Modelo de Árboles de Clasificación para la Identificación del Perfil del Alumno según el Riesgo Crediticio en la Universidad Peruana Unión.

De los Santos Almazán, Jean Carlo[1]; Quispe Condori Daniel Ángel[2]

Dr. Palza Vargas, Edgardo; Dr. Mamani Apaza, Guillermo; Mg. Acuña, Erika

Resumen

El presente trabajo tiene como objetivo la construcción de un modelo de árboles de clasificación para la identificación del perfil de los niveles del alumno con alto y bajo riesgo crediticio en la Universidad Peruana Unión. Para ello se utilizó la herramienta de Business Intelligence SQL Analysis Services 2008 del SQL Server y el SPSS 15.0, tanto para la construcción del modelo, como para la validación de éste respectivamente. El centro de aplicación fue el área financiera de la Universidad Peruana Unión.

La metodología utilizada fue el CRISP-DM que es una metodología para proyectos de minería de datos.

El modelo ha identificado los perfiles de los alumnos según el riesgo crediticio que éste tenga con las siguientes fases: Comprensión del negocio, análisis de los datos, preparación de los datos, modelamiento, evaluación e implantación.

Palabras clave: Árboles de clasificación, riesgo crediticio.

Classification Trees Model for Identifying the Student's Profile according to Credit Risk of Universidad Peruana Unión

Abstract

This work aims at building a classification tree model to identify the profile of the levels of students with high and low credit risk in Universidad Peruana Union. We used BI tool SQL Analysis Services 2008 SQL Server and SPSS 15.0 for both model building, and for the validation of this respectively. The main area for which this work is directed to the area of finance at the Universidad Peruana Union.

The methodology for this work was the CRISP-DM, one of the methodologies used in data mining project. Once built the model is able to identify the profiles of students based on credit risk it has. With this information, the area of finances can make plans to reduce credit risk.

Keywords: Classification trees, credit risk.

Introducción

Manfredo Añez, de la Universidad de Buenos Aires en el año 2000 definió al análisis crediticio como un arte, ya que no hay esquemas rígidos y que por el contrario es dinámico y exige creatividad por parte del responsable de créditos o negocio. Sin embargo es importante dominar las diferentes técnicas de análisis de créditos, así mismo es necesario contar con la información necesaria y suficiente de nuestros clientes que nos permita minimizar el número de incógnitas para poder tomar la decisión correcta.

El presente trabajo tiene como objetivo general el desarrollo de un modelo de árboles de clasificación, que sirva para la identificación del perfil de alumnos según el riesgo crediticio que

[1]*E.A.P. de Ingeniería de Sistemas, Universidad Peruana Unión. jeancarlodls@live.com*

[2]*E.A.P. de Ingeniería de Sistemas, Universidad Peruana Unión. angel_js147_33@hotmail.com*

éste pueda tener.

El proceso de minería de datos toma un gran valor para este tipo de análisis, el mismo que facilita la elaboración de un modelo de árbol de clasificación. Para esto existen cinco etapas : Comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación. En la etapa de comprensión del negocio se comprende los objetivos y requisitos del proyecto desde una perspectiva empresarial, en la siguiente fase de comprensión de los datos se establece un primer contacto con el problema, en la fase de preparación de los datos se seleccionan los datos que van a intervenir en la creación del modelo de árboles de clasificación, en la fase de modelado se lleva a cabo la creación del modelo de árbol de clasificación y en la última fase de evaluación se valida el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema.

