Sistema de enseñanza para la técnica de clasificación de árboles de decisión

A. Franco-Arcega, F.A. Castro-Espinoza y P. Cortes-García

Área Académica de Computación
Universidad Autónoma del Estado de Hidalgo
Carretera Pachuca-Tulancingo Km. 4.5, C.U.
Pachuca, Hidalgo, México, C.P. 46600

Resumen Los Árboles de Decisión son una de las técnicas de clasificación supervisada más utilizada para resolver problemas. Por este motivo, es necesario que usuarios, principalmente de las áreas relacionadas a computación, comprendan el procedimiento que se lleva a cabo para construir un Árbol, con la finalidad de poder aplicar este tipo de técnicas en la solución de problemas cotidianos. Actualmente, se ha detectado que el transmitir este conocimiento utilizando técnicas de enseñanza tradicionales no es una tarea fácil, ya que en la definición de un problema real puede existir una gran cantidad de atributos que lo describa, lo que conlleva a tener más operaciones matemáticas para generar un Árbol. En particular, en este trabajo se propone automatizar la herramienta de estudio sobre la técnica de clasificación supervisada denominada Árboles de Decisión desarrollando un sistema de enseñanza computacional. Este sistema tendrá los objetivos de mostrar cómo se construye un Árbol de Decisión paso a paso, analizando cada una de la operaciones matemáticas realizadas para tal fin.

Palabras clave: sistema de enseñanza, árboles de decisión, minería de datos.

1. Introducción

En la actualidad, la tecnología computacional ayuda a resolver problemas y satisfacer necesidades cotidianas. El ámbito académico es uno de los principales entornos en donde la tecnología introduce sus beneficios. Expertos de la computación se han interesado en brindar a los alumnos material académico que los lleve de la mano para la comprensión de ciertos temas complejos. Algunos trabajos [1,2,3] han propuesto diferentes herramientas para la comprensión de temas relacionados al procesamiento de datos con técnicas de computación inteligente.

En materias como Minería de Datos, Reconocimiento de Patrones o Inteligencia Artificial se trabaja con este tipo de técnicas inteligentes, con las cuales se resuelven problemas de predicción o descripción [4]. Los Árboles de Decisión son una de las herramientas más utilizadas para darle solución a problemas de tipo predictivo. Un Árbol de Decisión es una técnica que ayuda a tomar decisiones

adecuadas entre muchas posibilidades, aproxima funciones de valores discretos de destino y puede ser representado como un conjunto de reglas para mejorar la legibilidad humana. Estos métodos de aprendizaje se encuentran entre los más populares de los algoritmos de inferencia inductiva y se han aplicado con éxito a una amplia gama de tareas [5]. Un Árbol de Decisión es una estructura que contiene nodos y ramas. Los nodos de un árbol pueden ser de dos tipos: internos o hojas. Los nodos internos se caracterizan por tener uno o más atributos de prueba, los cuales ayudan a decidir qué camino es más satisfactorio para recorrer el árbol. Por su parte, los nodos hoja contienen etiquetas de clase, las cuales definen la decisión a tomar.

Uno de los problemas que se detecta en el ámbito académico es que no es fácil trasmitir el conocimiento de este tipo utilizando técnicas de enseñanza tradicionales, como papel y lápiz, y por consecuencia es difícil de comprender por los alumnos. Mientras más atributos están involucrados en la definición del problema, más operaciones matemáticas son utilizadas para la generación de un Árbol de Decisión y por tanto para una efectiva toma de decisiones. Por esta razón, se han desarrollado diversos sistemas que de manera automática ejecutan todas las operaciones matemáticas para llevar a cabo el procedimiento de alguna técnica en específico. En particular, en este trabajo se propone desarrollar un sistema de enseñanza de la Técnica de Árboles de Decisión, con el fin de que el alumno pueda apreciar cómo se construye un árbol paso a paso, analizando cada una de las operaciones matemáticas realizadas.

El documento está organizado de la siguiente forma: en la sección 2 se describen los principales sistemas que permiten la creación de Árboles de Decisión. En el apartado 3 se describe el sistema propuesto en el presente artículo detallando sus funcionalidades y las etapas de creación de un Árbol de Decisión. En la sección 4 se presentan los resultados preliminares del sistema propuesto. Finalmente, en el apartado 5 se presentan las conclusiones y trabajos futuros del presente trabajo de investigación.

