



UNIVERSIDAD DE ESPECIALIDADES ESPÍRITU SANTO

FACULTAD DE ECONOMÍA Y CIENCIAS EMPRESARIALES

TÍTULO:

***ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS***

**TRABAJO DE TITULACIÓN QUE SE PRESENTA COMO
REQUISITO PREVIO A OPTAR EL GRADO DE INGENIERÍA EN
CIENCIAS EMPRESARIALES CON MENCIÓN EN FINANZAS**

NOMBRE DEL ESTUDIANTE:

ANDRES MANUEL AVILES VARGAS

NOMBRE DEL TUTOR:

Ing. Billy Andrade García, MBA

SAMBORONDÓN, DICIEMBRE, 2016

***ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS***

**Análisis comparativo de modelos de puntuación de crédito generados por la
minería de datos en instituciones financieras**

Andrés Avilés Vargas

Universidad de Especialidades Espíritu Santo – Ecuador, amaviles@uees.edu.ec,
Facultad de Economía y Ciencias Empresariales, Edificio F, Universidad Espíritu
Santo, Km 2.5 Vía Puntilla Samborondón.

Resumen

El presente trabajo de investigación tiene como propósito analizar los modelos de puntuación de crédito generados por la minería de datos en instituciones financieras. Para esto se comparó y analizó los modelos y las metodologías utilizados con información bancaria real en 10 artículos científicos. La metodología exploratoria utilizada permitió observar y describir la variable más influyente en la construcción de los modelos, a su vez como la variable menos influyente y la confiabilidad estadística de predicción. Los resultados permitieron la elaboración de un cuadro comparativo con el fin de determinar las características de los modelos con mayor capacidad de predicción.

Palabras claves: Modelos de Puntuación de Crédito, Minería de Datos.
Redes Neuronales Artificiales, Regresión Logística, Banca.

Abstract

The purpose of this research is to analyze the credit scoring models generated by data mining in financial institutions. For this it was compared and analyzed the models and methodologies used with real bank information in 10 scientific articles. The exploratory methodology used allowed to observe and describe the most influential variable in the construction of the models, in turn as the least

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

influential variable and the predictive statistical reliability. The results allowed the development of a comparative table to determine the characteristics of the models with greater predictive capacity.

Keywords: Credit Scoring Model, Data Mining, Artificial Neural Networks. Logistic Regression, Banking.

Introducción

El sector bancario a nivel global se ha visto deteriorado en el último año. El declive de los precios del petróleo y las materias primas ha sostenido eminentes los riesgos en las economías de mercados emergentes, mientras que la mayor incertidumbre en torno a la transición del modelo de crecimiento de China ha aumentado los efectos de contagio en los mercados mundiales. Estos acontecimientos han implicado un endurecimiento de las condiciones financieras, la reducción del apetito del riesgo y el incremento de los riesgos de crédito, y han entorpecido el saneamiento de los balances, debilitando así la estabilidad financiera (Fondo Monetario Internacional , 2016).

En la actualidad, el sector bancario ecuatoriano sigue las tendencias de declive a nivel global de la industria. De hecho, entre junio del 2015 y junio del 2016 la banca privada del Ecuador registró utilidades por 97,89 millones de dólares según las cifras publicadas por la (Superintendencia de Bancos, 2016), un monto que representa una caída del 38% en comparación a las utilidades del año anterior. De acuerdo al mismo informe la banca privada del Ecuador vio un incremento de solvencia y liquidez, a su vez de una mejora de eficiencia en el uso de gastos operativos. Por otra parte, vio en decadencia el porcentaje de activos

Comentado [U1]: Corregir sangría. Quitar.

***ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS***

productivos, los niveles de rentabilidad, de cobertura y de intermediación; y en concordancia con la tendencia global, un alza en la tasa de morosidad.

Los bancos comerciales en el Ecuador son considerados como entidades financieras competitivas y de vital importancia. En el fin de proporcionar varios servicios financieros a los clientes, los bancos gestionan diferentes tipos de riesgos. De acuerdo con Arunkumar (2006) “la toma de riesgo es visto a menudo como el conductor básico para el comportamiento financiero y la rentabilidad”. Sin embargo, la aprobación de créditos en los bancos ecuatorianos es de naturaleza subjetiva. Esto implica que se revisa manualmente cada solicitud de préstamo, imponiendo sesgos, incluyendo conocimientos personales e intuición del gestor de crédito. Este método ha sido remplazado en bancos internacionales por los modelos de puntuación de crédito, o un híbrido de las dos metodologías, subjetiva y objetiva para la toma de decisión adecuada. Los modelos de puntuación de crédito permiten a los gerentes de bancos o gestores de crédito identificar objetivamente las cuentas que probablemente sean solventes o buenos riesgos crediticios de las aquellas con probabilidad de incumplimiento.

Hoy en día, el futuro de la industria bancaria depende en gran medida de la dinámica de gestión de riesgos (Bekhet, 2014). De hecho, según Huang (2007) “a menudo, los gerentes de bancos se enfrentan al problema de intentar aumentar el volumen de crédito mientras reducen la probabilidad de incumplimiento”. Debido a esto las instituciones bancarias buscan herramientas de gestión de riesgos más eficientes y modelos de apoyo a las decisiones, complementados con herramientas analíticas para sobrevivir en entornos empresariales inciertos y cambiantes.

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

Por otra parte, desde la década de 1990 el concepto de la banca se ha desplazado a bases de datos centralizadas, transacciones en línea y cajeros automáticos de manera global. Los datos según Moin & Ahmed (2012) “podrían ser uno de los recursos más valiosos de cualquier banco, pero solo si se sabe cómo exponer un valioso conocimiento oculto de los datos rústicos”. La minería de datos permite extraer dicho conocimiento a partir de datos históricos con el fin de predecir situaciones futuras mediante la elaboración de modelos predictivos.

