Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

Tipo de artículo: Artículo original
Temática: Tecnologías de la información y las telecomunicaciones
Recibido: 26/05/2015 | Aceptado: 14/09/2015

Desempeño de algoritmos de minería en indicadores académicos: Árbol de Decisión y Regresión Logística

Performance of data mining algorithms in academic indicators: Decision Tree and Logistic Regression

Iván Menes Camejo[1, 2*], Gloria Arcos Medina[1], Plácido Moreno Beltrán [1,2], Katherine Gallegos Carrillo[1]

[1] Escuela Superior Politécnica de Chimborazo. Panamericana Sur km 1[1/2], Riobamba - Ecuador. {kgallegos, garcos, placido.moreno }@espoch.edu.ec

[2] Grupo de Investigación AINTO – Aplicaciones Informáticas para la Toma de Decisiones – ESPOCH]

[*] Autor para correspondencia: imenes@espoch.edu.ec

Resumen

La minería de datos se orienta a la presentación prospectiva de información, y para ello, es necesario escoger un algoritmo apropiado que ofrezca los mejores resultados, según el tipo de datos y los objetivos del proyecto. En este documento se presenta un estudio de desempeño de los algoritmos de minería de datos: Árbol de Decisión y Regresión Logística, aplicados a los datos continuos y discretos generados por la función académica de una institución de educación superior. Se buscó determinar el algoritmo con el mejor desempeño a través del uso del método científico y técnicas de estadística descriptiva e inferencial, y los resultados presentan que: no existe una diferencia significativa en el uso de RAM de los algoritmos, el algoritmo de Árbol de Decisión tiene menor tiempo de respuesta, y mayor precisión que el de Regresión Logística, mientras que este último tiene un mejor uso de CPU, concluyendo que el algoritmo de Árbol de Decisión es el de mejor desempeño para el escenario planteado.

Palabras clave: análisis de desempeño, indicadores académicos, árbol de decisión, regresión logística, minería de datos.

Abstract

Data mining is aimed at prospective reporting, for which is necessary to choose the most appropriate algorithm, i.e. the one that provides the best results, depending on data types and project objectives. In this paper a study of performance of two data mining algorithms is presented, namely Decision Tree and Logistic Regression, which have been applied to continuous and discrete data generated by the academic function of an institution of higher education. We sought to determine the algorithm with the best performance by means of the scientific method and

descriptive and inferential statistical techniques. The results show that the decision tree algorithm is the best algorithm in terms of indicators of response time, CPU usage, RAM usage and accuracy.

Keywords: *performance analysis, academy indicators, decision tree, logistic regression, data mining.*

Introducción

Data Mining es el proceso de analizar datos usando metodologías automatizadas para encontrar patrones escondidos (MacLennan, 2008). Los procesos de minería de datos apuntan al uso del conjunto de datos generado por un proceso o negocio, con el fin de obtener información que apoye a la toma de decisiones de los niveles ejecutivos (Fayyad, y otros, 1996) (Han, 2006); a través de la automatización del proceso de encontrar información predecible en grandes bases de datos y la respuesta a preguntas que tradicionalmente requerían un intenso análisis manual (Vallejos, 2006). Por su definición, la minería de datos es aplicable a los procesos educativos (Huebner, 2013), tal es así que a nivel investigativo se ha formado una rama denominada Minería de Datos Educativa, en la que se hace uso de las técnicas de búsqueda de patrones y predicción, para hallar información que aporte a mejorar la calidad educativa (Merceron, y otros, 2005).

Para aplicar un proyecto de minería de datos, en cualquier tipo de escenario, es necesario realizar un estudio de los algoritmos disponibles, para determinar aquel que mejor se acople a las necesidades del proyecto a realizar (Hernández, y otros, 2004); por tal motivo, se ha desarrollado un estudio sobre el desempeño de los algoritmos de Árbol de Decisión y Regresión Logística de Microsoft, aplicado a los datos académicos de una institución de educación superior.

