Anales Científicos, 78 (1): 26-33 (2017)
ISSN 0255-0407 (Versión impresa)
ISSN 2519-7398 (Versión electrónica)
DOI: http://dx.doi.org/10.21704/ac.v78i1.811

Presentado:29/09/2016
Aceptado: 11/01/2017

Predicción del rendimiento académico aplicando técnicas de minería de datos

Prediction of academic performance applying data mining techniques

[1]Cesar Higinio Menacho Chiok

Resumen

El rendimiento académico de los estudiantes es uno de los temas de mayor preocupación que deben abordar las instituciones educativas superiores. Las Técnicas de minería de datos (TMD) aplicadas a los datos generados en los ambientes educativos, están demostrando ser herramientas eficaces para predecir el rendimiento académico de los estudiantes; con la finalidad de identificar los factores que más influyen en su aprendizaje y apoyar a los profesores a mejorar el proceso de enseñanza a través de realizar acciones pedagógicas más eficientes y oportunas. Esta investigación tiene como objetivo aplicar las TMD de regresión logística, árboles de decisión, redes bayesianas y redes neuronales usando los datos académicos de los estudiantes matriculados en el curso de Estadística General de la UNALM en los semestres 2013 II y 2014 I, con la finalidad de predecir la clasificación final (Desaprobado o Aprobado) de los futuros estudiantes matriculados en el curso. Se usa la matriz de confusión para comparar y evaluar la precisión de los clasificadores. Los resultados indican que la red Naive de Bayes obtuvo la mayor tasa de buena clasificación (71,0%).

Palabras claves: Rendimiento académico, Modelos predictivos, Técnicas de minería de datos, Matriz de confusión.

Abstract

Student's academic performance is one of the issues of greatest concern to higher education institutions. Data mining techniques (TMD) applied to data generated in educational environments are proving to be effective tools for predicting students' academic performance; With the aim of identifying the factors that most influence their learning and supporting teachers to improve the teaching process through more efficient and timely pedagogical actions. This research aims to apply logistic regression TMDs, decision trees, Bayesian networks and neural networks using the academic data of the students enrolled in the General Statistics course of the UNALM in the semesters 2013 II and 2014 I, in order to predict the final classification (Disapproved or Approved) of future students enrolled in the course. The confusion matrix is used to compare and evaluate the accuracy of the classifiers. The results indicate that the Naive network of Bayes obtained the highest rate of good classification (71.0%).

Key words: Academic Performance, Predictive Models, Data Mining Techniques, Confusion matrix.

1. **Introducción**

Uno de los retos que tienen que enfrentar las instituciones de educación superior para ofrecer una mayor calidad educativa, es mejorar el rendimiento académico de los estudiantes (Kumar & Chadha, 2011). La calidad de la educación puede ser medido por medio del rendimiento académico de los estudiantes (Rubyl & David, 2015). El bajo rendimiento académico de los estudiantes que lleva a la desaprobación de los cursos, es uno de los problemas que debe enfrentar las instituciones educativas superiores. Una manera de enfrentar este problema, es contar con información oportuna sobre el desempeño académico de los alumnos matriculados en los cursos al inicio de un semestre. Las técnicas de minería de datos (TMD), están siendo ampliamente aplicadas para predecir el rendimiento académico de los estudiantes, con la finalidad de detectar los factores que más influyen en su proceso de aprendizaje y permitan a los profesores realizar acciones pedagógicas más efectivas de asesoramiento y acompañamiento sobre los estudiantes que las requieran.

En las últimas décadas, hay un gran interés por aplicar las TMD en los ambientes de la educación superior, lo que ha generado una nueva comunidad de investigación educativa denominada Minería de Datos Educativa (en inglés: Educational Data Mining). La MDE, aplica las TMD para analizar y explorar los datos de los entornos educativos con la finalidad de comprender mejor el desempeño de los estudiantes y las condiciones en las cuales ellos aprenden (Ramaswami, 2009a), para lograr una mayor calidad de la enseñanza (Baker, 2008), para mejorar la calidad de la enseñanza en una institución de educación superior (Goyal & Vohra, 2012) y para transformar los datos en información útil para apoyar la toma decisiones y responder preguntas en la investigación educativa (Heiner, Baker, & Yacef, 2006). Según (Kabakchieva, 2013), hay un creciente uso de las técnicas de minería de datos en las universidades con la finalidad de analizar los datos educativos, con la finalidad de extraer información y conocimiento para apoyar la toma de decisiones educativas.