El presente trabajo tiene como objetivo elaborar un análisis sobre las aplicaciones de los modelos de puntuación de crédito generados por la minería de datos en instituciones bancarias a nivel global. Con el fin de realizar un aporte empírico a la literatura acerca de la relación entre los modelos generados por la minería de datos y el sector bancario. A través de este estudio se pretende comparar diferentes modelos de puntuación de crédito generados por la minería de datos publicados en artículos científicos. Con el propósito de analizar las variables más significativas que influyen en la aprobación de las decisiones de préstamos individuales, las variables menos influyentes en la toma de decisión y la confiabilidad estadística del poder de predicción generado por estos modelos.

Marco Teórico

Minería de datos

La minería de datos se refiere a la extracción de conocimientos provenientes de grandes cantidades de datos. Los datos pueden ser categorizados como datos espaciales, datos de series temporales, datos multimedia, datos web y datos de texto. El proceso consiste en extraer patrones interesantes, previamente

Comentado [U2]: A los títulos se les puede quitar la sangría

Comentado [U3]: Estos títulos se puede subrayar.

***ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS***

desconocidos, implícitos y potencialmente útiles (Moin & Ahmed, 2012). Utilizando la información contenida en el almacén de datos, la minería de datos intenta proporcionar respuestas a preguntas sobre una organización que un tomador de datos no había pensado con anticipación.

Varias tendencias recientes han incrementado el interés por la minería de datos, debido a la disminución del costo de almacenamiento de datos y la creciente facilidad de recolección de datos. Con mayores capacidades de almacenamiento de datos y costos decadentes la minería de datos ha brindado a las organizaciones una nueva representación de hacer negocios.

De acuerdo con Provost & Fawcett (2013) “la minería de datos es un arte”. Como todo arte maduro existe un proceso bien comprendido que sitúa una estructura sobre el problema, permitiendo consistencia, repetitividad y objetividad razonable. Una codificación útil con respecto al proceso es dada por el proceso estándar de industria cruzada para la minería de datos o mejor conocido como (CRISP-DM) por sus siglas en ingles (Kurgan, 2006). El cual está ilustrado en la figura 1.1.

ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

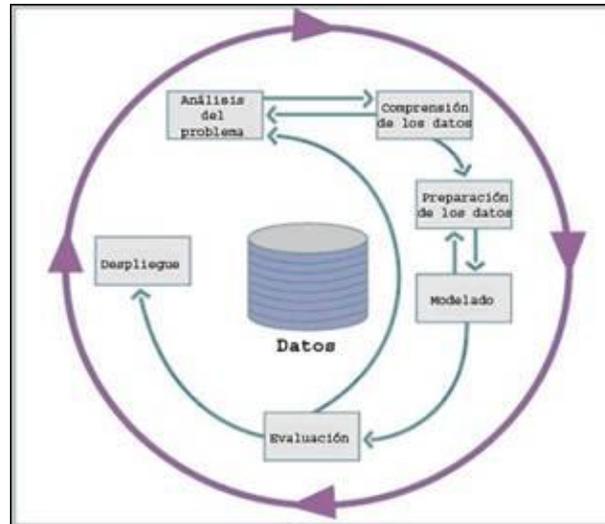


Figura 1.1: Proceso estándar de industria cruzada para la minería de datos (CRISP-DM).

Comentado [U4]: Fuente

Fuente: De *Data Science for business*, por F. Provost y T. Fawcett, 2013 O'Reilly, p.27. Copyright 2013 Foster Provost and Tom Fawcett.

Este diagrama de proceso hace explícito el hecho de que la iteración es la regla y no la excepción. Pasar por el proceso una vez sin haber resuelto el problema es, en términos generales, no un fracaso. Frecuentemente todo el proceso es una exploración de los datos, y la primera iteración que proporcionan los científicos de datos resulta fructífera. Incrementando el conocimiento a medida que se incrementan las iteraciones.

Modelos predictivos

De acuerdo con Koh & Gerry (2002) “las aplicaciones más comunes e importantes de la minería de datos involucran la predicción”. Un modelo predictivo es un modelo que utiliza diferentes variables o atributos basados en la

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

base de datos para determinar y predecir un atributo en específico (Bradley, 1999). Tal y como la experiencia de aprendizaje de un ser humano, las observaciones son usadas para crear un modelo que subraya las características de una ocurrencia.

En la aplicación los modelos predictivos resuelven problemas que se diferencian de dos maneras. Si la variable a predecir es de carácter numérico o continuo, se resuelve por medio de la regresión. Mientras que, si la variable a predecir es de carácter discreto, se resuelve por medio de la **clasificación**. Entre las técnicas que se utilizan para realizar modelos predictivos supervisados constan las redes neuronales artificiales, regresiones lineales, regresión logística, arboles de decisión y vectores máquina de soporte.

Comentado [U5]: ¿Qué técnicas se utiliza para realizar modelos predictivos?

Técnicas de minería de datos para generar modelos predictivos

Redes Neuronales Artificiales

Las redes neuronales artificiales se consideran una poderosa alternativa a los métodos convencionales de pronóstico y clasificación debido su capacidad de capturar relaciones no lineales y complejas (Bekhet, 2014). Según Cao & Parry (2009) “estos modelos tienen una capacidad biológicamente inspirada que imita las capacidades de procesamiento del cerebro humano”.

