quien solicita el crédito es una mujer joven, con un salario entre 0 – 250 dólares, soltera, no posee vehículo ni cuenta corriente, tiene una cuenta de ahorros y tiene una hipoteca que pagar; la decisión arrojada fue 'SI' aprobar cuando el sistema tendría que negar dicha solicitud de crédito.



Información de un error encontrado.

IX. Comparación con el algoritmo de clasificación ID3

La presente sección tiene como fin el de comparar el algoritmo de clasificación aquí propuesto (J48) con el algoritmo de clasificación ID3, el cual también es parte de los árboles de clasificación. Como se mencionó en la sección IV, el algoritmo J48 representa una evolución del algoritmo ID3; esto lo podemos notar en las pruebas que se realizaron sobre los mismos datos con el uso de ambos algoritmos.

	Algoritmo	
	J48	ID3
Tiempo de construcción del modelo	0.01 segundos	0.03 segundos
Instancias clasificadas correctamente	87%	77,1667
Instancias clasificadas incorrectamente	13%	20.5 %
Genera árbol	Si	No
Instancias no clasificadas	No	Si

Comparación entre algoritmos J48 e ID3.

Como se puede observar en la tabla anterior, el algoritmo J48 presenta una superioridad notable frente al algoritmo ID3; estas pruebas fueron realizadas sobre la misma base de datos. Como punto de partida se puede observar que el tiempo para la construcción

de modelo es mucho menor en el algoritmo J48 con respecto al ID3, además se puede observar que el total de instancias clasificadas correctamente es superior en el algoritmo J48, en el algoritmo J48 podemos observar el árbol gráficamente y no se dan instancias no clasificadas mientras que en algoritmo ID3 es todo lo contrario.

X. Conclusiones

- Mas que un sistema que nos arroje una decisión final en la cual nos podamos basar para aceptar o rechazar un crédito; el sistema ofrece un valor añadido para reducir la incertidumbre de los resultados de decisiones para poder mejorar el servicio de calidad.
- El algoritmo utilizado (J48) resulta muy conveniente a la hora de analizar este tipo de problemas ya que permite construir modelos fáciles de interpretar.
- Mediante el uso de weka podemos determinar la manera de como la base de aprendizaje se va alimentando y puede sacar los resultados una de manera eficiente.
- Analizando los resultados arrojados por weka se pudo observar que los atributos más importantes a la hora de solicitar créditos bancarios es el del número de hijos, siendo muy altas las probabilidades de un crédito al contar con un solo hijo.
- Analizando los mismos datos con los algoritmos J48 e ID3 respectivamente, se pudo determinar que el algoritmo J48 sí representa una evolución del algoritmo ID3 como se afirma en la sección IV.
- La utilización de los árboles de decisión nos permiten determinar patrones de comportamiento en los datos de la muestra analizada.

XI. Referencias

- [1] Banco Nacional de Fomento. [En línea] Mayo de 2010. Disponible en <<http://www.bnf.fin.ec/>>.
- [2] **Sierra Araujo, Basilio.** Aprendizaje Automático: conceptos básicos y avanzados. Aspectos piráticos utilizando el software WEKA. s.l. : Pearson, Prentice Hal, 2006.
- [3] **García Jiménez, María y Álvarez Sierra, Aránzazu.** Análisis de Datos en WEKA – Pruebas de Selectividad. 2010. Disponible en <<http://www.it.uc3m.es/jvillena/irc/practicass/06-07/28.pdf>>
- [4] **Collada Pérez, Sonia y Gálvez Carranza, Rubén.** Clasificación de e-mails: Detección de Spam. 2010. Disponible en <<http://www.it.uc3m.es/jvillena/irc/practicass/07-08/DeteccionSpam.pdf>>
- [5] Minería de datos aplicada a la Formación de Equipos de Proyectos de Software. André Apuero, Margarita, Baldaquín, María Gulnara y Muñoz Castillo, Vanesa D. 121, La Habana : AHCNET, 2010. Disponible en <<http://www1.ahcnet.net/actualidad/revista/Documents/rev121.swf>>
- [6] **Liu, Chuping, y otros.** A study of machine learning techniques to detect Mortgage. California : University of Southern California, 2007. Disponible en <<http://www.scf.usc.edu/~rizwankh/rizz/567.pdf>>
- [7] **Fathi Eletter, Shorouq, Ghaleb Yaseen, Saad y Awad Elrefae, Ghaleb.** Neuro-Based Artificial Intelligence Model for Loan Decisions. Amman : American Journal, 2010. ISSN 1945-5488. Disponible en <<http://www.scipub.org/fulltext/ajeba/ajeba2127-34.pdf>>
- [8] **Bastos, Joao.** Credit scoring with boosted decision trees. Lisboa : s.n., 2008. Disponible en <http://mpra.ub.uni-muenchen.de/8156/1/MPRA_paper_8156.pdf>

[9] PATRONES DE MOROSIDAD PARA UN PRODUCTO CREDITICIO USANDO LA TÉCNICA DEL ÁRBOL DE CLASIFICACIÓN CART. **Salinas Flores, Jesús**. s.l. : Revista de la Facultad de Ingeniería

Industrial, 2005, Vol. 8.1. ISSN: 1810-9993.

Disponible en

<<http://revistas.concytec.gob.pe/pdf/id/v8n1/a06v8n1.pdf>>