Modelo de Árboles de desción para pronosticar la morosidad de los alumno de la Universidad Peruana Unión.

Vargas, Hovanna; Ccapa, Lesly
Dr. Palza Vargas, Edgardo; Dr. Mamani Apaza, Guillermo.

Resumen

La presente investigación tiene por objetivo determinar un Modelo de Árboles de decisión que permite el pronóstico de las características de morosidad de los alumnos de la Universidad Peruana Unión. La metodología utilizada es CRISP-DM, creada por especialistas para proyectos de minería de datos. Al aplicar el modelo de árboles de decisión se logró identificar las características de un alumno moroso; distribuidas en cinco variables predominantes: Ayuda Institucional, Ingreso de los padres, Monto de Crédito, Tarjetas de crédito y la Situación laboral del padre.

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

Present investigation aims to determine a decision tree model that allows prediction of the characteristics of late payment by students of the Universidad Peruana Union. The methodology used is CRISP-DM, created by specialists for mining projects. In applying the decision tree model was able to identify the characteristics of a delinquent student, divided into five predominant variables: Institutional Support, Income of parents, amount of credit, credit cards and the Father's work status.

Keywords: Classification trees, credit risk.

I. INTRODUCCION

Actualmente en el mundo cada año aumentan las entidades crediticias, el mercado es cada vez más competitivo; por lo tanto una entidad crediticia debe ejercer control efectivo sobre el proceso de evaluación de sus clientes con el fin de otorgarle o negarle el crédito solicitado.

La cartera de crédito al consumo implica el manejo de un gran número de clientes. Las entidades financieras requieren procesar un gran número de solicitudes de crédito, por tanto es importante que la administración deba conocer el comportamiento de sus clientes.

El riesgo de crédito es el tipo de riesgo más importante al que debe hacer frente cualquier entidad financiera. Un indicador del riesgo crediticio es el nivel de morosidad de la entidad, es decir, la proporción de su cartera que se encuentra en calidad de incumplimiento.

Una institución que tenga altas tasas de morosidad y de préstamos incobrables no es viable en la perspectiva futura. La morosidad y el incumplimiento de los clientes de la devolución de los créditos otorgados, ocasionan a las empresas costos y los convierte en empresas ineficientes, afectando su situación financiera y económica.

Establecer cual son las causas o determinantes del índice de morosidad, no es problema sencillo de resolver, dado que no existen muchos estudios con evidencia empírica. De acuerdo con el punto de vista tradicional, el cliente incumple sus pagos porque el uso indebido del préstamo lo coloca en

incapacidad de pagar. Sin embargo el incumplimiento generalizado es frecuentemente un reflejo de la renuencia a pagar por parte del prestatario.

En el Perú la Universidad Peruana Unión tiene como clientes a sus alumnos, y los servicios que brinda son esencialmente educativos de formación profesional y especialización a nivel de pregrado y posgrado brindados a través de sus Facultades Pre- grado y Posgrado.

La Universidad se encuentra con un porcentaje alto de alumnos morosos en las diferentes facultades y escuelas, lo cual está provocando una gran preocupación en el directorio general, debido a que está alterando el pago a tiempo a su personal, los recursos destinados al mantenimiento de su infraestructura y las ganancias a la institución, esto sucede debido que no se está controlando la morosidad en los alumnos en el departamento de finanzas.

Es por esto la necesidad de generar un modelo que muestre el pronóstico de los alumnos morosos de la Universidad Peruana Unión y apoye a la toma de decisiones al área de Finanzas; lo cual es una ventaja en el negocio.

Con esta investigación se contribuirá a establecer la automatización de las actividades relacionadas con la morosidad de la universidad, ayudara a definir el comportamiento crediticio de los alumnos de la universidad. Los datos detallados de los alumnos morosos permitirán tomar un mejor control del problema de la morosidad y aplicar las precauciones para evitarlas en el futuro.

II. MODELO DE PREDICCION

Para el pronóstico de morosidad de los alumnos de la Universidad Peruana Unión se construyó y validó una encuesta para la recolección de datos, la cual fue tomada a los alumnos de las diferentes escuelas de la Universidad.