La percepción multicapa o (MLP) por sus siglas en ingles es el modelo de red neuronal *feed-forward* más popular, utilizado frecuentemente en el reconocimiento de patrones. Un modelo típico de red neuronal *feed-forward* está compuesto por una arquitectura de tres capas: capa de entrada, capa oculta y capa de salida, como se expresa en la figura 1.2. La capa de entrada se encarga de alimentar las variables de entrada o de carácter predictor a la capa siguiente. Cada

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

neurona oculta recibe una suma ponderada de todas las entradas en la capa de entrada, mientras se aplica una función de transferencia como: log sigmoide, tangente hiperbólica, *soft max* a la suma ponderada. De igual forma, cada neurona oculta transfiere un resultado ponderado a la neurona de salida, es decir cada neurona variable dependiente (Cao & Parry, 2009). El resultado de la neurona de salida es la solución del problema planteado.

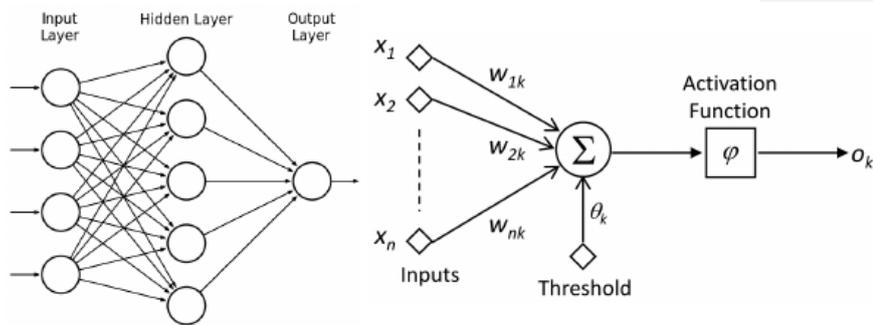


Figura 1.2: Arquitectura de red neuronal multicapa feed-forward

Comentado [U6]: Fuente

Fuente: De *Artificial Neural Networks for RF and Microwave Design—From Theory to Practice*, por Qi-Jun Zhang y Kuldip C. Gupta, 2003, IEEE transactions on microwave theory and techniques, p.1341. Copyright 2003 IEEE Microwave Theory and Techniques Society.

En la actualidad las redes neuronales artificiales se han utilizado con éxito en aplicaciones financieras como: puntuación de crédito, predicción de fracaso corporativo y calificación de bonos. Así como también tareas como predicción de movimientos de precios de acciones y fluctuaciones de tipos de cambio (Bekhet, 2014).

**ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACIÓN DE
CRÉDITO GENERADOS POR LA MINERÍA DE DATOS EN
INSTITUCIONES FINANCIERAS**

Arboles de decisión

Un árbol de decisión o también conocido como árbol de clasificación es un esquema de clasificación que genera un árbol y un conjunto de reglas, que representan el modelo de diferentes clases, a partir de un conjunto de datos dado (Chitra & Subashini, 2011). El nodo principal o conocido como nodo raíz, en el cual se toma la decisión inicial desencadena en nodos internos o ramas de árbol. Específicamente cada rama del árbol de decisión es una pregunta de clasificación, mientras que las hojas del árbol representan las particiones del conjunto de datos o individuos con su clasificación y respectiva probabilidad de efecto. El conjunto de datos disponibles para el desarrollo de métodos de clasificación se divide generalmente en dos conjuntos disjuntos, un conjunto de entrenamiento y un conjunto de prueba. El primero indicado es usado para derivar el clasificador, mientras que el último se utiliza para medir la precisión del modelo clasificador.

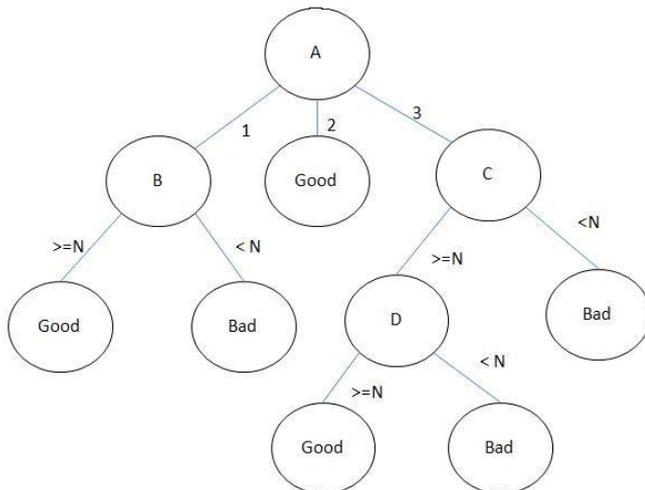


Figura 1.3: Árbol de clasificación o decisión.

Comentado [U7]: Clasificación o decisión?

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

Fuente: De *Data Science for business*, por F. Provost y T. Fawcett, 2013

O'Reilly, p.64. Copyright 2013 Foster Provost and Tom Fawcett.

El árbol de decisión representa un mapeo de las observaciones sobre un elemento a la conclusión sobre su valor objetivo como un modelo predictivo en la minería de datos y el aprendizaje automático (Zhang, X., C., & Zheng, 2010). Esta técnica mira todas las posibles preguntas que podrían separar el conjunto de datos de formación original en segmentos que son casi similares con respecto a las diferentes clases que se predice.

La ventaja de este método es que es un modelo fácil de entender y explicar cómo se expresa en la figura 1.3, sin embargo, la limitación de este modelo es que, no puede generalizarse en una estructura diseñada para todos los contextos.

Regresión Logística o Análisis discriminatorio

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable dependiente dicotómica basada en una o más variables categóricas. En lugar de ajustar los datos a una línea recta, la regresión logística utiliza una curva logística (Chitra & Subashini, 2013). Con el fin de estimar la probabilidad de que acontezca un suceso. A diferencia de la regresión lineal que utiliza el proceso de regresión para predecir un valor numérico de la variable dependiente.