2. Trabajos relacionados

Actualmente, se han desarrollado algunos sistemas que permiten la construcción de Árboles de Decisión. Sin embargo, ninguno de estos sistemas muestra paso a paso el proceso que se lleva a cabo para esta generación, ya que en los sistemas simplemente se visualiza como queda el árbol final. Al usuario no se le presenta el procedimiento de construcción, por lo cual no podrá comprender que operaciones matemáticas se realizaron para llevar a cabo el proceso.

Precision Tree es un sistema que permite desarrollar Árboles de Decisión que proveen una estructura formal en la que las decisiones y los acontecimientos casuales están relacionados en secuencia de izquierda a derecha. Las decisiones, los eventos al azar, y los resultados finales se representan mediante nodos conectados por ramas [6]. El resultado es una estructura de Árbol con la raíz en la izquierda y los caminos diversos a la derecha. Aunque visualmente se entiende

una decisión, ya que identifica las mejores opciones y comunica los resultados, el proceso de construcción no se muestra paso a paso.

El sistema GATree hace uso de algoritmos genéticos para evolucionar o construir árboles binarios de Decisión que ajuste al concepto de destino. Por ello GATree adopta una representación natural del espacio de búsqueda usando Árboles de Decisión reales y no de cadenas binarias [7]. GATree es un constructor de Árboles de Decisión que se basa en algoritmos genéticos, el cual asume que si se tienen amplios recursos, entonces se podrá esperar un árbol cada vez de mejor ajuste. GATree puede proporcionar un conjunto de Árboles de Decisión que son totalmente diferentes, pero que son opciones cercanas al espacio de solución. Al igual que el sistema anterior, GATree no permite al usuario visualizar el proceso de construcción, simplemente muestra el árbol terminado.

SMILES es un sistema de aprendizaje automático que integra muchas características diferentes de las técnicas de otras máquinas y paradigmas de aprendizaje. En particular, SMILES tiene un manejo de los recursos sofisticado y muy eficaz. De esta manera, SMILES combina y mejora el reciente interés en combinación de hipótesis sensibles a los costes de aprendizaje [8]. Sus aplicaciones son básicamente para la Minería de Datos y cualquier otra tarea de aprendizaje donde los Árboles de Decisión podrían ser útiles. Este sistema aunque es muy robusto en cuanto al manejo de los recursos, no es apropiado para la enseñanza de la técnica de árboles, ya que el usuario no comprendería todas las características de él, además de que no observaría de manera detallada la generación de un árbol.

See5 es una sofisticada herramienta de Minería de Datos para descubrir patrones que definen categorías, los cuales pueden ser utilizados para hacer predicciones [9]. See5 ha sido diseñado para analizar bases de datos sustanciales que contienen miles de millones de registros y decenas a cientos de atributos de tipo numérico, hora, fecha o campos nominales. Para maximizar la interpretabilidad, See5 expresa sus resultados como Árboles de Decisión o como conjuntos de reglas if-then. Aunque See5 es fácil de usar y no supone ningún conocimiento especial de Estadística o Machine Learning, no le permite al usuario entender el proceso interno que se lleva a cabo para llegar al resultado final.

Angoss es un software de análisis predictivo que ayuda a las empresas a descubrir información valiosa en sus datos, como el descubrimiento de oportunidades para aumentar las ventas y la rentabilidad, y reducir el riesgo [10]. Angoss permite procesar atributos discretos (categóricos) y continuos. Además, el usuario puede definir algunas preferencias en el crecimiento del Árbol (por ejemplo, árboles binarios o n-arios), tipos de valor de agrupación y los intervalos abiertos. Este sistema muestra un informe acerca de las reglas de cada nodo y las estadísticas para cada decisión (nodos hoja). Debido a que Angoss está enfocado a la solución de análisis en empresas, no es viable para ser utilizado en la enseñanza de esta técnica de predicción.

Weka es una herramienta desarrollada en lenguaje Java por la Universidad de Waikato, Hamilton, New Zeland [11]. Weka permite la extracción de conocimiento desde bases de datos y debido a que fue desarrollado bajo licencia GNU-

GPL es una de las suites más utilizadas actualmente en las áreas de aprendizaje automático y minería de datos. Esta herramienta permite al usuario aplicar diversos algoritmos de clasificación , agrupamiento, selección de variables, entre otros. Sin embargo, Weka no muestra al usuario como se lleva a cabo el procesamiento de las bases de datos, únicamente permite cargar en memoria la base de datos y elegir el algoritmo a aplicar, para después sólo mostrar el resultado final del procesamiento.