El modelo fue construido con 800 datos recolectados mediante la encuesta, teniendo diferentes variables que permitieron identificar las características de morosidad de los alumnos. La herramienta que se utilizó fue el SQLServer 2008 de Microsoft.

Para el testeo del modelo de árbol de decisión para la morosidad se trabajó con el 70% de la población.

Árboles de Decisión. (Según Departamento de Informática Universidad Nacional de San Luis (UNSL) San Luis. Argentina Octubre de 2006.) El aprendizaje de árboles de decisión es un método que ha sido utilizado en numerosas tareas de aprendizaje inductivo. Es un método de aproximación de funciones robusto a la presencia de datos erróneos y es capaz de aprender expresiones disyuntivas.

Existe toda una familia de algoritmos de aprendizaje de árboles de decisión que incluye a algoritmos muy conocidos como ID3, ASSISTANT y C4.5. Esta familia de algoritmos, referenciada a veces como TDIDT (Top-Down Induction of Decision Trees) se caracteriza por buscar en un espacio de hipótesis completamente expresivo que evita las dificultades de los espacios de hipótesis restringidos. Su sesgo inductivo es un sesgo de preferencia por árboles pequeños sobre árboles grandes.

(Crossland M.D, 1995), menciona que los árboles de decisión son herramientas excelentes para ayudar a realizar elecciones adecuadas entre muchas posibilidades. Su estructura permite seleccionar una y otra vez diferentes opciones, que pueden tener diferentes alternativas que al ser exploradas pueden ser una posible decisión.

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionada por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos.

El algoritmo genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como *nodos*. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el

algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

El algoritmo utiliza la *selección de características* para guiar la selección de los atributos más útiles. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si utiliza demasiados atributos de predicción o de entrada al diseñar un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la *entropía* y las *redes Bayesianas*.

Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está *sobre-ajustado* o *sobreentrenado*. Un modelo sobre-ajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobre-ajustar un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol. (Microsoft, 2008).

Teoría de la Información: Según (Rodilla, 2005)

Determinando la (im) pureza de una partición por entropía:

Un experimento puede tener m resultados distintos v_1,…, v_m que pueden ocurrir con probabilidades $P(v_1)$,…, $P(v_m)$, entonces la cantidad de información I que se obtiene al conocer el resultado real del experimento es:

$$I(P(v_1), \dots P(v_m)) = \sum_{i=1}^{m} -P(v_i) log_2 P(v_i)$$

Para ejemplificar esta idea, consideremos el experimento de arrojar una moneda, el cual tiene como resultados posible $cara$ y $sello$. Si conocemos de antemano que la moneda fue alterada para que siempre caiga $cara$, la entropía (información) I del resultado del experimento será:

$$I(P(cara), P(sello)) = I(1,0) = -1log_2 1 - 0log_2 0$$

Este resultado significa que, dado que ya sabemos que la moneda caerá $cara$, la información que obtengamos al conocer el resultado del experimento será nula. Si en cambio utilizamos una moneda totalmente balanceada, que produce cualquiera de los dos resultados en forma equiprobable, tendremos que:

$$I\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} log_2 \frac{1}{2} - -\frac{1}{2} log_2 \frac{1}{2} = 1$$

Como podemos observar la entropía tiene su valor más bajo (0) cuando existe total certeza en el resultado del experimento, mientras que el mayor valor de entropía es alcanzado en el caso de mayor incertidumbre (eventos equiprobables). Entre estos dos valores extremos tendremos toda una serie de distribuciones de probabilidad válidas caracterizadas por tener una entropía baja cuando existen eventos altamente probables. Así por ejemplo, si la moneda es alterada para caer cara en un 99% de los casos, tendremos que I(99 100 ; 1 100) = 0;08.

Metodología de investigación: La metodología que se utilizó para esta investigación fue el modelo de CRISP-DM que es un modelo de minería de datos, el cual está estructurado en seis fases.