En el sector bancario se han utilizado la regresión logística, la regresión lineal, programación lineal y árboles de clasificación para desarrollar sistemas de puntuación de crédito (Dong, K., & Yen, 2010). Entre las cuales la regresión logística sobresale. Sin embargo, la capacidad de predicción de esta técnica es

Comentado [U8]: O análisis discriminatorio

Comentado [U9]: Agregar: Redes Neuronales y Análisis Discriminante.

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

inferior a otros métodos como redes neuronales artificiales y máquinas de soporte vectorial.

Metodología

Para efectos de este estudio, se realizó una investigación bibliográfica de 10 artículos científicos donde se han aplicado técnicas estadísticas multi-variantes lineares y no lineares mediante la minería de datos con el fin de elaborar modelos de puntuación de crédito. Los datos fueron extraídos de la sección de metodología y resultados de los 10 artículos científicos previamente mencionados. Se empleó una metodología exploratoria debido a que se observó y describió las variables más significativas que influyen en la aprobación de créditos individuales, las variables menos influyentes y la confiabilidad estadística de predicción de acuerdo a estos modelos. El método utilizado fue el cuantitativo debido a que se analizó la sección de resultados de los artículos científicos.

Se analizó individualmente la estructura de cada modelo en detalle y el proceso implementado. Los datos fueron analizados e ingresados en una hoja de cálculo donde se detalla, la técnica estadística utilizada, la variable más influyente del modelo, la menos influyente y la confiabilidad estadística del poder de predicción del modelo. Así mismo, se elaboró un análisis comparativo con los resultados.

Resultados

Análisis de modelo de puntuación de crédito utilizado en artículo científico #1

El siguiente modelo de puntuación de crédito se elaboró utilizando un conjunto de datos que incluía aplicaciones de crédito aceptadas tanto como aplicaciones de crédito rechazadas, por parte de diferentes bancos jordanos en el

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

periodo 2006-2011. El conjunto de datos utilizado consto de 492 casos. Entre los cuales 292 datos eran créditos aceptados y 200 rechazados.

El conjunto de datos contuvo un total de 12 diferentes variables, tanto categóricas como numéricas. El subconjunto de datos se dividió aleatoriamente en SPSS, dividiendo la información en 359 casos para entrenamiento, 64 para validación y 69 para prueba de poder de predicción. Además, se categorizó como 0 a las aplicaciones rechazadas y 1 a las aceptadas.

El modelo utilizó como método de clasificación una red neuronal artificial multicapa *feed-forward*. En el cual se utilizó la versión 18 del software SPSS para determinar la arquitectura de la red neuronal. Obteniendo como resultado 12 variables independientes en la primera capa, 9 neuronas ocultas con una función hiperbólica con el fin de entrenar la información. Finalizando en la capa final con una neurona de salida, en este caso la variable dependiente, activada con la función *softmax* del SPSS software. A través de este proceso la información se entrena con el fin de encontrar el patrón de aplicaciones de crédito aceptadas al igual que el de las rechazadas. Por otra parte, con el fin de determinar que variable es la más óptima para el proceso de aprendizaje, se utilizó un algoritmo de optimización de pendiente de gradiente para estimar los pesos sinápticos de la red neuronal con la tasa de aprendizaje.

La figura #2 muestra la clasificación obtenida por la red neuronal, indicando el porcentaje de acierto de clasificación que realizó el modelo.

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

Muestra	Observacion	Prediccion		
		0	1	Porcentaje correcto
Entrenamiento	0	124	14	89,9%
	1	11	210	95,0%
Total				93,0%
Prueba	0	21	3	87,5%
	1	3	42	93,3%
Total				91,3%
Validacion	0	29	9	76,3%
	1	2	24	92,3%
Total				82,8%

Figura #2: Resultados probabilisticos de paper#1

Fuente: Credit Risk Management for the Jordanian Commercial Banks: A Business Intelligence Approach, por Bekhet, H. A. y Eletter, S. F. K, (2014), Review of Development Finance, p.26. Copyright 2014 Africagrowth Institute.

Los resultados revelan que el verdadero poder predictivo del modelo es de 91,3% logrando clasificar correctamente 87,5% de las aplicaciones rechazadas y 93,3% de las aceptadas. Mientras que el estudio revela que la variable con mayor importancia en el modelo para la toma de decisión del crédito es la ratio de deuda/ingreso. Relegando a su vez al género como la menos importante de acuerdo al modelo.

Un modelo con mucha aplicabilidad en otras instituciones bancarias considerando que 75% de eficacia de predicción es ya aceptable.

Análisis de modelo de puntuación de crédito utilizado en artículo científico #2

Este modelo de puntuación de crédito se basó en 497 casos de créditos aprobados y créditos no aprobados en el periodo 2006-2011 de un banco comercial iraní. El modelo utilizó 18 variables independientes tanto categóricas como numéricas. Del subconjunto de datos se tomó un 70% para el entrenamiento de la información. Mientras que solo de 90 casos un 20% de validación y solo un

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

10%. Un proceso de validación un tanto cuestionable. Por otra parte, se asignó un 0 para categorizar a los clientes aceptados y un 1 a los rechazados.

La clasificación fue realizada por una red neuronal artificial multicapa *feed-forward*. La cual empleo la función de tangente hiperbólica en la capa neuronal oculta y una función softmax en la capa resultante o final.

La figura #3 revela los resultados obtenidos por la red neuronal y el respectivo porcentaje de acierto de clasificación.

Muestra	Observacion	Prediccion		
		0	1	Porcentaje correcto
Entrenamiento	0	209	31	87,1%
	1	14	119	89,5%
Total				87,9%
Prueba	0	3	1	75,0%
	1	0	5	100,0%
Total				88,9%
Validacion	0	4	1	80,0%
	1	1	3	75,0%
Total				77,8%

Figura #3: Resultados probabilísticos de clasificación

Fuente: *Measuring Credit Risk of Bank Customers Using Artificial Neural Network*, por Nazari, M. y Alidadi, M., (2013), *Journal of Management Research*, p.23. Copyright 2013 Nazari, M., & Alidadi, M.