La elaboración de un modelo para el análisis del riesgo crediticio es importante para la toma de decisiones y debería estar regido en uno de los estándares establecidos en el acuerdo de BASILEA II, en la que se presenta la clasificación de riesgo de crédito en cinco categorías: Categoría A o "riesgo normal", Categoría B o "riesgo aceptable, superior al normal", Categoría C o "riesgo apreciable", Categoría D o "riesgo significativo", Categoría E o "riesgo de incobrabilidad". El modelo de árboles de clasificación propuesto sirve para tener una comprensión de las características de los diferentes perfiles del alumno de la Universidad Peruana Unión con respecto a la categoría del riesgo crediticio que éste pueda tener ya que un punto importante al momento de analizar el riesgo crediticio es la información del cliente que se pueda tener. El modelo de árboles de clasificación permite explotar los datos de los alumnos de la UPeU, representándolo en un gráfico de fácil entendimiento que muestra las diferentes características según el riesgo crediticio del alumno. La gerencia de finanzas podrá tener una administración más eficiente sobre el riesgo crediticio, tomando acciones preventivas y correctivas para disminuir éste tipo de riesgo.

Árboles de clasificación en relación al riesgo financiero. Salinas Flores, Jesús de la Universidad Mayor de San Marcos en el año 2005 en su investigación denominada "Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART" planteó como objetivo encontrar un patrón de comportamiento de la morosidad a partir de la información obtenida al momento de solicitar un crédito para un producto crediticio y a su vez dar a conocer una nueva técnica estadística muy útil para este campo que es el árbol de clasificación CART.

Con el desarrollo de su investigación Salinas detectó patrones diferentes para los morosos y no morosos. Además llegó a la conclusión que usando el algoritmo CART se consiguen detectar las variables más influyentes sobre la morosidad; en su investigación resultaron ser: la ubicación geográfica, la antigüedad laboral, la edad, la carga familiar y el estado civil. De su análisis obtuvo que un 93.75% de clasificación correcta para los morosos y un 86,42% para los no morosos.

Otra investigación fue la realizada por Luis Ernesto Domínguez Velásquez de La Paz, Bolivia el 2006. Esta fue denominada "Minimización del riesgo crediticio mediante la evaluación de la solicitud del cliente". En este documento el autor mencionó en sus conclusiones que "el resultado obtenido es excelente ya que se obtuvo una clasificación en el conjunto de aprendizaje de 92.45%. Además, el árbol construido tiene la ventaja sobre otros sistemas de clasificación como las redes neuronales artificiales, de que las reglas producto de los sistemas de inducción son entendibles para un analista humano, y además las variables irrelevantes son eliminadas del modelo, pues no figuran en los árboles…".

Otro artículo de la Revista Colombiana de Estadística titulado "Aplicación de árboles de decisión en modelos de riesgo crediticio" desarrollado por Paola Andrea Cardona Hernández y publicado en Diciembre del 2004

Materiales y Métodos

Para la elaboración del modelo de árboles de clasificación se utilizó la metodología CRISP-DM, propuso en el año 1999 cuando un importante consorcio de empresas NCR, AG, SPSS , OHRA, Tera data, y Daimer - Chrysler, plantean el desarrollo de una guía de referencia de libre distribución, aplicando las seis fases de esta metodología.

Fase de Comprensión del negocio. En esta fase nos reunimos con la parte administrativa de finanzas, para poder entender el problema que presenta dicha área. El problema que se encontró en el área de finanzas fue el riesgo crediticio que tienen los alumnos de la UPeU. Una vez determinado el problema, se procedió a reunirnos con la persona encargada de dar plazos de créditos a los alumnos, esta reunión se llevó con el fin de entender la forma de analizar el crédito y las variables que se toma en cuenta al otorgar los créditos correspondientes (Figura 1).

Figura 1 - Variables que intervienen en el análisis de otorgamiento de crédito en el área de finanzas

Fase de compresión de los datos. Para la obtención de estos datos por cada alumno, se procedió a contactarnos con la parte informática del área de finanzas para solicitar dichos datos.