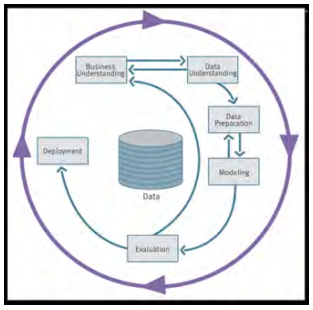


Figura 1. Modelo CRISP-DM

Fases del modelo CRISP-DM:

1. Comprensión del Negocio: establecimiento de los objetivos del negocio, evaluación de la situación, generación del plan del proyecto.

2. Comprensión de los Datos: recopilación inicial de datos; descripción, exploración y verificación de la calidad de datos.

3. Preparación de Datos: selección de datos construcción e integración de los datos.

4. Modelado: Aplicación de las técnicas de minería de datos; selección y diseño de la evaluación, construcción del modelo de árboles de decisión y la evaluación respectiva.

5. Evaluación: Evaluación de los resultados del modelo, de acuerdo a las necesidades del negocio y establecimiento de los pasos a seguir.

6. Despliegue: Integración el resultado del modelo a las actividades del negocio, planificación, monitorización y revisión del proyecto.

III. ANÁLISIS DEL MODELO UTILIZADO

Análisis de la eficiencia del modelo de arboles de decisión.

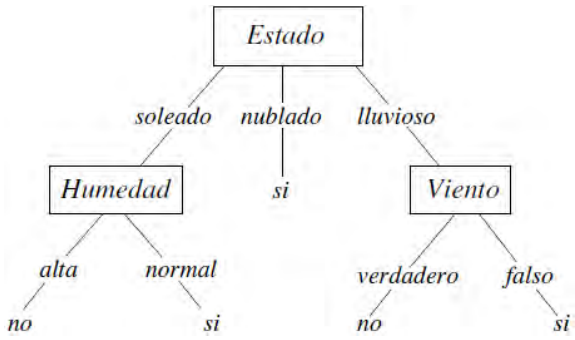


Figura Nro. 02. Ejemplo de Jugar Tenis.

Este ejemplo de árboles de decisión, se trata de decidir si vamos a jugar tenis dependiendo, de las condiciones atmosféricas siguientes: nubosidad, humedad y viento.(cielo=soleado, temperatura=caliente,humedad=alta, viento=fuerte)

Día	Cielo	Temperatura	Humedad	Viento	Jugar_tenis
D1	Soleado	Alta	Alta	Débil	No
D2	Soleado	Alta	Alta	Fuerte	No
D3	Lluvioso	Alta	Alta	Débil	Sí
D4	Lluvioso	Media	Alta	Débil	Sí
D5	Lluvioso	Fría	Normal	Débil	Sí
D6	Lluvioso	Fría	Normal	Fuerte	No
D7	Lluvioso	Fría	Normal	Fuerte	Sí
D8	Soleado	Media	Alta	Débil	No
D9	Soleado	Fría	Normal	Débil	Sí
D10	Lluvioso	Media	Normal	Débil	Sí
D11	Soleado	Media	Normal	Fuerte	Sí
D12	Lluvioso	Media	Alta	Fuerte	Sí
D13	Lluvioso	Alta	Normal	Débil	Sí
D14	Lluvioso	Media	Alta	Fuerte	No

Entropía ([9+,5-])= - (9/14) log2 (9/14) -(5/14) log2 (5/14) = 0'940.

Supongamos que S es un conjunto de entrenamiento con 14 ejemplos

A.9 ejemplos positivos y 5 negativos ([9+,5-]).
B.Unos de los atributos, Viento, puede tomar los valores Débil y Fuerte.

C.La distribución de ejemplos positivos y negativos según los valores de Viento son.

	Positivos	Negativos
Débil	6	2
Fuerte	3	3

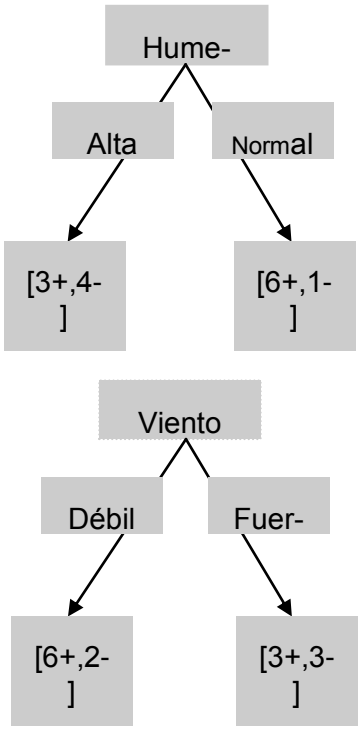
La ganancia de información que obtenemos si clasificamos los 14 ejemplos según el atributo Viento es:

Ganancia(S, A) = Entropía(S) - $\sum_{ve\ Valores\ (A)} \frac{|S_v|}{|S|} Entropia(S_v)$

= Entropía (S) – $|8|/14\ Entropia\ (S\_Débil\ )$ - $|6|/14\ Entropia\ (S\_Fuerte\ )$

= 0.940- $\frac{|8|}{14}$ 0.811- $\frac{|6|}{14}$ 1.00

=0.048



Ganancia(S, Humedad)

= 0.940–(7/14)*0.985

– (7/14)*0.592

= **0.151**

Ganancia(S, Viento)

= 0.940–(8/14)*0.811

– (6/14)*1.00

= **0.048**

Aplicación Del Modelo De Arboles De Decisión:

El pronóstico de las características de morosidad del alumno se mide en dos escalas: Moroso y No moroso. Para identificar las características de morosidad del alumno se ha considerado cinco variables predominantes: Ayuda Institucional, Ingreso de los padres, Monto de Crédito, Tarjetas de crédito y la Situación laboral del padre.

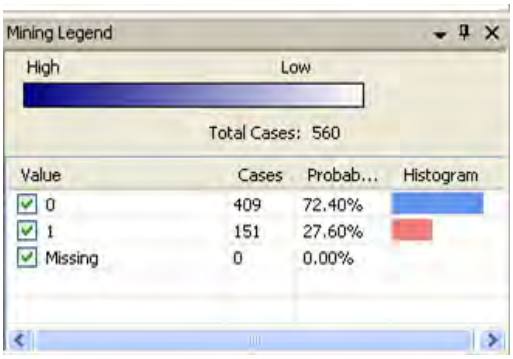


Figura 3. Cuadro de Pronostico.

En la Figura 3. Se observa el porcentaje de los datos analizados con pronóstico de morosidad donde el 72.40% de los alumnos no son morosos y el 27.60% son morosos, estos datos se analizaron en un 70% de toda la data.

La variable predomínate en las características de los alumnos morosos es el sueldo de los padres, el cual se divide en dos porcentajes: menores y mayores igual de 1600 soles. En la figura 4 se muestra como resultado de la principal característica un total de 419 registros, donde el 36.09% de los alumnos tienen las características de morosidad.

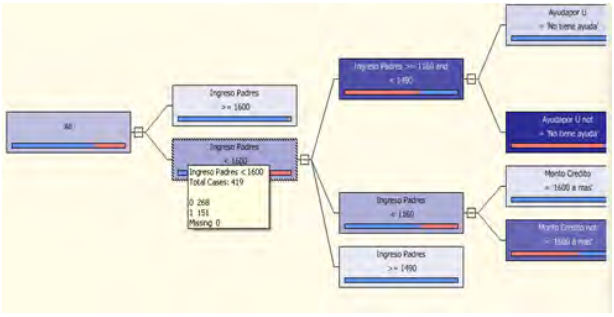


Figura 4. Modelo de morosidad, variable sueldo de padres.

La característica de morosidad ingreso de padres se divide en tres nodos, se observa en la figura 5, que la mayor cantidad de morosos con un porcentaje de 65.44% se encuentran en el nodo de ingreso de padres de un rango entre 1160 y 1490 soles; el cual cuenta con un total de 139 casos.

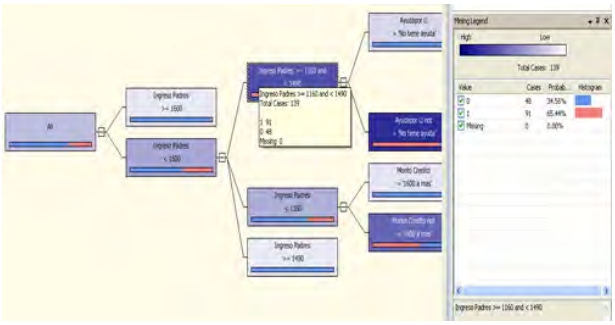


Figura 5. Modelo de morosidad, variable sueldo de padres.