Estudios anteriores como el de la Universidad de Minho en Portugal (Cortez, y otros, 2006), tomó los datos de los estudiantes de secundaria de dos instituciones públicas del mismo país, para realizar la aplicación de técnicas de predicción. Se probaron tres diferentes propósitos de minería y cuatro métodos de minería de datos. Los resultados obtenidos revelaron que es posible alcanzar una alta precisión en la predicción, dados los datos de dos períodos académicos. En la universidad de Awadh, en India, se condujo un estudio sobre el rendimiento de los estudiantes basados en un grupo de 60 alumnos de diferentes carreras (Kumar, y otros, 2011). Se usó la tarea de clasificación de minería sobre la base de datos de los estudiantes para predecir la división de los mismos. Información como la asistencia, pruebas, seminarios y tareas se recolectaron para predecir el rendimiento al final del período académico.

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas son plasmadas en algoritmos, que después se aplican sobre un conjunto de datos para obtener resultados (Fayyad, y otros, 1996) (Moreno, y otros, 2001). Cada algoritmo está diseñado para aceptar o arrojar diferentes tipos de datos, por lo que desde esa perspectiva se pueden descartar los algoritmos que no aceptan los tipos de datos existentes en cada proyecto (Chapman, 2000). En este escenario, por el uso de datos tanto discretos como continuos, los algoritmos de Árbol de Decisión y Regresión Logística de Microsoft son escogidos sobre los otros implementados por Data Tools de Microsoft Analysis Services (SSDT), una herramienta asociada al desarrollo de bases de datos y proyectos de inteligencia de negocios.

El algoritmo de Regresión Logística es un tipo de análisis estadístico orientado a la predicción de una variable categórica en función de otras variables consideradas como parámetros predictores (Fernández, 2011). Específicamente el algoritmo implementado por Microsoft resulta ser una variante del algoritmo de red neuronal. Este tipo de algoritmo debido a que acepta cualquier tipo de entrada, es considerado como flexible y se ajusta a varias tareas analíticas dentro de la minería de datos, entre las que se pueden mencionar predicción, clasificación y explorar y ponderar los factores que contribuyen a un resultado específico (Microsoft Corporation, 2012).

Por su parte, los árboles de decisión y reglas que usan divisiones invariantes tienen una forma de representación simple, haciendo del modelo de inferencia relativamente sencillo para el entendimiento del usuario (Microsoft Corporation, 2012). Un modelo de árboles de decisión tiene un nodo primario único que representa el modelo y sus metadatos. Debajo del nodo primario aparecen árboles independientes que representan los atributos de predicción que se seleccionan.

Una variable común a los algoritmos descritos anteriormente, es el desempeño, debido a que se define como la característica relacionada con el tiempo de respuesta, uso de recursos (RAM y CPU) y confiabilidad de las operaciones (IBM Corporation, 2003). La confiabilidad de un algoritmo de predicción está dada por la precisión con la que un modelo resultante define el conjunto de datos de entrada. Este factor es cuantificable con la herramienta de análisis Data Tools de Microsoft (Microsoft Corporation, 2012). Se buscó determinar el algoritmo con mejor desempeño entre: Árbol de Decisión y Regresión Logística de Microsoft, sobre datos continuos y discretos de indicadores académicos de una institución de educación superior.

El presente artículo está divido en dos secciones principales: la primera presenta la metodología computacional en la que se describe la variable de comparación, junto con las técnicas y herramientas usadas para la obtención de valores de los indicadores definidos para el desempeño, así como los escenarios preparados para el estudio comparativo; y la

siguiente sección de resultados en donde se analiza los datos obtenidos del tiempo, uso de RAM, uso de CPU y precisión de los algoritmos de minería Árbol de Decisión y Regresión Logística, haciendo uso de estadística descriptiva para definir el conjunto de datos resultante y el test no-paramétrico de estadística inferencial denominado "Prueba de los rangos con signo de Wilcoxon" para evaluar las hipótesis de trabajo, en donde se obtuvo que: el algoritmo Árbol de Decisión tiene más precisión y menor tiempo de respuesta que Regresión Logística, mientras que este último hace un menor uso de CPU en sus operaciones, en cuanto al uso de RAM no se detectaron diferencias estadísticamente significativas entre los datos obtenidos.

Metodología computacional

Con el objetivo de determinar el desempeño de los algoritmos de minería de datos citados anteriormente, se procedió de acuerdo a los siguientes pasos:

a) Definir los parámetros de comparación, esto incluye definir la variable de desempeño y sus indicadores.
b) Diseñar el ambiente de pruebas y escenarios de los cuales se han tomado las mediciones necesarias correspondientes a una muestra de la población definida.
c) Realizar el análisis descriptivo de los resultados obtenidos de cada indicador.
d) Evaluación estadística de resultados mediante contrastes de hipótesis.