Tabla 1 - Varibles más representativas del modelo

Tipo Riesgo	Nivel de riesgo crediticio del alumno	Alto riesgo Bajo riesgo
Financiamiento	Tipo de financiamiento de los estudios del alumno	Ayuda de sus padres Beca ayuda institucional Auto sostén Beca Ley
ModalidadPago	La cantidad de cuotas que paga el alumno	5 cuotas 2 cuotas Al contado
MontoCrédito	El monto de crédito que pide el alumno al momento de la matrícula.	0-300 301-600 601-1200 1201-1600 1600 a más
Sectoreconomico	Sector económico al que pertenece el cliente	Actividades de servicio Actividades de producción Otros
Estudiaenotra	Si estudia o no en otra institución	Si No
SitlaboralPadre	Situación laboral del padre del alumno	Dependiente Independiente Jubilado Otro
IngresoPadres	La cantidad de ingreso que percibe el padre del alumno	
EdadPadres	Edad del padre del alumno	

Tabla 2 - Probabilidad de predicción del modelo

	Serie, Modelo	Puntuación	Población correcta	Probabilidad de predicción
	DATA	1.00	85.00%	97.92%
	Modelo Ideal		85.00%	

Fuente: BI del SQL Server 2008

En la tabla 1 se muestra la lista de datos. Se nos otorgó los datos de 800 alumnos del total de alumnos de la Universidad Peruana Unión.

Fase de preparación de los datos. Para esta fase se procedió a utilizar todos los datos, ya que estos son provenientes de fuentes con datos exactos y no provienen de encuestas hechas a los alumnos, en las cuales puede haber datos incorrectos.

Luego se procedió a pasar los datos a la herramienta de Microsoft Excel 2007, para más adelante exportarlas a las otras herramientas que se utilizaron en las fases posteriores.

Fase de modelado. En esta fase se procedió a construir el modelo de árboles de clasificación con ayuda de la herramienta SQL Analysis Services. En la Figura 2, se muestra el proceso de construcción del modelo usando dicha herramienta. Esta herramienta tiene disponible ocho algoritmos para la solución de problemas de minería de datos los cuales son aplicados según la necesidad. Por la naturaleza de la investigación el más apropiado es el algoritmo de árbol de decisión de Microsoft el que se basa para la construcción de los nodos del árbol en la correlación existente. Luego el orden de prioridad de cada una de las clases dado a través de los métodos: Puntuación interestingness, Entropía de Shannon, Bayesiano con prioridad K2, Dirichlet bayesiano con prioridad uniforme (predeterminado).