En la Figura 6. Se aprecia otra ramificación del árbol con la variable ayuda económica; con 106 casos, donde el 81.07% de los alumnos morosos tiene esa característica.

Figura 6. Modelo de morosidad, variable ayuda económica.

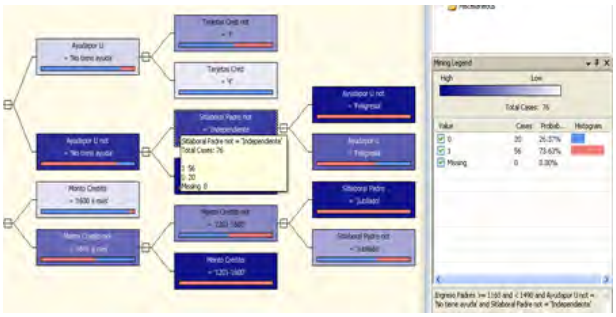


Figura 6. Modelo de morosidad, variable ayuda económica.

Otra característica de morosidad de los alumnos es la variable tarjetas de crédito, en la Figura 7, se observa que 41.68% de alumnos con un total de 12 casos presenta esta característica.

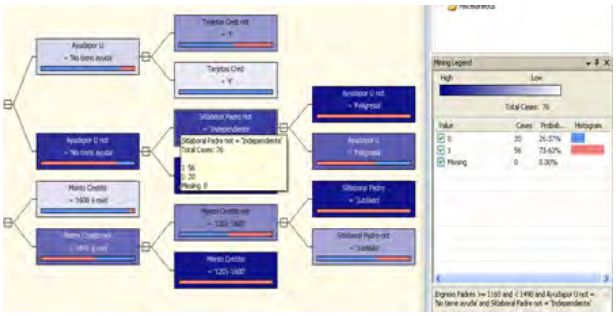


Figura 7. Modelo de morosidad, variable tarjetas de crédito.

La ultima característica predominante en la morosidad de los alumnos es la situación laboral independiente de los padres con un total de 76 casos, con el 73.63% de los alumnos morosos. Figura 8.

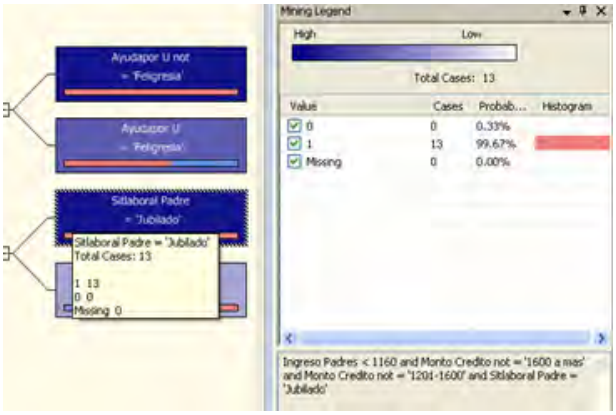


Figura 8. Modelo de morosidad, variable laboral de padres.

En la Figura 9. Se aprecia la estructura de minería de datos de PRUEBAHO, que nos muestra que de un 30% de la población, el 79.37% tienen características de morosidad con un 36.58% de probabilidad de predicción.

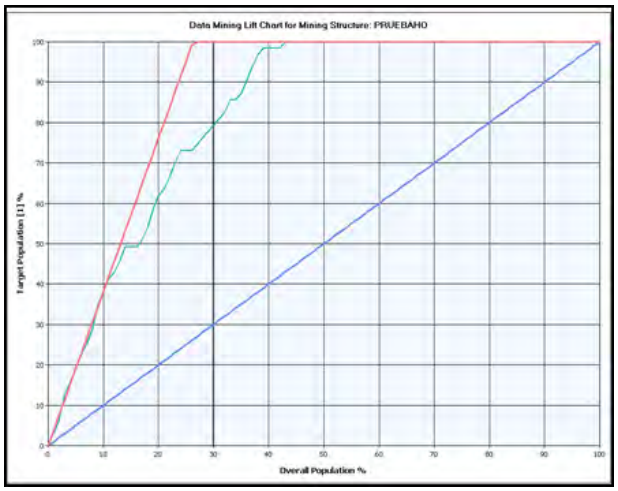


Figura 9. Estructura de Minería de Datos.