[1]Facultad de Economía y Planificación, Universidad Nacional Agraria La Molina, Lima, Perú. E-mail: cmenacho@lamolina.edu.pe

El Dpto. de Estadística Informática (DEI) de la Facultad de Economía y Planificación de la UNALM, en los últimos años está realizando esfuerzos para mejorar el proceso de enseñanza y aprendizaje del curso de Estadística General que ofrece a todos los estudiantes de las diferentes especialidades; tales como, la implementación de materiales y autoevaluaciones en la plataforma Moodle, monitoreo de clases, toma de asistencia y otros más. Cada semestre el DEI ofrece 9 grupos que hace un total de 450 alumnos matriculados. Los resultados semestrales indican que en promedio el 40% de los alumnos desaprueban el curso. A los esfuerzos que realiza el DEI, será de gran ayuda contar con modelos predictivos que se conviertan en herramientas analíticas para apoyar a los docentes a identificar al inicio del semestre posibles estudiantes con problemas de rendimiento, y así poder brindarles un seguimiento y asesoramiento adecuado.

La presente investigación tiene como objetivo aplicar las TMD de regresión logística, árboles de decisión, redes bayesianas y redes neuronales usando los datos académicos de los estudiantes matriculados en el curso de Estadística General de la UNALM de los semestres 2013 II y 2014 I, con la finalidad de predecir la clasificación final (Desaprobado o Aprobado) de los estudiantes matriculados en el curso. Se aplicarán métricas a partir de la matriz de confusión para validar e identificar el mejor modelo que permita predecir la clasificación de nuevos estudiantes matriculados en el curso; y apoyar a los profesores a identificar a los estudiantes con problemas académicos para brindarles un asesoramiento oportuno y efectivo.

Las TMD han demostrado ser herramientas eficaces para predecir el rendimiento académico de los estudiantes. En (Pimpa, 2013), se ha realizado un estudio para analizar los factores que afectan el rendimiento académico. Se aplican los algoritmos de árboles de decisión y redes bayesianas con el programa Weka y con validación cruzada 10 folds. Los datos corresponden a 1600 estudiantes de los registros académicos del 2001 al 2011 de la universidad de Tailandia. El árbol de decisión tuvo una precisión del 85,2%. En (Gart & Sharma, 2013) se comparan varias TMD (C4.5, ID3, CART y J48, Naive de Bayes, Redes Neuronales, k-medias y k-vecino más cercano) para predecir el rendimiento académico que apoyen a las instituciones de educación superior a mejorar las habilidades y el desempeño de los estudiantes. En (Dole & Rajurkar, 2014), se aplica el algoritmo Naive de Bayes y árbol de decisión para predecir la graduación y la nota final de los estudiantes (Aprobado y desaprobado) sobre una colección de datos (2010-2011) entre el primer año y los datos tomados durante la preparatoria de 257 estudiantes y considerando 11 atributos. En Cocea & Weibelzahl (2007) para predecir el éxito en un curso de informática, se comparan cinco algoritmos de clasificación regresión lineal y SVM para datos continuos y Naïve Bayes, Bayes Net TAN y redes bayesianas para datos discretos con 125 y 88 observaciones respectivamente. En (El Din Ahmed & Sayed, 2014), se aplica el árbol de decisión ID3 para evaluar el rendimiento de los estudiantes, sobre 10 atributos de la base de datos de los estudiantes (1547 registros, período: 2005-2010), con la finalidad de predecir el rendimiento final. El estudio ayudará a mejorar el rendimiento académico de los estudiantes, identificando aquellos que necesitan atención especial y tomando acciones oportunas; consiguiendo reducir la tasa de desaprobación.