Variable de comparación

Los algoritmos de predicción pueden ser analizados por su complejidad espacial o la complejidad temporal al momento de analizar los datos ingresados, pero debido a que se busca los mejores resultados posibles con el algoritmo seleccionado, se analizará el desempeño, ya que, por su definición no sólo se asocia a características de ejecución sino también de efectos del algoritmo sobre las entradas proporcionadas. El desempeño está relacionado con las características de tiempo de ejecución y respuesta, uso de recursos y confiabilidad de las operaciones (IBM Corporation, 2003). Al ser el desempeño una variable de tipo compleja se deben definir los indicadores de la misma, así como los pesos que proporcionarán prioridad a cada uno de los criterios.

Tiempo de respuesta: El tiempo de respuesta corresponde a un indicador de la categoría de velocidad y estará medido en segundos. Este valor ha sido recogido de la herramienta Data Tools. A menor tiempo de respuesta de un algoritmo frente a una misma estructura de datos mejor puntuación tendrá. Del 100% se asigna a este indicador un 10% del peso total.

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

Uso del CPU: El uso del CPU corresponde a la categoría de uso de recursos del computador; ha sido medido en porcentaje desde el monitor del sistema. Un algoritmo tendrá mejor desempeño con respecto al uso de CPU cuanto menor sea su valor. Se asigna a este indicador un 10% del peso total de criterios.

Uso de Memoria: El uso de memoria medido en Megabytes en el monitor del sistema, al igual que el uso del CPU, será mejor en cuanto un algoritmo utilice menos memoria frente a la misma estructura de datos. Se asigna a este criterio un 10% del total.

Precisión: La precisión es el indicador más importante a la hora de decidir qué algoritmo tiene un mejor desempeño por lo que se le asigna un 70% del total de la decisión final. Este valor será obtenido de la herramienta Data Tools que ofrece la opción "Gráfico de precisión" y es la que ayuda a comparar los modelos calculando la efectividad a través de una población normalizada. Una mayor puntuación es mejor (Microsoft Corporation, 2012). La precisión es una magnitud adimensional que demuestra cuán bien un modelo describe el conjunto de datos de entrada.

Ambientes de prueba

El servidor sobre el que se practicaron las pruebas tiene las características de hardware y software que se describen en la Tabla 1.

Tabla 1. Especificaciones del servidor

Característica	**Especificación**
Procesador	Intel® Core™ i7-4702MQ @2.20GHz
Memoria RAM	8GB
Disco Duro	1TB
Sistema operativo	Windows 8.1
Motor de base de datos	Microsoft SQL Server 2012 Standard

Este ambiente de pruebas está orientado a cubrir los objetivos relacionados a los ejes del proceso académico que son: ingreso, matriculación, promoción y graduación. Los requerimientos se listan a continuación:

- Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.
- Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.
- Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia); por carrera, niveles, asignaturas, áreas de conocimiento y facultad.
- Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

- Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.

Para realizar la medición de indicadores se ha tomado como población el total de modelos de minería que satisfacen los requisitos. Debido a que cada requisito exige diferentes niveles de detalle, se tomaron en cuenta 528 modelos, siendo éste el número de la población, y el tamaño de la muestra se define mediante la fórmula (Fernández, 1996):

$$n = \frac{N * p * q * Z^2}{e^2 * (N-1) + p * q * Z^2} = 76$$

Los valores de la fórmula anterior se muestran en la Tabla 2.

Tabla 2. Valores para determinar la muestra

Variable	Definición	Valor
N	Tamaño de la población.	528
p	Variabilidad positiva	0,5
q	Variabilidad negativa	0,5
Z	Valor obtenido mediante niveles de confianza. Es un valor constante.	1,7 => 91%
e	Límite aceptable de error que, generalmente cuando no se tiene su valor, suele utilizarse un valor que varía entre el 1% (0,01) y 9% (0,09).	9%

Cada uno de los algoritmos ha sido aplicado a una estructura de datos que satisface los requisitos definidos. Estas estructuras son un conjunto de datos obtenidas con lenguaje SQL e integradas a la herramienta Data Tools a través de la opción de "Vista de Datos". Los atributos usados para cada estructura de datos se describen en la Tabla 3.