El modelo reveló un 88,9% de poder predictivo de los casos para los casos. Clasificando correctamente el 75% de los créditos aceptados o buenos créditos y un 100% de los créditos rechazados. Además, el modelo indicó que la frecuencia de crédito individual es el mayor influyente para la toma de decisión del modelo, mientras que estatus de la cuenta del cliente en el banco es la menos importante. Sin embargo, este modelo cuenta la falencia de la pequeña muestra efectuada como prueba para revelar el poder predictivo del modelo. Por lo que los resultados deberían ser tomados con cautela.

**ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

Análisis de modelo de puntuación de crédito utilizado en artículo científico #3

El conjunto de datos para la elaboración de este modelo se basó en 633 créditos entregados por parte de un banco comercial tunecino en el periodo 2010-2012. Categorizando con un 1 a los clientes que no pagaron sus deudas a tiempo y un 0 a los que lo hicieron correctamente. Además, el modelo empleó 4 variables para realizar la clasificación: edad, monto, historial de crédito y ocupación. Con el fin de determinar si el cliente entra en default o no.

Comentado [U10]: Dos puntos seguidos “:”

El modelo utilizó la regresión logística. Dado que la regresión logística es solamente utilizada con gran cantidad de datos fue necesario observar la ausencia de multi-colinealidad por parte del modelo, que no se presentó debido al uso de solo 4 variables exploratorias independientes.

La figura #4 revela los resultados obtenidos por el modelo, indicando el poder predictivo del mismo.

	Default	No-default	Total
Default	364	4	368
No-default	2	263	265
Total	366	267	633

Figura #4: Matriz de clasificación de observaciones

Fuente: *The Consumer Loan's Payment Default Predictive Model: An Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank*, por Abid, L., Masmoudi, A. y Zouari-Ghorbel, S., (2016), *Asian Economic and Financial Review*, p.34. Copyright 2016 Springer Science+Business Media New York.

El modelo clasificó correctamente 89% de los préstamos, indicando una sensibilidad de 99.41% y una especificidad de 98.47%. Revelando un alto poder predictivo por parte del modelo, sin importar su sencillez constructiva. Además, el

Comentado [U11]: clasificó

ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

modelo **mostró** al historial de crédito del individuo como la variable más influyente y significativa, mientras que la edad **resultó** como la de menos importancia.

Comentado [U12]: mostró

Comentado [U13]: resultó

Análisis de modelo de puntuación de crédito utilizado en artículo científico #4

Este modelo se basó en **82,093** créditos recibidos por clientes de un banco de Samán en el periodo 2000-2008. El modelo **empleó** 20 variables independientes para clasificar correctamente los créditos aceptados y los rechazados. Las variables utilizadas fueron: Edad, antecedentes, año de entrada al banco, ciudad, tipo de préstamo, año de inicio de préstamo, año final del préstamo, tiempo del contrato, genero, educación, ocupación, estatus marital o civil, tipo de colateral, método de pago, interés, gastos personales, monto del préstamo.

Comentado [U14]: en

Comentado [U15]: empleó

El modelo utilizó una red neuronal multicapa *feed-forward*, de 2, 3 y 4 capas. Utilizando diferentes estrategias para la construcción de las redes. En total se generaron 61 neuronas de entradas, procesadas por una función hiperbólica en la capa oculta y la función softmax en las 5 neuronas de salida representadas como activas, deudas dudosas, sobresaliente, atrasado y refinanciado.

La figura #5 revela la confiabilidad estadística obtenida por las diferentes estructuras de redes neuronales, mientras que la figura #5.1 revela los conjuntos de datos distribuidos en las 5 neuronas resultantes.

Estrategia utilizada	Numero de capas	Probabilidad de acierto en entrenamiento	Probabilidad de acierto en validacion	Probabilidad de acierto en prueba
Rapida	2	85,64%	85,77%	86,15%
Rapida	3	86,36%	86,33%	86,71%
Rapida	4	85,55%	84,70%	85,78%
Dinamica	4	85,97%	85,47%	85,98%
Multiple	4	86,16%	86,20%	86,46%

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

Figura #5: Probabilidades de acierto de modelos generados

Fuente: *A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks*, por Yaghini, M., Zhiyan, T. y Fallahi, M., (2011), *Int'l Conf. Data Mining*, p.5. Copyright 2011 Yaghini, M., Zhiyan, T. y Fallahi, M.

Estado	Predicciones				
	Activo	Deuda dudosa	Sobresaliente	Atrasado	Refinanciado
Activo	7032	0	0	0	5291
Deuda dudosa	0	3518	331	513	1
Sobresaliente	1	770	732	1227	1
Atrasado	1	495	351	1435	0
Refinanciado	2131	0	0	0	57975

Figura #5.1: Resultados matriz de clasificación

Fuente: *A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks*, por Yaghini, M., Zhiyan, T. y Fallahi, M., (2011), *Int'l Conf. Data Mining*, p.5. Copyright 2011 Yaghini, M., Zhiyan, T. y Fallahi, M.

El modelo que generó la mayor capacidad de predicción fue la multicapa de 3 capas con una estrategia rápida. Esto es debido a que la estrategia rápida requiere más memoria y más tiempo para ejecutar los algoritmos funcionales. Además, el modelo considero el historial de crédito como la variable de mayor influencia, mientras que la edad demostró ser irrelevante y menor significancia en la elaboración del modelo.