3. Sistema de enseñanza propuesto

Los Árboles de Decisión son una herramienta valiosa que permite solucionar diversos problemas de clasificación supervisada o predicción. Para que un usuario pueda saber como aplicar esta herramienta en alguna situación real, es necesario que comprenda su funcionamiento. Un Árbol de Decisión ayuda a la toma de decisiones efectivas, ya que sus nodos hoja contienen etiquetas de clase que son asignadas a un nuevo caso. Es importante que el usuario entienda como se construye un árbol para que visualice los patrones que ayudan en la toma de decisiones, así como para observar que atributos son importantes en su problema.

El Sistema propuesto, denominado SETECAD, muestra paso a paso la construcción de un Árbol de Decisión tomando en cuenta un orden jerarquizado, es decir, por niveles de nodos, los cuales se identificarán fácilmente. Además, las operaciones matemáticas que son utilizadas para expandir los nodos internos, también se muestran en una ventana del sistema.

El sistema que permite enseñar la técnica de Clasificación Supervisada de Árboles de Decisión fue desarrollado en lenguaje Java, utilizando elementos gráficos y de dibujo de este lenguaje. Se tiene una interfaz principal y diferentes ventanas que muestran variados elementos que sirven para la comprensión del proceso de construcción de un Árbol de Decisión. Estas ventanas son: Lectura de datos, creación de un Árbol de Decisión, desarrollo computacional, formulario, ayuda e impresión. En las siguientes subsecciones se describen cada una de las funcionalidades del sistema SETECAD. La figura 1 muestra la interfaz principal del sistema SETECAD.

3.1. Lectura de datos

Esta ventana permite abrir el archivo que contiene los datos que van a ser utilizados para la construcción del Árbol de Decisión, el cual debe contener como primera línea el nombre de los atributos de dichos datos. Si el conjunto de datos está descrito por n atributos, las n-líneas siguientes serán destinadas para la declaración del tipo de atributos y clase, es decir, los atributos categóricos se declararán con los posibles valores de dicho atributo, los atributos de tipo continuo se declaran con la palabra reservada #NUMERO y la clase con la palabra CLASE antecedido de las etiquetas de clases en dichos datos. La Figura 2 muestra un ejemplo de un documento de entrada que contiene un conjunto de datos. Una vez cargada la información se almacena en memoria y se detalla en pantalla

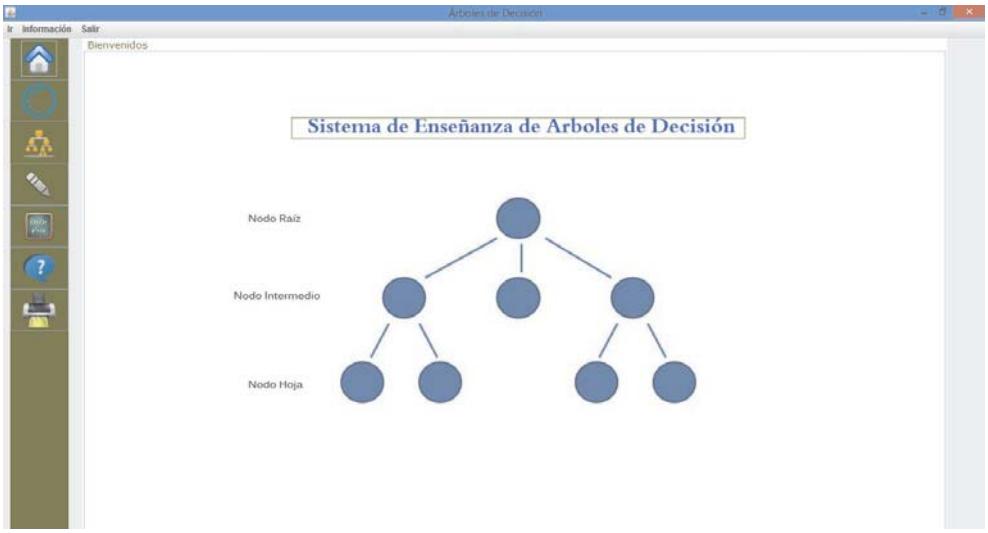

Figura 1. Interfaz principal del sistema SETECAD

el contenido de esta información, en específico, se muestra el número total de objetos contenidos en el archivo, el número de atributos y si dichos atributos son categóricos o continuos, el número de clases y finalmente, se muestra una porción de los datos antes almacenados. Esto se muestra en la figura 3