Tabla 3. Atributos de estructuras de datos

Indicadores	Atributos
Ingreso	Cédula [cadena], nombres [cadena], edad de inscripción [entero], ciudad de procedencia [cadena], provincia de procedencia [cadena], país de procedencia [cadena], sexo [cadena], estado civil [cadena], carrera de inscripción [cadena], instituto de procedencia [cadena], titulo [cadena].
Matriculación	Clave [entero], código de estudiante [cadena],nombres [cadena], nacionalidad [cadena], sexo [cadena], edad de inscripción [entero], total de materias escogidas

	[cadena], total de créditos escogidos [cadena].
Deserción	Clave [entero], materia [cadena], horas teóricas [entero], horas prácticas [entero], área de estudio [cadena], número de matrícula [entero], asistencia [entero], nivel de la materia [entero], nombres de estudiante [cadena], sexo de estudiante [cadena], nacionalidad de estudiante [cadena], estado civil del estudiante [cadena], nombres del docente [cadena], sexo del docente [cadena], estado civil del docente [cadena], nacionalidad del docente[cadena], tipo de docente [cadena], tipo de título del docente, forma de deserción [cadena].
Promoción	Clave [entero], materia [cadena], horas teóricas de la materia [entero], horas prácticas de la materia [entero], área de estudio [cadena], número de matrícula [entero], nivel de la materia [entero], nota de promoción [entero], nombres de estudiante [cadena], sexo del estudiante [cadena], nacionalidad del estudiante [cadena], nombres del docente [cadena], nacionalidad del docente [cadena], sexo del docente [cadena], estado civil del docente [cadena], tipo de docente [cadena].
Repitencia	Clave [entero], materia [cadena], horas teóricas de la materia [decimal], horas prácticas de la materia [decimal], área de materia [cadena], número de matrícula [entero], asistencia [entero], nivel de la materia [entero], nombres del estudiante [cadena], sexo del estudiante [cadena], nacionalidad del estudiante [cadena], nombres del docente [cadena], sexo del docente [cadena], estado civil del docente [cadena], nacionalidad del docente [cadena], tipo de docente [cadena], nota final [decimal].
Eficiencia terminal	Clave [entero], nombres del estudiante [cadena], sexo del estudiante [cadena], nacionalidad [cadena], área de proyecto de graduación [cadena], promedio de notas [decimal], nota promedio de grado [decimal], créditos finales [decimal], edad de ingreso [entero], tiempo para terminar pensum [entero], eficiencia terminal [entero].

Los escenarios surgen al momento de agregar un algoritmo a la estructura de datos; es decir, se definen con la implementación del algoritmo de árbol de decisión y regresión logística respectivamente. Los pasos a seguir para realizar este proceso son:

a) Seleccionar la definición del origen de la estructura de datos (relacional).

b) Seleccionar el algoritmo de Regresión Logística para escenario 1 y Árbol de Decisión para escenario 2.

c) Seleccionar el conjunto de datos de origen.

d) Especificar los tipos de datos.

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

e) Definir el conjunto de aprendizaje y de pruebas para el algoritmo.

f) Dar un nombre al modelo y estructura de minería de datos.

Resultados y discusión

Tiempo de Respuesta

El tiempo de respuesta fue medido en segundos con la ayuda de la herramienta Data Tools de Microsoft. La figura 1 presenta un gráfico resumen de los resultados.