IV. CONCLUSION

Usando el algoritmo de árboles de decisión se detectaron patrones para los morosos y los no morosos. El principal patrón detectado es que el ingreso económico de los padres sea menor a 1600 nuevos soles, que no tenga ayuda de la universidad, que la situación de los padres es independiente (jubilado), y que tengan un monto creditico 1201 y 1600 nuevos soles con la universidad. Este patrón caracteriza al 33,75% de los alumnos son morosos.

Aplicando, este tipo de investigaciones en este el rubro, que es el brindar créditos educativos, estaremos previniendo futuros endeudamientos y falta de pagos en las instituciones. Además sabremos a qué tipo de clientes podremos otorgarle un crédito de acuerdo a las variables definidas de acuerdo al análisis realizado.

El estudio realizado posee un margen de error, no se puede afirmar del todo que un cliente que no posee la cantidad requerida de ingresos tienda a ser deudor, pero posee los indicios. Además los que poseen una elevada cantidad de ingreso económico, tiendan a ser deudores por más que el modelo de estudio diga lo contrario.

REFERENCIAS

Breiman L, Friedma J, Olshen R, Stone C. 1984. Classification and regression trees. Editorial Pacific Grovic. 485p.

Vallejos Sofia. 2006. Diseño y Administración de Datos. Argentina: Editorial Corrientes. 352p.

Lara G. 2008. La Técnica del árbol para la toma de decisiones. México: Univalle. 350p.

Rodilla, V. 2005. Inteligencia Artificial e Ingeniería del Conocimiento. México. Mc Graw Hill. 550p.

Pérez C. 2007. Data mining: Soluciones con Enterprise Miner. Editorial. Paraninfo. 455p.

Vitt Elizabeth, L.M., Misner Stacia. 2002. Business Intelligence: Técnicas de análisis para la toma de decisiones estratégicas.

Matich, D. J. (marzo del 2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Informatica Aplicada a la Ingenieria de Procesos - Orientacion I (págs. 12-13). Univercidad Tecnológica Nacional - Facultad Regional Rosario Depart de Ingen y Quimica.

Martinez González, D. (2004-2005). Redes Neuronales Artificiales y Mapas Auto Organisados. Curso (Sistemas Expertos e Inteligencia Artificial) 2004-2005 . Ciudad Univercitaria de Burgos, Burgos.

Departamento de Informática Universidad Nacional de San Luis (UNSL) San Luis. Argentina Octubre de 2006. Aprendizaje de árboles de decisión y Minería de Datos.

Referencias electrónicas

GrupoAnts. 2009. Arboles de Inducción. [Articulo en linea] SlideShare. [Consultado en 18 de Octubre de 2010] Formato html. Disponibilidad libre en: http://www.slideshare.net/EliteAstarothJG/arboles-de-induccion.

Grupo de Estudios en Metodologías de Ingeniería de Software. 2009. Ingeniería de Proyectos de Explotación de Información. [Articulo en linea] SlideShare. [Consultado el 20 de octubre del 2010] Formato html. Disponibilidad libre en: http://posgrado.frba.utn.edu.ar/investigacion/articulos-y-comunicaciones/WICC-2010-172-176.pdf

Microsoft. 2008. Algoritmo de árboles de decisión de Microsoft. [Consultado el 15 de Octubre] Formato html. Disponibilidad libre en: http://technet.microsoft.com/es-es/library/ms175312.aspx

MicroStrategy, Corp. 2009. Microstrategy en la versiona 8i [http://www.microstrategy.com/] (Consultado el 25 de Octubre del 2010).

Agradecimientos

Se agradece al equipo de trabajo Nils Ferro Quintanilla, Jonahtan Ander Marlo Salazar por haber mejorado el articulo Modelo de Árboles de Clasificación para pronosticar la morosidad de los alumnos de la Universidad Peruana Unión.