En (Thai, 2007), se compara la precisión de los algoritmos de árboles de decisión y redes Bayesianas, para predecir el rendimiento académico de los estudiantes no graduados y graduados sobre su nota final. Se aplica el programa con Weka, con los algoritmos árboles de decisión J48 y M5P y red Bayesiana; considerando diferentes valores de los parámetros. Los resultados indican una precisión del 73,0% cuando la clase es (bajo, muy bajo, bueno y muy bueno), 94,0% cuando la clase es (Desaprobado, Aprobado). En (Kabra & Bichkar, 2011), se aplica el algoritmo de árbol C4.5 con información pasada del rendimiento académico, con la finalidad de construir un modelo predictivo capaz de identificar anticipadamente a los estudiantes del primer ciclo de ingeniería con probable desaprobación, y apoyar al profesor a proveer un adecuado asesoramiento. Se aplica el programa Weka a los datos de admisión, académicos y demográficos de 346 estudiantes.

2. Materiales y métodos

2.1 Materiales

Los datos corresponden a los registros académicos de la Oficina de Estudios de la UNALM, para una muestra de 914 estudiantes matriculados en los ciclos 2013 II y 2014 I en el curso de Estadística General. Las variables consideradas para el estudio se describen en la Tabla 1.

Tabla 1. Descripción de variables

Variables	Descripción	Valores
X1	Situación del curso de EG	1=Nuevo, 0=Repitente
X2	Sexo	1=Masculino, 0=Femenino
X3	Promedio ponderado	1=Menor o igual a 11,5; 0=Mayor a 11,5
X4	Situación académica	1=Normal, 0=Observado
X5	Nr. veces que llevo Matemáticas	1=No_Rep_Mat, 0=Rep_Mat
X6	Nota en matemáticas	1=11 y 12, 0=Más de 12
X7	Nr. veces que llevo Diferencial	1=No_Rep_Mat, 0=Rep_Mat
X8	Nota en diferencial	1=11 y 12, 0=Más de 12
X9	Situación en Integral	1=Llevo el curso, 0=No llevo el curso
X10	Créditos cursados	1=Menor o igual a 15, 0=Mayor a 15
Y	Resultado final	1=Desaprobado, 0=Aprobado

Terminología para la definición de variables

Una base de datos (D), con p atributos y n instancias, puede ser expresada como una matriz de datos X_{ij}; i=1,2,...,n y j=1,2,..., p, que representa a la i-ésima observación (instancia) y j-ésima variable (atributo). Sea X_j^k, $K=1, 2,..., q$ la j-ésima variable cualitativa independiente con su k-ésimo valor y con "q" posibles valores. La variable dependiente para un problema de clasificación, se define como cualitativa Y^g ;g=1,2,.., m, que corresponde la g-ésima clase. La variable dependiente Y^g, se denomina variable o atributo clase, conteniendo m clase o categorías. En la aplicación del estudio, la variable Y^g representa el resultado final de un estudiante con dos clases: Y^1 =*Desaprobado,* Y^0 =*Aprobado.*

2.2 Métodos

Las técnicas de minería de datos aplicadas a la tarea de la predicción, tienen como objetivo desarrollar un modelo que permita predecir el valor de la variable de entrada (variable dependiente) en función de un conjunto de variables predictoras (variables independientes). Cuando se aplican las TMD a un problema de clasificación, la variable dependiente es cualitativa y corresponde a un aprendizaje supervisado, puesto que los datos se encuentran previamente clasificados. En el dominio educativo, un modelo predictivo del rendimiento académico tiene como finalidad estimar el valor de la variable dependiente referida a una nota o calificación que describe el resultado final del estudiante en un curso determinado. En esta investigación, las técnicas de minería de datos para la clasificación que se proponen aplicar con la finalidad de predecir el rendimiento final de los estudiantes en el curso de EG (Desaprobado y Aprobado) son: regresión logística, árboles de decisión, redes neuronales y redes bayesianas. A continuación se presentará una descripción general sobre el proceso de aprendizaje y la inferencia para la predicción de la clasificación de nuevas instancias en el ámbito educativo para cada una de las técnicas propuestas.