Análisis de modelo de puntuación de crédito utilizado en artículo científico #5

Este modelo se basó en 316 casos de préstamos aceptados y rechazados por parte de un banco iraní. A parte un subconjunto de 90 casos se utilizó para realizar la prueba de poder predictivo del modelo. Se utilizó 5 variables independientes exploratorias con el fin de predecir si el cliente es sujeto de crédito o debe ser rechazado. Las variables consideradas fueron: tasa de interés, periodo

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

de amortización del préstamo, historial del cliente con el banco, tiempo de retraso en el pago del préstamo y la industria en la que labora el cliente.

El método considerado para la clasificación fue la regresión logística empleada en la versión 18 del software SPSS. La figura #6 detalla los resultados obtenidos por el modelo.

Observaciones	Predicciones		
	Mal cliente	Buen cliente	Porcentaje correcto
Mal cliente	41	6	87,2%
Buen cliente	4	39	90,7%
Total			88,9%

Figura #6: Resultados probabilísticos de clasificación de clientes

Fuente: *Credit Risk Modeling for Commercial Banks*, por Karimi, A., (2014), *International Journal of Academic Research in Accounting, Finance and Management Sciences*, p.189. Copyright 2014 HRMARS

El modelo cuenta con un poder de predicción de 88.9%. Clasificando correctamente un 87.2% a los malos clientes o clientes rechazados. Y un 90.7% de exactitud hacia los buenos clientes. Es importante resaltar la importancia de clasificar correctamente a los mal clientes por sobre los buenos, debido al costo implícito o riesgo por parte del banco al clasificar un cliente como bueno en diferencia de negarle el crédito a un cliente bueno. Además, el modelo reveló que la tasa de interés fue la más significativa para el resultado obtenido y el tiempo de retraso de un préstamo la menos indicativa.

Análisis de modelo de puntuación de crédito utilizado en artículo científico #6

Este modelo se construyó con 302 casos de créditos sobresalientes y créditos en default, provistos por una cooperativa de crédito de Brasil en los periodos 2003-2007. El modelo utilizó 26 variables numéricas y categóricas.

Comentado [U16]: utilizó

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

El modelo utilizó un algoritmo C4.5 de árbol de decisión en conjunto con la implementación de la herramienta J4.8 existentes en el software WEKA. Generando 41 hojas de las reglas creadas por el algoritmo.

La figura #7 muestra el subconjunto de datos puestos a prueba para determinar la capacidad de predicción del modelo.

Observaciones	Predicciones		
	Default	Sobresalientes	Probabilidad correcta
Default	121	11	91,67%
Sobresalientes	8	181	95,77%
Total			94,08%

Figura #7: Resultado de clasificación de clientes

Fuente: *Credit analysis using data mining: application in the case of a credit unión, por Sousa, M. D. M. y Figueiredo, R. S, (2014), JISTEM-Journal of Information Systems and Technology Management, p.391. Copyright 2014 TECSI FEA USP.*

El árbol de clasificación logro un poder predictivo de 94.08%. Clasificando correctamente el 91.67% de los defaults. Mientras que un 95.77% de los sobresalientes. Además, el modelo revelo que la garantía de liquidez por parte del sujeto de crédito representa la variable con mayor significancia para no entrar en default, mientras que el lugar de nacimiento es la de menor significancia en el modelo.

Análisis de modelo de puntuación de crédito utilizando el artículo científico #7

Este modelo tuvo bases utilizando 924 casos de créditos otorgados a clientes en el periodo 2003-2006 por parte de un banco comercial tunecino. Se utilizaron 24 variables independientes financieras y no financieras. Mientras que

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

el resultado se categorizo como 0 si el cliente pagaba a tiempo su préstamo y 1 si no lo hacía.

El modelo utilizó el método de K-vecinos cercanos, el cual es un método clasificador que utiliza diferentes valores como K o constantes para determinar la probabilidad de que un elemento pertenezca a una clase. En este caso el modelo optó por 4 valores diferentes de K (2, 3, 4, 5). Generando tres modelos diferentes con las diferentes constantes. Uno con únicamente variables no financieras, uno con variables meramente financieras y un condensado de ambas.

Comentado [U17]: No se entiende la importancia de la K.

En la figura #8 se revela el modelo de capacidad de predicción mayor generado en el artículo científico.

Observaciones	Predicciones		Porcentaje correcto
	0	1	
0	409	49	89,3%
1	56	410	88,0%
Total			88,6%

Figura #8: Matriz de clasificación probabilística

Fuente: *Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian Banks*, por Abdelmoula, A. K., (2015), *Accounting and Management Information Systems*, p.91. Copyright 2015 Abdelmoula, A. K.

El modelo tiene un poder de predicción de 88.6%. Clasificando correctamente un 89.3% de los pagos a tiempo y un 88% de los que hicieron default. Este modelo utilizo K=3 como constante y fue generado por las variables financieras únicamente. Entre las que se destacó con mayor significancia para el modelo, la ratio de gasto/ingreso. Mientras que la variable que menos influyo fue el monto de colateral presentado.

Comentado [U18]: ¿?

**ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

Análisis de modelo de puntuación de crédito utilizando el artículo científico #8:
The Comparison of Credit Risk between Artificial Neural Network and Logistic
Regression Models in Tose-Taavon Bank in Guilan, (2015).

El siguiente modelo se construyó con 376 casos provistos por un banco comercial en Guilan en el periodo 2009-2014. El subconjunto de datos se dividió en 302 buenos clientes y 74 malos clientes.

El método empleado en este modelo fue la regresión logística utilizando como variables exploratorias el monto del préstamo, la tasa de interés, número en meses de pagos, garante, monto mensual de ingresos del cliente.

Comentado [U19]: Equivalente a análisis discriminatorio

Comentado [U20]: número

El estudio reveló que las variables más significativas e influyentes en el modelo son la tasa de interés y el número de pagos mensual en ese orden. Indicando un poder de predicción del modelo de 82.7%.