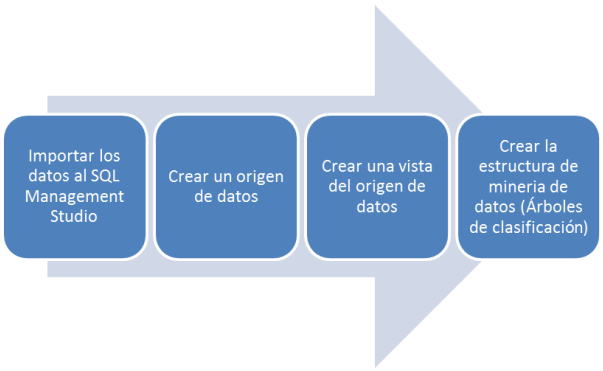


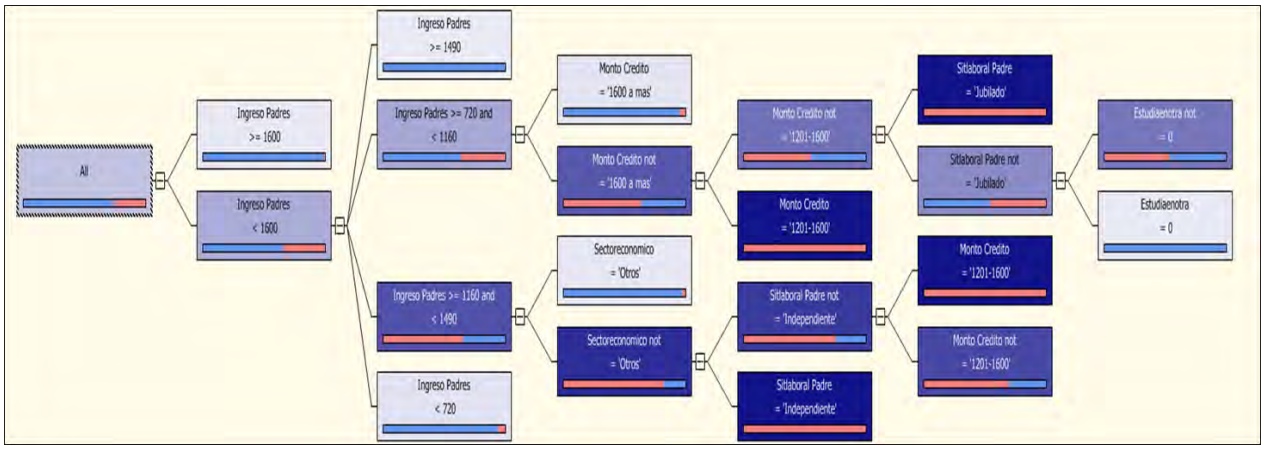
Figura 2. Proceso de construcción del modelo usando la herramienta SQL Server 2008 con Analysis Services.

Fase de evaluación. En esta fase se procedió a utilizar la herramienta SPSS 15.0 para observar las variables que tienen mayor relación con la variable riesgo crediticio. Este análisis se comparó con el árbol de clasificación construido, para determinar si el perfil del alumno según el nivel de riesgo crediticio tiene que ver con las variables de mayor relación analizadas en el SPSS 15.0 .

Otro punto para la validación de nuestro modelo de árboles de clasificación es el gráfico de precisión representado en la Figura 3, fue extraída del diseñador de minería de datos de la herramienta de BI del SQL Server y sirve para validar la precisión y comparar la habilidad de predicción de los modelos de minería de datos.

Este gráfico de elevación se creó trazando los resultados de las consultas de predicción de un conjunto de datos de prueba según los valores conocidos de la columna de predicción, en este

Modelo final de árbol de clasificación con los 7 niveles especificando la línea de dependencia para el riesgo crediticio



Fuente: BI del SQL Server 2008

caso el grado de riesgo crediticio, que se encuentra dentro del conjunto de datos.

En dicha figura podemos observar estas afirmaciones. El gráfico muestra dos líneas, como se representa en la Tabla 2, la línea roja expresa los datos reales y la línea de color azul expresa los resultados que producirían un modelo ideal, son las predicciones perfectas que nunca están equivocadas.

Figura 3. Gráfico de precisión del modelo de árboles de clasificación (BI del SQL Server 2008)

Ahora, con una población correcta al 85% podemos obtener una probabilidad de predicción del 97.92%, esto quiere decir que nuestro modelo de árboles de clasificación se acerca mucho al

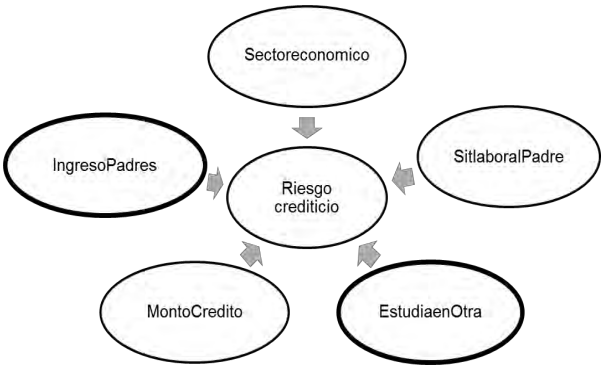


Figura 4 - Red de dependencias con el nodo principal dependiente Riesgo Cre-

Modelo Ideal propuesto; y así podemos comparar los distintos porcentajes de la población. Y con esto, podemos afirmar que al 95% de la población correcta y al compararlo con el modelo ideal, nuestro árbol de clasificación mostrado en esta investigación obtiene una probabilidad de predicción de 97.84%.