Regresión logística binaria (RL)

La Regresión Logística (RL), permite estudiar la dependencia funcional entre una variable dependiente categórica Y (con dos clases) y un conjunto de "p" variables independientes o predictoras X= (X_{1i}, X_{2i},...,X_{pi}) que pueden ser cuantitativas o categóricas. El modelo de una regresión logística binaria, permite predecir en términos de la probabilidad la ocurrencia del evento de interés (Y=1, Desaprobad). Así se tiene la ecuación:

$$P\,(Y=1/X)=\frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X1_i+\beta_2 X_{2i}+...+\beta_p X_{pi}}}$$

El proceso de inferencia con la RL, consiste aplicar la ecuación estimada al vector de datos X^K_0 para predecir la clasificación del rendimiento académico de un estudiante (Desaprobado o Aprobado). La ecuación estimada será:

$$P\,(Y=1/X^K_0)=\frac{e^{\hat\beta_0+\hat\beta_1 X_{10}+\hat\beta_2 X_{20}+...+\hat\beta_p X_{P0}}}{1+e^{\hat\beta_0+\hat\beta_1 X_{10}+\hat\beta_2 X_{20}+...+\hat\beta_p X_{P0}}}$$

Árboles de decisión (AD)

Un árbol de decisión es un modelo jerárquico para el aprendizaje supervisado, que puede ser aplicado para un problema de regresión o clasificación. Un árbol de decisión es un modelo no paramétrico, puesto que no se asume ninguna forma paramétrica para las densidades de la variable clase y la estructura de árbol no se fija a priori, sino que se va generando durante el proceso de aprendizaje y que depende de la complejidad del problema inherente a los datos (Alpaydin, 2010). El modelo para un árbol de clasificación, se presenta como una estructura jerárquica para mostrar y establecer las relaciones entre la variable de dependiente y el conjunto de variables predictoras. El AD está compuesto por el nodo raíz que se presenta en la parte superior, un conjunto de nodos internos asociados cada uno a una variable predictoras, y cuyas ramas representan validaciones o decisiones de los valores de la variable y un conjunto de nodos hojas o nodos terminales, que están etiquetadas con algún valor de la clase de la variable dependiente. Uno de los AD de clasificación de mayor aplicación, es el algoritmo C4.5 que fue propuesto por (Quinlan, 1993). El C4.5, va generando el árbol con la estrategia "Divide y Vencerás", que consiste en un proceso recursivo y descendentemente para ir asignado los atributos a los nodos, dividiendo en tantas ramas como valores tenga la variable. El C4.5 utiliza la métrica de la Razón de Ganancia para la selección del atributo y realiza la poda del árbol después de haberlo construido (post poda) posibilitando tener árboles más consistentes y evitando el problema de sobre ajuste. Puede trabajar con atributos nominales y continuos; pero los atributos continuos son convertidos en intervalos discretos (discretización automática) y puede construir árboles cuando algunos de los ejemplos presentan valores perdidos.

La inducción de un árbol, consiste en el proceso de su construcción a partir de un conjunto de entrenamiento. Considerando un conjunto de entrenamiento D, con N instancias (observaciones) agrupadas en m clases para la variable dependiente Y^g; g=1, 2, ... m. Se define N^g como el número de instancias que corresponde a la clase g, N_h que contiene el nodo h, N en el nodo raíz y $N^g{}_h$ para el nodo h y clase g, tal que $\sum N^g{}_h = N_h$ Se define la probabilidad para clase g (a priori) p^g =(N^g/N)y para el nodo h y la clase g, por $p^g{}_h$ =($N^g{}_h$/N_h). El algoritmo C4.5, utiliza la medida de Entropía (H) para el cálculo de la Razón de ganancia y elegir el atributo que se asignará a un nodo para dividir las instancias en tantas ramas como valores tenga. Cuando menor sea el valor de la Entropía para un atributo, menor será la incertidumbre y mayor la información que proporciona el atributo. La Razón de la Ganancia de información (RG) para una variable, se

define como el cociente entre la ganancia de información y su entropía, escogiendo la variable con la mayor ganancia de información. La RG se define por:

$$RG\ (X^k_j) = [GI(X^k_j)/H(X^k_j)]$$

Donde:

$GR\ (X^k_j) = H(Y^g)\text{-}H_p(X^k_j)$, es la Ganancia de Información. Mide la ganancia que se obtiene por usar la variable. Se escoge la variable con el valor más alto, indicando que la reduce la incertidumbre de la clasificación de la clase objetivo.

$H\ (Y^g)\sum_g p^g Log_2 p^g$, es la entropía para la variable clase.

$H\ (Y^k_j)\sum_k p^k_k Log_2 p^k_k$, entropía para una variable con respecto a una clase.

$Hp\ (Y^k_j)\sum_k p^k_k Log_2 p^g\ (X^k_j)$, es la entropía media ponderada.

El proceso de inferencia con el C4.5, con la finalidad de predecir la clasificación del resultado final del rendimiento académico para un estudiante como Desaprobado o Aprobado; se realiza a través de evaluar el conjunto de reglas que genera el C4.5, haciendo los recorridos desde el nodo raíz hasta cada uno de los nodos terminales.

Redes bayesianas (RB)

Las Redes Bayesianas o redes probabilísticas se fundamentan en la teoría de la probabilidad y combinan la potencia del teorema de Bayes con la expresividad semántica de un grafo. Las RB permiten representar un modelo causal por medio de una representación gráfica de las independencias y dependencias entre las variables que forman parte del dominio de la aplicación. Las RB, tienen una parte cualitativa representada por el grafo, compuesto por nodos que representan las variables y arcos que unen los nodos que representan valores de dependencia entre las variables. La parte cuantitativa comprende la incertidumbre del problema, representada por medio de probabilidades condicionales que reflejan posibles relaciones causa – efecto entre los nodos (variables) de la RB. La RB más simple para problemas de predicción y clasificación, es conocida como Näive Bayes (NB) que tiene una estructura de red fija de un grafo acíclico dirigido. La estructura de una Red Bayesiana Naive Bayes, está compuesta por un nodo raíz que corresponde a la variable dependiente Y^g (variable clase) y un conjunto de variables independientes $X=(X_1,X_2,...,X_p)$ que están representadas en los nodo de la red. Un clasificador Näive de Bayes, asume que Y^g es padre de todas las variables del conjunto X y a su vez estas variables son independientes entre sí dada la variable clase.

El proceso de la construcción de una red bayesiana Naive de Bayes, se basa realizar el aprendizaje paramétrico que consiste en estimar las probabilidades con el conjunto de entrenamiento D. La probabilidad a priori de la variable clase: $P(Y^g)=(N^g/N)$; para cada variable predictora X^k_j, la probabilidad condicional dada la variable clase: $P(X^k_j/Y^k)=(N^k_h/N_h)$; la probabilidad total: $P(X^k_j)=\sum_{k=1}^{m} P(Y^g)\ P(X^K_j/Y^g)$, que indican que tan probable para la variable predictora X^k_j ocurra su valor k-ésimo.

El proceso de inferencia, con la finalidad de predecir la clasificación del resultado final del rendimiento académico para un estudiante X^k_0 como Desaprobado o Aprobado, se basa en estimar la probabilidad a posteriori (Teorema de Bayes), que permita determinar a la clase (C_g) más probable a la que pueda pertenecer. La probabilidad a posteriori se expresa:

$$P(C_k / X^k_0) = \prod_{j=1}^{p} P(Y^g)P(X^k_j / C_g); \quad g = 1,2,...,m$$

Redes neuronales (RN)