3.2. Creación de un árbol de decisión

En esta ventana se construye gráficamente el Árbol, paralelamente a los cálculos que se realizan en la ventana de Desarrollo Computacional, es decir, conforme se van realizando las iteraciones se va actualizando el gráfico que muestra el árbol construido. Esta construcción inicia en el nodo raíz, eligiendo al atributo que contenga mayor ganancia de información (ver sección 3.4), y posteriormente se dividen los objetos de entrenamiento de acuerdo a dicho atributo seleccionado. En la Figura 4 se observa cómo se grafica el nodo raíz una vez expandido, enumerando los nodos generados en el árbol de dicisión. La Figura 5 muestra el árbol actualizado una vez que se ha expandido el nivel dos de éste.

3.3. Desarrollo computacional

Esta ventana imprime el procedimiento metodológico y matemático para la construcción de un Árbol de Decisión. La parte metodológica muestra los pasos que se llevan a cabo para la expansión de los nodos, especificando al usuario las instrucciones realizadas en el proceso de expansión. Además el usuario puede observar las operaciones matemáticas calculadas para tal fin. La Figura 6 muestra una parte de lo que se puede observar en esta ventana de desarrollo computacional.

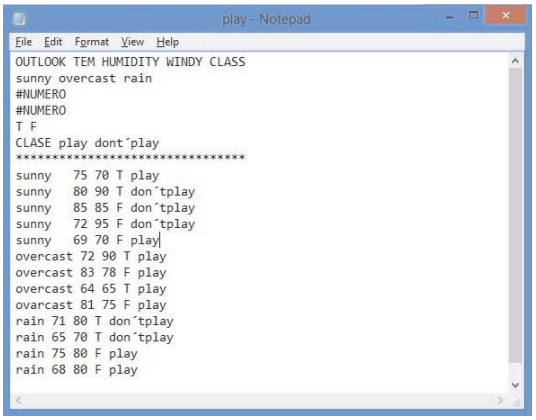

Figura 2. Formato del archivo

3.4. Formulario

En cada nodo interno de un Árbol de Decisión se debe seleccionar un atributo para dividir este nodo. La teoría de la información, la cual está basada en la entropía, se toma como referencia para elegir al mejor atributo en este sistema propuesto. Entre más pequeño sea el valor de la entropía, menor será la incertidumbre y más útil será el atributo para la clasificación. Para encontrar la medida en como un conjunto S es dividido de acuerdo a sus clases se suman las frecuencias de la proporción de cada clase i, siguiendo la ecuación 1.

$$\text{ínf}(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) \tag{1}$$

Si se considera este proceso para encontrar la medida en cómo se particiona el conjunto S, de acuerdo a un número n de posibles valores de un atributo en específico, se calcula la suma ponderada de cada subconjunto de objetos. La ecuación 2 define esta operación.

$$\inf_x(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{ínf}(S_i) \tag{2}$$

Utilizando las medidas anteriores se puede obtener la ganancia de información de un atributo X, basada en la disminución de la entropía del conjunto S. La ecuación 3 muestra la fórmula de la ganancia de información.

$$gain\,(X) = \text{ínf}\,(S) - \inf_x\,(S) \tag{3}$$

Cuando un atributo contiene una gran cantidad de posibles valores, especialmente los atributos numéricos, éstos deben procesarse de una manera diferente. La ganancia de información de un atributo de este tipo debe ser normalizada de