Fase de implementación: en esta fase se procedió a realizar los planes de implementación propiamente dicha, el plan de monitoreo y mantención, se elaboró el informe final y se realizó una revisión del proyecto.

Resultados y discusión.

La tabla 3, se muestra las correlaciones de las 7 variables que influyen más al riesgo crediticio en

esta investigación extraída luego del procesamiento con la herramienta SPSS. Las variables encerradas en las celdas con borde de color rojo son las resultantes del análisis hecho por el BI del SQL Server. Como podemos ver 5 de estas 7 variables se muestran en nuestra red de dependencias (MontoCredito, IngresoPadres, EstudiaenOtra, SectorEconomico y SitLaboralpadre) que se muestra en la Figura 4.

Tabla 3 - Resultados del análisis de correlación

Pos	Variable	Coeficiente de co-rrelacion	Valor
1	Montocredito	R de pearson	0.291
		Sig	0.000
2	IngresoPadres	R de pearson	0.289
		Sig	0.000
3	EstudiaenOtra	R de pearson	0.278
		Sig	0.000
4	Edadpadre	R de pearson	0.248
		Sig	0.000
5	SectorEconomico	R de pearson	0.229
		Sig	0.000
6	Sitlaboral	R de pearson	0.182
		Sig	0.000
7	SitLaboralpadre	R de pearson	0.179
		Sig	0.000

Fuente: SPSS

En este análisis podemos observar que las 2 variables con más relación a nuestro modelo son IngresoPadres y EstudiaenOtra con un R de Pearson de 0.289 y 0.278 correspondientemente; confirmando la sugerencia hecha por la herramienta del SQL Server.

Los altos coeficientes de correlación y el Sig. tan bajo (0.000) muestran el alto grado de correspondencia de estas variables independientes con el nivel de riesgo crediticio. Además podemos afirmar con estos valores que existe una correlación positiva moderada al nivel de 0,01 y es significativa.

Además se puede afirmar que existe mucha relación con los valores del análisis de correlación con esta herramienta y la expuesta por la herramienta de Business Intelligence del SQL Server 2008.

Análisis del modelo. Del proceso para la construcción del árbol de clasificación generado con los 7 niveles correspondientes podemos afirmar que:

Se realizó el proceso de aprendizaje de las variables de entrada con respecto a la variable de predicción con un porcentaje de casos al 90%.

La clasificación del riesgo crediticio viene a ser dada por 2 categorías: Riesgo crediticio alto y riesgo crediticio bajo.

El modelo de árboles de clasificación generado obtuvo una puntuación de 0.97 y representando una predicción con una probabilidad de confiabilidad del 94.68% expresado en la herramienta de BI del SQL Server 2008 sobre la minería de datos.

Sobre el árbol de clasificación:

Se analizaron 720 casos de la muestra total obtenida.

Sobre estos casos:

El 72.64% de los alumnos representan a los alumnos con un ingreso de los padres menor a 1600 Nuevos Soles, el porcentaje restante supera esta cantidad.

Sobre estos alumnos, los padres con ingresos entre 1160 Nuevos Soles y menores a 1490 Nuevos Soles, el 65.24% poseen mayor grado de riesgo crediticio.

Sin embargo, también encontramos un alto porcentaje de los padres con ingreso menor a 1160 Soles alcanzando un 32.9% de los casos sobre los alumnos antes mencionados.

En esta rama de alumnos los que poseen un monto de crédito mayor a 1600 son los que poseen mayor porcentaje de no cubrir el crédito resultando en casi un 60% de los casos. Y de estos los que oscilan entre 1201 y 1600 representan un punto crítico ya que representan el 100% de los casos incumplidos por parte de los alumnos.

De los casos restantes de la rama antes

mencionada los padres en situación de jubilación también alcanzan el 100% de los casos y los que no logran un 70% de poseer un alto grado de riesgo crediticio.