Las Redes Neuronales Artificiales (RNA), son modelos matemáticos y computacionales inspirados en sistemas biológicos, adaptados y simulados en computadoras. El modelo de una RNA, se basa en la estructura de un grafo dirigido, compuesto por un conjunto de neuronas que se interconectan a través de arcos dirigidos que establece la relación entre nodos $j \rightarrow i$, y tiene asociados un peso numérico W_{ij} que determina la fuerza y el signo de la conexión. Las neuronas se organizan por niveles o capas que pueden ser de tres clases: entrada, oculta o salida, con un número determinado de neuronas en cada una de ellas. Las RNA, tienen definidas dos funciones: una función de activación y una función de salida; siendo las funciones más usadas: escalón, identidad, logística, gausiana y tangente hiperbólica.

El modelo de RNA más conocido es el Perceptrón Multicapa (PM), que comprende de una capa de entrada, una oculta y una de salida. El PM, es una red que está compuesto por un conjunto de nodos de entradas que se asocia a cada una de las variables predictoras $X=(X_1, X_2,..., X_p)$, un conjunto de nodos intermedios que representan las interconexiones y los nodo de salida asociados a los valores de variable dependiente. El aprendizaje para un conjunto de entrenamiento (D) con el PM, se usa el algoritmo Back Propagación que se basa en una usar una función de entrada en el nodo h:

$Net_h = w_0 + \sum_{i-1}^{p} w_{ih} x_j$ y funciones de transferencia o salida que pueden ser, la identidad: $Y^g_h = {}_{Neth}$ o la sigmoidal: $Y^g_h = [1/(1+e^{-Neth})]$.

El proceso de inferencia con la RN, con la finalidad de predecir la clasificación del resultado final del rendimiento académico para un estudiante X^k_0 como Desaprobado o Aprobado, se basa en la propagación de la red. La expresión será: $\hat{Y}^g_0 = [1/(1+e^{-Neth})]$, donde

$Net_0 = w_0 + \sum_{j-1}^{p} w_{j0} x_{j0}$

Técnicas para evaluar clasificadores

La evaluación de las técnicas de clasificación, es importante porque permite validar la bondad de ajuste del modelo sobre el conjunto de entrenamiento. Así mismo, permiten comparar entre varias técnicas de clasificación y seleccionar la que tenga la mayor precisión. Para la evaluación de las TMD, se propone usar la matriz de confusión, área bajo la curva ROC y el coeficiente Kappa.

Matriz de confusión. La matriz de confusión es una tabla de contingencia que muestra la distribución de la clasificación observada (real) y la predecida (clasificador) para las distintas categorías de la variable clase. En la Tabla 2 se muestra la matriz de confusión para el caso de dos clases.

Tabla 2. Matriz de confusión

Clasificación observada	Clasificación predecida		Total (Observado)
	Positiva (Clase 0)	Negativa (Clase 1)	
Positiva (Clase 0)	VP	FN	VP + FN
Negativa (Clase 1)	FP	VN	FP + VN
Total (Predecido)	VP + FP	FN + VN	N

Donde: $N=VP+VN+FP+FN$

El VP (verdaderos positivos) y El VN (verdaderos negativos), es el número de observaciones que predice correctamente el clasificador como la clase positiva y negativa. El FP (falsos positivos) y El FN (falsos negativos), es el número de observaciones que se predice incorrectamente como la clase positiva siendo de la clase negativa y como la clase negativa siendo de la clase positiva respectivamente. A partir de la matriz de confusión, se determinados las métricas:

S= [$(VP+VN)/N$] La tasa de buena clasificación. Mide la proporción de observaciones que el clasificador predice correctamente la clase positiva y negativa

e= [$(FN+FP)$/N] La tasa de mala clasificación. Es la proporción de observaciones que el clasificador predice incorrectamente