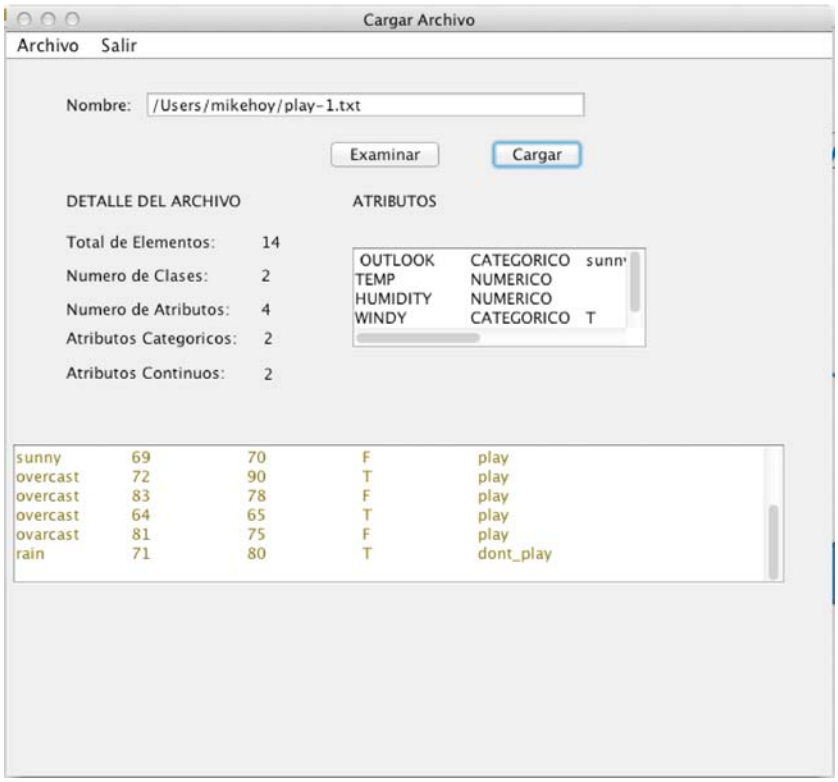

Figura 3. Ventana de lectura de datos del sistema

acuerdo al número de posibles valores que tenga el atributo. Esta normalización se hace con la ecuación 4. Finalmente, la proporción de ganancia de información (Gain Ratio) se efectúa aplicando la ecuación 5.

$$split(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2 \left(\frac{S_i}{S} \right) \tag{4}$$

$$ratio\,(X) = \frac{gain\,(x)}{split\,(x)} \tag{5}$$

El atributo con la mayor ganancia de información es el elegido para la división del nodo. Las fórmulas expuestas en esta sección se muestran en la ventana de Formulario del sistema de enseñanza propuesto.

3.5. Ayuda

Al dar clic en el botón de ayuda se abre un archivo que al usuario le permitirá saber como es el funcionamiento de dicho sistema. La Figura 7 muestra dicho documento.

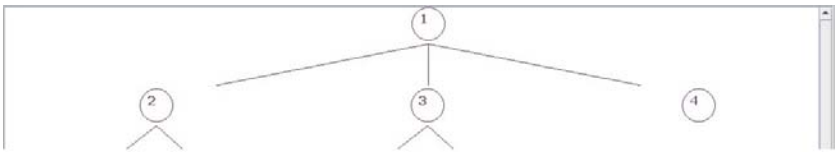

Figura 4. Construcción con un nivel

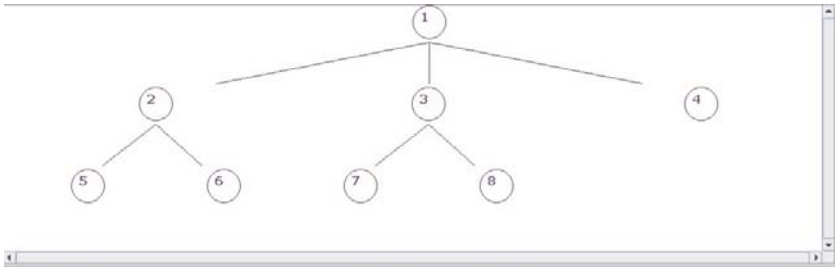

Figura 5. Construcción con dos niveles

```
Valor 3 :                                                    2
  Calculo con Eq. 1 :                                        0.0
  Calculo con Eq. 2 :                                        6.730116670092565
Ganancia del atributo (Eq. 3) :                              0.48072261929232607

Número de Atributo :                                         2
Tipo de Atributo :                                           Numérico
Paso 1.1: Ordenar los valores númericos del atributo
Paso 1.2: Para cada par de valores del atributo X, Y obtener el promedio entre ellos (Xp) y calcular la Ganancia
          particionando los objetos en el nodo con la condición if/else (valorAtributo <= Xp)
  Promedio (Xp) 1 :                                          64.5
    Valores Atributo <= Xp
      Calculo con Eq. 1 :                                    0.0
      Calculo con Eq. 2 :                                    0.0
    Valores Atributo > Xp
      Calculo con Eq. 1 :                                    0.666278442414676
```