Además, los alumnos con padres no jubilados todavía representan un 52% de los casos de incumplimiento de créditos.

Conclusiones

A través de los Árboles de Clasificación se ha definido que las características de los alumnos con alto grado de Riesgo Crediticio son los que cuentan con padres cuyos ingresos son menores a 1160 Nuevos Soles y presentan una situación laboral independiente, además ostentan un crédito mayor a 1600 Nuevos Soles, no cuentan con sostenimiento económico propio y estudian en otras instituciones. Por el contrario los alumnos con menor riesgo crediticio dependen directamente del ingreso de los padres ya que los que tiene padres con ingresos mayores a 2000 soles poseen una porcentaje bajo de riesgo de casi un 85%.

Los árboles de clasificación es una técnica de minería de datos y podemos afirmar que su interpretación y análisis es práctica ya que muestra de manera ordenada y de fácil entendimiento las características de los alumnos con alto y bajo riesgo crediticio de la Universidad Peruana Unión.

El tener los datos representados en un árbol de clasificación permite que el análisis por parte de la gerencia sea correcto y lleve menos tiempo la identificación de los diferentes perfiles según el riesgo crediticio de los alumnos de la Universidad Peruana Unión.

El modelo de árboles de clasificación que se construyó con la herramienta SQL Analysis Services se validó con la herramienta SPSS 15.0, la cual tuvo resultados semejantes en lo que se refiere a la relación de las variables independientes con la variable dependiente.

Se logró construir el modelo de árboles de clasificación que permitió realizar la identificación de los perfiles según el riesgo crediticio de los alumnos de la Universidad Peruana Unión.

Con el modelo de árboles de decisión se ha conseguido detectar las variables que más influyen sobre el riesgo crediticio alto/bajo del alumno de la Universidad Peruana Unión. Estas variables son: Ingreso de los padres, Monto de crédito, Sector económico, situación laboral del padre y si estudia en otra institución.

Recomendaciones

Realizar modelos de árboles de clasificación con distintas herramientas de minería de datos para hacer una validación y comparación de los distintos modelos de árbol que se construye, de esta manera se realice un mejor análisis y tomar el mejor modelo para el área de finanzas.

Para una mejor construcción de un árbol de clasificación, se recomienda la omisión de algunas variables, ya que existen variables que no tienen relación con el riesgo crediticio, de esta manera se optimizaría la construcción del árbol.

Se recomienda realizar un árbol de clasificación general, que se pueda utilizar como punto de partida para la construcción de futuros árboles de clasificación y así poder utilizarlo en diferentes tipos de empresas.

Se recomienda hacer un trabajo conjuntamente con el área de finanzas, y que el área de finanzas forme parte del proyecto desde la parte inicial hasta la parte final. Debe existir una comunicación adecuada entre amabas partes, tanto como el equipo de proyecto y el área de finanzas.

Se recomienda en una futura investigación crear una interfaz exclusivamente para los gerentes de finanzas, para que éstos puedan utilizarlo.

Para la construcción del modelo se recomienda usar una mayor cantidad de datos, si es posible con los datos de todos los alumnos de la Universidad Peruana Unión. Mientras más sea la cantidad de alumnos, el modelo de árboles de clasificación más se ajustará a la realidad de la

universidad.

Agradecimientos :

Se agradece al equipo que ha replicado el trabajo para certificar los resultados.

El equipo que lo conforma es :

-Willy Medina Bacaya.

-Rosa Liz Valle Yanavilca.

-Kelly Sobrado León.

-Eber Ortiz Mas.

Referencias

Azevedo A. 2008. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. [Artículo en línea]. IADIS European Conference Data Mining 2008. 4 pp. ISBN: 978-972-89-24-63-8. [Consultado en 1 de octubre de 2010]. Formato pdf. Disponible en: <http:// www.iadis.net/dl/ final_uploads/200812P033.pdf >.