Análisis de modelo de puntuación de crédito utilizando el artículo científico #9

Este modelo se basó en un subconjunto de datos provistos por un repositorio de información. Esto se debe a que se quiere proteger la información del banco y la información privilegiada del consumidor. El número de casos fue de 1000 en los que 700 fueron sujetos de crédito y 300 no. Además, se utilizaron 20 variables numéricas y categóricas.

Comentado [U21]: fue

Comentado [U22]: fueron

El método utilizado fue regresión logística. Empleada en la versión 18 del software SPSS. Sin embargo, el modelo cuenta con una posible sobrealimentación debido a la amplia cantidad de variables utilizadas. La figura #9 muestra la clasificación resultante del modelo a su vez como la probabilidad de predicción del modelo.

ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

Observaciones	Predicciones		
	Buen cliente	Mal cliente	Probabilidad correcta
Buen cliente	624	76	89,1%
Mal cliente	160	140	46,7%
Total			76,4%

Figura #9: Resultados de clasificación de clientes

Fuente: Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring, por Islam, M., Zhou, L. y Li, F., (2009), Independent thesis Advanced level (degree of Master (One Year)), p.15. Copyright 2009 Spring.

La tabla #9 señala que el modelo genera una probabilidad de un falso positivo de 10.9%. Clasificando correctamente el 89.1% de los buenos clientes. Sin embargo, el modelo genera una probabilidad de un falso negativo del 53.3%, lo cual es muy alto. Un resultado poco aplicable en el sector bancario debido al alto costo que este modelo representa por la mala clasificación de malos créditos como bueno. Además, el modelo **presentó** el estado de la cuenta bancaria como la variable más significativa e influyente del modelo, mientras que la variable menos significativa no fue mencionada.

Comentado [U23]: presentó

Análisis de modelo de puntuación de crédito utilizando el artículo científico #10

Este modelo se basó en 2000 casos de entrenamiento de un banco no descrito producto de la sensibilidad de la información provista. El periodo de tiempo tampoco fue descrito. Los 2000 casos fueron subdivididos en 10 agrupaciones de 200 cada una. 12 variables fueron descritas en el modelo. La agrupación con la menor cantidad de errores de suma cuadrada fue la elegida para realizar el árbol de clasificación como método elegido. En este caso la agrupación con la constante $K=3$.

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

El árbol se desarrolló utilizando el algoritmo C5.0 en SPSS Clementine. Dividiendo 60% de la información como entrenamiento y un 40% para prueba de poder predictivo.

La figura #10 revela la clasificación resultante del modelo propuesto con su respectiva probabilidad de predicción.

Observaciones	Predicciones		
	Buen Cliente	Mal cliente	Probabilidad correcta
Buen cliente	116	4	96,7%
Mal cliente	9	71	88,8%
Total			93,5%

Figura #10: Clasificación de clientes realizada por el modelo

Fuente: *Data mining application in banking sector with clustering and classification methods*, por Çaliş, A., Boyacı, A. y Baynal, K, (2015), *Industrial Engineering and Operations Management (IEOM)*, p.6. Copyright 2015 IEEE.

El modelo muestra un 93.5% de poder predictivo general. Clasificando correctamente el 96.7% de los buenos clientes, mientras que el 88.9% de los malos clientes. Así mismo, el modelo implica que la variable más influyente y de mayor significancia es el estatus de educación, siendo la universidad un valor relevante y de mayor importancia para la decisión en diferencia de educación secundaria o primaria. La variable menos influyente fue omitida.

Análisis de tabla comparativa de modelos

En la figura #11 se muestra un condensado de las variables más influyentes en los modelos previamente mencionados, las variables menos influyentes, la técnica utilizada y el poder de predicción del modelo. La tabla esta ordenada en orden descendente de acuerdo a su confiabilidad probabilística o poder de predicción.

ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE CREDITO GENERADOS POR LA MINERIA DE DATOS EN INSTITUCIONES FINANCIERAS

	Técnica Utilizada	Variable mas influyente	Variable menos influyente	Confiabilidad probabilist
Paper #6	Arboles de clasificacion	Garantia de liquidez del cliente	Lugar de nacimiento del cliente	94,1%
Paper #10	Arboles de clasificacion	Nivel de educacion	No mencionada	93,5%
Paper #1	Redes neuronales artificiales	Ratio deuda/ingreso	Genero	91,3%
Paper #3	Regresion logistica	Historial de credito	Edad	89,0%
Paper #2	Redes neuronales artificiales	Frecuencia de prestamo individual	Estatus de cuenta del cliente	88,9%
Paper #5	Regresion logistica	Tasa de interes	Tiempo de retraso de prestamo	88,9%
Paper #7	K-vecinos cercanos	Ratio deuda/ingreso	Colateral	88,6%
Paper #4	Redes neuronales artificiales	Historial de credito	Edad	86,7%
Paper #8	Regresion logistica	Tasa de interes	Garante	82,7%
Paper #9	Regresion logistica	Estado de la cuenta bancaria	No mencionada	76,4%

Figura #11: Cuadro comparativo de técnicas, variables y confiabilidad

Fuente: Autor

Los resultados revelan que los árboles de clasificación **obtienen** la mayor capacidad predictiva de entre todos los métodos analizados superando el 93% de confiabilidad. Sin embargo, las redes neuronales artificiales, la regresión logística y los vecinos cercanos son técnicas válidas y con alta probabilidad de clasificación acertada para los modelos de puntuación de crédito. Este estudio contrasta con el de Lemos (Lemos, 2005) en el que indica a las redes neuronales artificiales con una mejor probabilidad predictiva que los arboles de clasificación.