Figura 6. Desarrollo computacional

Figura 7. Manual de usuario

4. Resultados preliminares

El sistema SETECAD se les ha proporcionado, como apoyo educativo, a 5 estudiantes de la materia de Minería de Datos, de la Maestría en Ciencias Computacionales del Centro de Investigación en Tecnologías de Información y Sistemas (CITIS) de la Universidad Autónoma del Estado de Hidalgo (UAEH). Los avances académicos mostrados por los 5 estudiantes han sido importantes, en comparación con los obtenidos por otros estudiantes en anteriores versiones de la misma materia. Minería de Datos se imparte semestralmente a un promedio de entre 5 y 10 alumnos, sin embargo, estos resultados no podrán ser validados de manera experimental hasta no evaluar el sistema con grupos de estudiantes homogéneos, en cuanto a su nivel de conocimiento en Árboles de Decisión y de tamaño considerable. Por tal razón se está planificando una serie de experimentos que permitan validar la ventaja de utilizar el sistema SETECAD para enseñar Árboles de Decisión. Se propone utilizar SETECAD en la materia de Reconocimiento de Patrones cuyos grupos son alrededor de 30 estudiantes, en 4 grupos, y el procedimiento es dividir cada grupo en 2 partes, a una se le impartirá la materia sin permitirles utilizar SETECAD, y a los restantes que reciban como apoyo a su aprendizaje el sistema, y al final del tópico de Árboles de Decisión, realizar un análisis para determinar el nivel de aprovechamiento académico mostrado por ambos grupos que permita una evaluación experimental más realista de la utilización del sistema SETECAD.

5. Conclusiones

Las técnicas de clasificación supervisada son de gran ayuda para la solución de problemas de tipo predictivo. Los Árboles de Decisión son uno de los preferidos debido a su simplicidad y fácil entendimiento. Sin embargo, enseñar esta herramienta con técnicas de enseñanza tradicionales resulta complicado ya que son utilizadas diferentes operaciones matemáticas, las cuales se vuelven costosas cuando el número de atributos en un conjunto de entrenamiento a procesar es grande. En este trabajo se desarrolló un sistema computacional que permite generar Árboles de Decisión. Este sistema pretende mostrar el conocimiento de dicha técnica de enseñanza de una manera más dinámica, con la finalidad de comprender el por qué y de dónde sale cada resultado y al mismo tiempo darle al usuario la oportunidad de descubrir para que se utiliza. Además, SETECAD llevará de la mano a quien interactúe con él, para la comprensión de dicho algoritmo sin necesidad de realizar los cálculos manualmente. Como trabajo futuro se contempla la mejora del sistema propuesto, de acuerdo a los resultados obtenidos con los grupos de estudiantes de la materia de Reconocimiento de Patrones.

Agradecimientos. Los autores agradecen el apoyo brindado por PROMEP para el desarrollo de este trabajo con el recurso asignado en la carta de liberación PROMEP/103.5/12/8114, folio UAEH-PTC-554.

Referencias

1. Aybars Ugur, Ahmet Cumhur Kinaci.Web-Based Tool for Teaching Neuronal Network Concepts. Computer Appliccations in Engineering Education. Ed ISSUE 3.Volumen 18. PAG 449-457. 2010
2. Melvin Ayala, Malek Adjouadi, Mercedes Cabrerizo, Armando Barreto. A Windows-Based Interface for Teaching Image Processing. Computer Appliccations in Engineerring Education. Ed ISSUE 2. Volumen 18.PAG 213- 224. 2010
3. Reyes Juárez-Ramírez, Guillermo Licea, Alfredo Cristo Bal-Salas. Teaching Undergraduate Students to Model use cases using Tree Diagram. Computer Appliccations in Engineerring Education. Ed ISSUE 1.Volumen 18.PAG 77-86. 2010
4. Tom M. Mitchell. 1997. Machine Learning. Ed. Reviews. PAG 52
5. Tan, P.N., Steinbach, M. and Kumar, V. 2006. Introduction to Data Mining. Addison Wesley.
6. Data Mining Communitys Top Resource, `http://www.palisade.com/precisiontree/default.asp`
7. Data Mining Communitys Top Resource, `http://http://www.gatree.com/?page_id=6`
8. Data Mining Communitys Top Resource, `http://users.dsic.upv.es/~flip/smiles/`
9. Data Mining Communitys Top Resource, `http://www.rulequest.com/see5-info.html`
10. Data Mining Communitys Top Resource, `http://www.angoss.com/predictive-analytics-software/overview`
11. Weka - Data Mining with Open Source Machine Learning Software, `http://www.cs.waikato.ac.nz/ml/weka/`