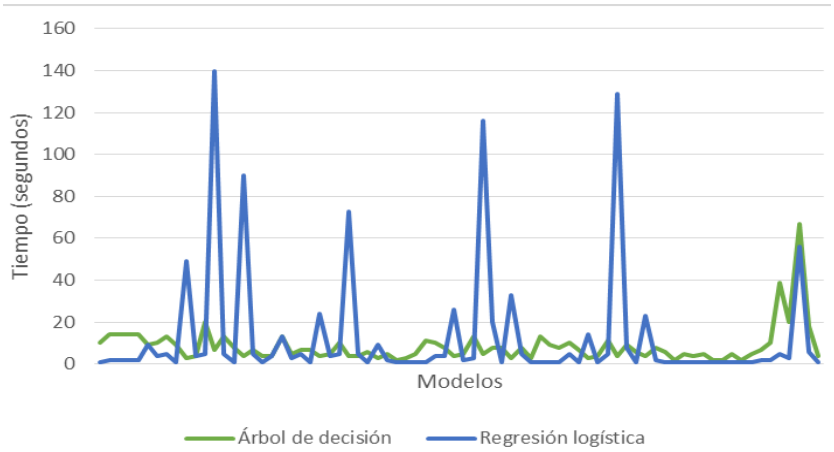

Figura 1. Tiempo de Respuesta

La figura anterior muestra que los valores del tiempo de respuesta para el algoritmo de regresión logística tienen mayor variación que los valores del algoritmo de árbol de decisión. Los datos descriptivos indican que: la media de los 76 valores de muestra para el algoritmo de árbol de decisión es 8,54 segundos, con una desviación estándar de 8,77; frente a 12,80 segundos del algoritmo de regresión logística con una desviación estándar de 28,42.

Uso del CPU

Este indicador fue tomado en porcentaje con la ayuda del monitor de sistema de Microsoft. La figura 2 presenta un gráfico resumen de los resultados obtenidos de este proceso.

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

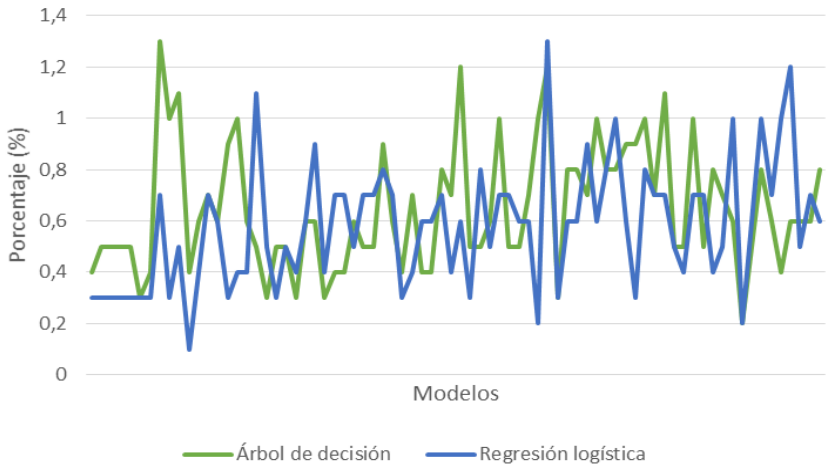

Figura 2. Uso de CPU

Los valores obtenidos muestran gran variabilidad en el uso de CPU en los dos algoritmos de estudio, siendo la media de los 76 valores tomados para el algoritmo de árbol de decisión de 0,65% con una desviación estándar de 0,25; mientras que el algoritmo de regresión logística tiene una media de 0,58% con una desviación estándar de 0,24.

Uso de RAM

Este indicador fue tomado en Megabytes con la ayuda del monitor de sistema de Microsoft. La figura 3 presenta un gráfico resumen de los resultados obtenidos de este proceso.

Figura 3. Uso de RAM

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

Los valores con respecto al uso de RAM mostrados en la gráfica anterior, no muestran una diferencia significativa entre los dos algoritmos de estudio; siendo que la media de los 76 valores tomados para el algoritmo de árbol de decisión es 133,18 MB con una desviación estándar de 22,77; mientras que para el algoritmo de regresión logística la media es de 133,68 MB con 22,76 de desviación estándar.

Precisión

El indicador precisión es un valor que se ha tomado de la herramienta Data Tools de Microsoft. La Figura 4 presenta un gráfico resumen de los resultados obtenidos de este proceso.