Entre las variables destacadas como de mayor influencia y significancia recurren la capacidad de pago de los clientes de acuerdo a la ratio de deuda/ingreso, así como el historial de crédito y la tasa de interés. Mientras que variables demográficas como la edad y genero muestran ser poca significancia para el desarrollo de modelos de estas características.

Conclusiones

A pesar de que la aprobación de préstamos en bancos ecuatorianos ha sido sujeta en su mayoría al criterio del oficial de crédito o al soporte de un modelo de puntuación de crédito tradicional, los bancos pueden mejorar sus procesos de

Comentado [U24]: Tenemos algún dato de costos? De pronto hacer redes neuronales es menos costoso que árboles de clasificación y tan solo se pierden 2 puntos porcentuales.... Costo/Beneficio.

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

aprobación de crédito y disminuir la tasa de morosidad de sus clientes empleando modelos de puntuación de créditos generados por la minería de datos. Este estudio propone un nuevo enfoque para evaluar las solicitudes de préstamos como un modelo de apoyo para la decisión del oficial de crédito. El artículo compara diferentes técnicas utilizadas en modelos de puntuación de crédito como: árboles de clasificación, redes neuronales artificiales, regresión logística y k-vecinos cercanos, las variables más influyentes en estos modelos, las variables menos influyentes y la capacidad de predicción de las técnicas y modelos empleadas. Los resultados muestran a los árboles de clasificación como las técnicas con mayor capacidad de predicción entre créditos buenos y créditos riesgosos, con un 94,1% y 93,5% de eficacia. Seguido de las redes neuronales artificiales con un poder predictivo de 91,3% y regresión logística con un 89%. Estableciendo a todas las técnicas por encima del 75% aceptable. Además, se revela que la relación gastos/ingresos, el historial crediticio y las tasas de interés son un factor de alta significancia para la determinar si debería otorgarse el crédito o no. Mientras que variables demográficas como edad y genero son de poca relevancia para los modelos.

Por otra parte, esta publicación propone un estudio más profundo que compare el costo de clasificación de créditos errónea generado por diferentes técnicas estadísticas relacionadas a la minería de datos.

Bibliografía

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. *Accounting and Management Information Systems*, vol. 14, no 1, p. 79.

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

- Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2016). The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank. *Journal of the Knowledge Economy*, 1-15.
- Arunkumar, R. &. (2006). Risk Management in Commercial Banks (A case study of Public and Private sector banks). In *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, (págs. 1-22). Delhi.
- Bekhet, H. A. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, vol. 4, no 1, p. 20-28.
- Bekhet, H. A., & Eletter, S. F. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, vol. 4, no 1, p. 20-28.
- Bradley, P. S. (1999). Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, vol. 11, no 3, p. 217-238.
- Çaliş, A., Boyaci, A., & Baynal, K. (2015). Data mining application in banking sector with clustering and classification methods. *Industrial Engineering and Operations Management (IEOM)*, (págs. 1-18). Dubai, United Arab Emirates (UAE).
- Cao, Q., & Parry, M. E. (2009). Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm. *Decision Support Systems*, vol. 47, no 1, p. 32-41.
- Chitra, D. K., & Subashini, B. (2013). Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no 8, p. 219-226.
- Chitra, K., & Subashini, B. (2011). Customer Retention in Banking Sector using Predictive Data Mining Technique. In *ICIT 2011 The 5th International Conference on Information Technology*, (págs. 1-4).
- Dong, G., K., L. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, vol. 1, no 1, p. 2463-2468.
- Fondo Monetario Internacional . (2016). *Informe sobre la estabilidad financiera mundial* .

**ANÁLISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

- Huang, C. L. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, vol. 33, no 4, p. 847-856.
- Islam, M., Zhou, L., & Li, F. (2009). *Application of artificial intelligence (artificial neural network) to assess credit risk: a predictive model for credit card scoring*. Blekinge, Sweden: Blekinge Institute of Technology, School of Management.
- Karimi, A. (2014). Credit Risk Modeling for Commercial Banks. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, vol. 4, no 3, p. 187-192.
- Koh, H. C., & Gerry, C. K. (2002). Data mining and customer relationship marketing in the banking industry. *Singapore Management Review*, vol. 24, no 2, p. 1.
- Kurgan, L. A. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, vol. 21, no 01, p. 1-24.
- Lemos, E. P. (2005). Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. *Revista de Administração*, vol. 40, no 3, p. 225-234.
- M., T. T., Aghajan Nashtaei, R., & Chirani, E. (2015). The Comparison of Credit Risk between Artificial Neural Network and Logistic Regression Models in Tose-Taavon Bank in Guilan. *International Journal of Applied Operational Research-An Open Access Journal*, vol. 5, no 1, p. 63-72.
- Moin, K., & Ahmed, D. Q. (2012). Use of data mining in banking. *International Journal of Engineering Research and Applications*, vol. 2, no 2, p. 738-742.
- Nazari, M., & Alidadi, M. (2013). Measuring credit risk of bank customers using artificial neural network. *Journal of Management Research*, vol. 5, no 2, p. 17.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.
- Sousa, M. D., & Figueiredo, R. S. (2014). Credit analysis using data mining: application in the case of a credit union. *JISTEM-Journal of Information Systems and Technology Management*, vol. 11, no 2, p. 379-396.

**ANALISIS COMPARATIVO DE MODELOS DE PUNTUACION DE
CREDITO GENERADOS POR LA MINERIA DE DATOS EN
INSTITUCIONES FINANCIERAS**

- Superintendencia de Bancos. (2016). *Analisis Financiero: sistema de bancos privados*. Quito .
- Yaghini, M., Zhiyan, T., & Fallahi, M. (2011). A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks. *Int'l Conf. Data Mining*, (págs. 1-6).
- Zhang, D., X., Z., C., L. S., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, vol. 37, no 12, p. 7838-7843.