Berzal F. 2002. ART un método alternativo para la construcción de árboles de decisión. [Tesis doctoral]. Asesor: Juan Carlo Cubero Talavera. Granada: Departamento de Ciencias de la computación e inteligencia artificial, Universidad de Granada. 339 pp. [Consultado en 10 de octubre de 2010].

Bonàs A., Llanes M., Usón I. Y Veiga N. 21 de Junio 2007. Universitat Pompeu Fabra – IDEC. Máster en Mercados Financieros

Campoverde F. 2007. Asesor Empresarial y Catedrático Universitario- Universidad Espíritu Santo- Guayaquil.

Cardona P. 2004. Aplicación de árboles de decisión en modelos de riesgo crediticio. Revista Colombiana de Estadística. 27 (2). 139 – 151 pp.

CUELLO GABRIEL 2006 "TÉCNICAS DE MINERÍA DE DATOS DENTRO DE CONTEXTOS METODOLÓGICOS Y DE EMPRESA" **< http://es.scribd.com/ doc/26802110/Instituto-Tecnologico-de-Buenos-Aires-Escuela-De >**

Díaz Z. 2007. Predicción de crisis empresariales en seguros no vida, mediante árboles de decisión y reglas de clasificación. Madrid: Editorial Complutense. 144 pp. ISBN: 978-84-7491-882-3.

Gallardo J. 2009. Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM). [Tesis doctoral]. Asesor: Oscar Marbán Gallego. Departamento de Lenguajes y sistemas informaticos e ingeniería de software, Facultad de informática. 317 pp. [Consultado en 5 de octubre de 2010] Formato pdf. Disponible en: <http://oa.upm.es/1946/1/ JOSE_ALBERTO_GALLARDO_ARANCIBIA .pdf>.

Gomes R., González J. Morosidad, Gestión de Cobros e Impagos, Concurso de Acreedores,

Análisis de Riesgo.

Inza P. 2005. Árboles de Clasificación para el riesgo crediticio. Departamento de Ciencias de la Computación e Inteligencia Artificial – Universidad del País Vasco.

Ledesma Z. 2007. Análisis del riesgo crediticio bancario en la economía cubana. Teoría y Praxis. 3: 77 – 87 pp.

Llombart Óscar Alonso 2005 BI : Inteligencia aplicada al negocio[Articulo en linea] **< http://es.scribd.com/doc/6811330/BI-Inteligencia-Aplicada-Al-Negocio >**

López C. 2006. Análisis avanzados de grandes volúmenes de datos en el sector seguros. [Artículo científico en línea]. Actuarios: 10 pp. [Consultado en 5 de octubre del 2010].

Pérez J. 2006. Árboles Consolidados: Construcción de un árbol de clasificación basado en múltiples submuestras sin renunciar a la explicación. [Tesis para doctor en informática]. Asesor: Javier Muguerza

Petterson S. 2008. Diseño, selección y síntesis de nuevos inhibidores de entrada del VIH. [Tesis de ingeniero químico]. Asesor: Dr. Jordi Teixidó. Barcelona: Departamento de química orgánica, Escuela Técnica Superior IQS. 204 pp. [Consultado en 5 de octubre de 2010]. Formato pdf. También disponible en: < http://www.tesisenxarxa.net/TESIS_URL/AVAILABLE/TDX-1215109-152530//tsps1.pdf >.

Rivero. Donostia: Facultad de Informática de la Universidad del País Vasco. 293 pp. [Consultado en 10 de octubre de 2010]. Formato pdf. También disponible en: < www.sc.ehu.es/acwaldap/ald.eng/PhDThesis/PerezPhd_spanish.pdf >.

Sacco L. 2009. Metodología SEMMA. Lea en Binario. [Consultado en 10 de octubre de 2010]. Formato html. También disponible en: < http://leaenbinario.blogspot.com/2009/11/semma.html >.

Salinas J. 2005. Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART. Revista Industrial Data. 8 (1): 29 – 36 pp. ISSN:1560 -9146.