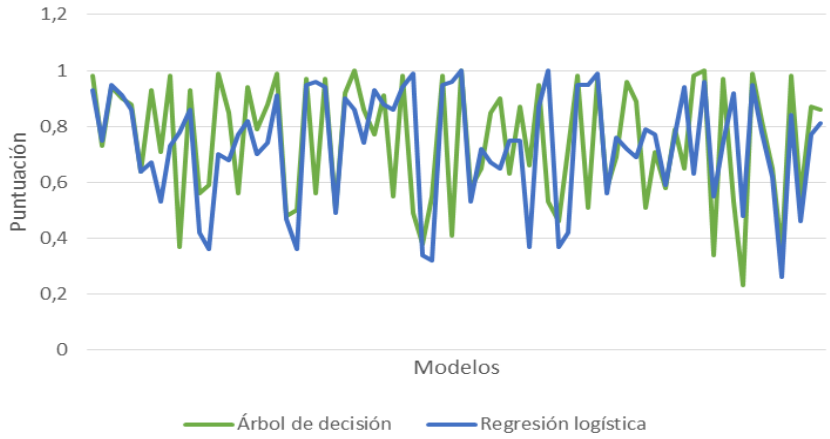

Figura 4. Precisión

Los valores de precisión de los algoritmos, mostrados en la gráfica anterior, muestran variabilidad en los dos casos de estudio, lo que se corrobora con los datos descriptivos: el algoritmo de árbol de decisión tienen una media de 0,75 en la precisión de los 76 modelos obtenidos con una desviación estándar de 0,21; mientras que el algoritmo de regresión logística tiene una media de 0,74 con desviación estándar de 0,20.

Contrastes de Hipótesis

Con el objetivo de determinar el algoritmo de mejor desempeño para ser aplicado sobre los datos académicos de la institución de educación superior, se han planteado cuatro hipótesis con respecto a los indicadores de desempeño definidos que son: tiempo de respuesta, uso de CPU, uso de RAM y precisión. La hipótesis nula en las pruebas de comparación de la media para cada indicador es que el desempeño de los dos algoritmos es el mismo; por tanto se podrá determinar el algoritmo con mayor desempeño en los casos en los que se rechace la hipótesis nula en estas pruebas de comparación de medias.

Revista Cubana de Ciencias Informáticas
Vol. 9, No. 4, Octubre-Diciembre, 2015
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 104-117

Dado que el conjunto de datos de muestra no cumple la hipótesis de normalidad, se ha aplicado el test no-paramétrico denominado "Prueba de los rangos con signo de Wilcoxon" para muestras relacionadas. Y los resultados se indican en la Tabla 4 a continuación:

Tabla 4. Resumen de prueba de hipótesis

Hipótesis nula (H_o)	**Significancia (2 colas)**	**Decisión**
La mediana de las diferencias entre el tiempo de respuesta de Regresión Logística y el tiempo de respuesta de Árbol de Decisión es igual a 0.	0,005	Rechazar la hipótesis nula
La mediana de las diferencias entre la precisión de Regresión Logística y la precisión de Árbol de Decisión es igual a 0.	0,050	Rechazar la hipótesis nula
La mediana de las diferencias entre el uso de CPU de Regresión Logística y el uso de CPU de Árbol de Decisión es igual a 0.	0,046	Rechazar la hipótesis nula
La mediana de las diferencias entre el uso de RAM de Regresión Logística y el uso de RAM de Árbol de Decisión es igual a 0.	0,086	Retener hipótesis nula

A partir de la comparación entre medias muestrales, para aquellas variables en las que se rechaza H_o, se puede determinar que hay suficiente evidencia estadística para afirmar que:

- El algoritmo de Árbol de Decisión con 8,54 segundos supera en tiempo de respuesta a Regresión Logística con 12,80 segundos.
- El algoritmo de Regresión Logística tiene una media de 0,58% con lo que supera al algoritmo de Árbol de Decisión, cuya media es de 0,65% con respecto al uso del CPU.
- El algoritmo de Árbol de Decisión tiene una media de 0,75 de precisión; mientras que el Algoritmo de Regresión Logística tiene una media de 0,74 de precisión, con lo que se define que Árbol de Decisión supera en la prueba de precisión.

Los resultados de tiempo de respuesta y precisión favorecen al algoritmo de Árbol de Decisión, mientras que Regresión Logística lo supera en uso de CPU. El uso de RAM resultó ser semejante para ambos algoritmos. Debido a que la precisión es el indicador de más peso para determinar el algoritmo de mejor desempeño, se escoge el algoritmo de Árbol de Decisión para ser aplicado en el análisis de indicadores académicos de la educación superior.

Conclusiones

El análisis del desempeño de los algoritmos Árbol de Decisión y Regresión Logística, bajo los indicadores de tiempo de respuesta, uso de CPU, uso de RAM y precisión, sobre datos de indicadores académicos, revela que la precisión de

dichos algoritmos es diferente. De esta forma se establece que el algoritmo Árbol de Decisión tiene mejor precisión, debido a que el valor su media muestral es mayor. Cabe resaltar que el indicador precisión es el más importante para establecer el desempeño de un algoritmo de minería de datos.

Los algoritmos no presentan diferencia significativa en el uso de RAM, lo que se puede adjudicar a que los algoritmos fueron sometidos a pruebas bajo las mismas estructuras de datos y el almacenamiento en RAM necesario para el proceso está relacionado con cantidad de datos de entrada proporcionado.

El uso del CPU de los algoritmos Árbol de Decisión y Regresión Logística es diferente y se determina que Regresión Logística tiene un menor uso de este recurso frente a Árbol de Decisión.

En tiempo de respuesta, el algoritmo de Árbol de Decisión tiene una menor media frente al de Regresión Logística, por lo que se puede concluir que lo hace más rápido frente a un mismo escenario, debido a que el segundo algoritmo usa un análisis para cada atributo de predicción mientras que el primero usa los datos de entrada como un solo conjunto para el análisis.

Bajo las circunstancias señaladas en los literales anteriores, se determinó como el algoritmo de mejor desempeño sobre datos académicos al algoritmo Árbol de Decisión.

Contar con los datos socioeconómicos de los estudiantes aportará con nuevos patrones dentro de la extracción de conocimiento, por lo que en futuros trabajos se insta a desarrollar un plan para integrar dicha información a un estudio de algoritmos de minería.

Agradecimientos

A la Escuela Superior Politécnica de Chimborazo por la información facilitada para la realización de este estudio.

Referencias

- CHAPMAN, Pete and et.al. *CRISP-DM 1.0.* Washington D. C., EEUU : SPSS, 2000. págs. 1-76.
- CORTEZ, Paulo; SILVA, Alice. *Using Data Minig to predict secondary school student performance.* Guimaraes, Portugal : s.n., 2006.
- FAYYAD, Usama; PIATESKY, Gregory; PADHRAIC, Smyth. *Knowledge Discovery in Databases.* 1996. págs. 37-54.
- FERNÁNDEZ, Santiago. *Regresión Logística.* Madrid : Universidad Autónoma de Madrid, 2011.

- FERNÁNDEZ, Pita. Investigación: Determinación del tamaño muestral. Unidad de Epidemiología Clínica y Bioestadística. A Coruña. Cad Aten Primaria 1996. págs. 1-6.
- HAN, Jiawei. *Introduction to Data Mining.* San Francisco : Morgan Kaufmann, 2006. págs. 1-20.
- HERNÁNDEZ, José; RAMÍREZ, José; FERRI, César. Introducción a la Minería de Datos Madrid: Pearson, 2004. págs. 3-39.
- HUEBNER, Richard. *A survey of educational data-mining research.* Norwich : Norwich University, 2013. pág. 13.
- IBM Corporation. Rational Unified Process. *Concepts: Performance Testing.* [En línea] 2003. [Citado el: 8 de 11 de 2013.] http://students.mimuw.edu.pl/~zbyszek/posi/ibm/RUP_Eval/process/workflow/test/co_perfo.htm.
- KUMAR, Brijesh y SAURABH, Pal. *Mining Educational Data to analyze student´s performance.* Rajasthan, India : IJACSA, 2011. págs. 63-69.
- MACLENNAN, Jamie. *Data Mining with Microsoft SQL Server 2008.* Indianapolis, EEUU, Wiley Publishing Inc. 2008. págs. 39-53.
- MERCERON, Agathe y KALINA, Yacef. *Educational Data Mining: a Case Study.* Sydney : University of Sydney, 2005. págs. 1-8.
- MICROSOFT CORPORATION. Algoritmo de Árboles de Decisión. [En línea] Microsoft Developer Network, 2012. [Citado el: 6 de 11 de 2013.] https://msdn.microsoft.com/es-ec/library/ms175312.aspx.
- MICROSOFT CORPORATION. Algoritmo Regresión Logística de Microsoft. [En línea] Microsoft Developer Network, 2012. [Citado el: 6 de 11 de 2013.] http://msdn.microsoft.com/es-es/library/ms174806.aspx.
- MICROSOFT CORPORATION. Algoritmos de minería de datos. [En línea] Microsoft Developer Network. [Citado el: 6 de 11 de 2013.] https://msdn.microsoft.com/es-es/library/ms175595.aspx.
- MORENO, María; QUINTALES; Luis y GARCÍA; Francisco. *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software.* Salamanca : Universidad de Salamanca, 2001. págs. 1-14.
- VALLEJOS, Sofía. Minería de Datos. Corrientes, Argentina, Universidad Nacional de Noreste, 2006, págs. 11-16.