El área bajo la curva ROC. Se puede usar como un índice conveniente de la exactitud global de la prueba, llamada AUC (Área bajo la curva ROC) como un índice de la performance del clasificador (la exactitud máxima correspondería a un valor del área bajo la curva de 1 y la mínima a un valor de 0,5). La AUC se calcula con la siguiente expresión:

$$AUC= [(1+TVP-TFP)/2], \qquad 0 \leq AUC \leq 1$$

Coeficiente de Kappa (k). Es un coeficiente estadístico propuesto originalmente por (Cohen, 1960) que permite medir la concordancia entre los resultados de dos o más variables cualitativas. El índice k, aplicado a la tabla de confusión permite evaluar si la clasificación observada es similar (concordante) con la clasificación predecida por el clasificador. Para dos categorías, el coeficiente de Kappa se calcula:

$$k=[(P_0-P_e)/(1-P_e)],$$

$$0 \leq k \leq 1 \quad \text{con } k= P_0[(VP+VN)/N] \text{ y } P_e[(a*c+b*d)/N^2]$$

Siendo: $a=VP+FP$, $b=FN+VN$, $c= VP+FN$, $d=FP+VN$

Donde: Po, es la proporción de aciertos. Pe, es la proporción de aciertos esperados bajo la hipótesis de independencia entre las dos variables. En la Tabla 3, se presenta la valoración del valor de k que utiliza la escala propuesta por (Landis and Koch, 1977)

Tabla 3. Valoración del coeficiente Kappa

kappa	Grado de concordancia
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

3. Resultados y discusión

Para la aplicación de las TMD propuestas, se usará el programa WEKA (Waikato Environment for Knowledge Analysis) desarrollado por la Universidad de Waikato de Nueva Zelanda. WEKA es un programa de uso libre (Licencia GNU) y está compuesto por un conjunto de algoritmos que implementan la mayoría de las técnicas de minería de datos. Para evaluar los modelos predictivos sobe el resultado (Desaprobado, Aprobado) de los estudiantes matriculados en el curso de Estadística General, se usará el método Validación Cruzada-10 folds y la matriz de confusión.

En primer lugar se aplica la regresión logística para estimar el modelo predictivo que mejor se ajuste a los datos, seleccionado las variables predictoras que sean estadísticamente significativas. En la Tabla 4, se muestra las variables seleccionadas significativas.

Tabla 4. Significación de las variables

Variables	Coeficiente	P-Valor
X1_Sit_Curso	-0,368	0,039
X2_Sexo	-0,317	0,038
X3_Prom_Acum	0,559	0,004
X4_Sit_Academica	-0,328	0,082
X5_Veces_Mat	-0,427	0,031
X6_Nota_Mat	0,08	0,070
X7_Veces_Dif	-0,462	0,010
X8_Nota_Dif	0,183	0,000
X9_Llevo_Integral	-0,616	0,000
X10_Cred_Cursados	-0,018	0455
Constante	-1,896	0,028

Las variables significativas para explicar el resultado final de un estudiante en el curso EG son: X1, X2, X3, X5, X7, X8, y X9.

Regresión logística

En la Tabla 5 se presenta el análisis de regresión logístico. Según el valor de la razón de ventaja (Odds Ratios), se identifican las variables más importantes para predecir el resultado de un estudiante en el curso de EG, el promedio ponderado y la nota en cálculo diferencias.

Tabla 5. Significación de los coeficientes de regresión

Variables	Coeficiente	P-Valor	Exp(Beta)
X1_Sit_Curso	-0,510	0,002	0,600
X2_Sexo	-0,333	0,028	0,717
X3_Prom_Acum	0,854	0,065	2,350
X5_Veces_Mat	-0,485	0,008	0,616
X7_Veces_Dif	-0,696	0,000	0,499
X8_Nota_Dif	0,238	0,000	1,269
X9_Llevo_Integral	-0,645	0,000	0,525
Constante	-2,236	0,003	0,107

Árboles de decisión

Se aplica el método C4.5 que se implementa en WEKA con el algoritmo J48. Se usa un nivel de confianza de 0,15 y con poda.

```
J48 pruned tree
------------------
X3_Prom_Acum = Menor_o_igual_a_11.5
|   X1_Sit_Curso = Repitente
|   |   X8_Nota_Dif <= 12: Desaprobo (84.0/36.0)
|   |   X8_Nota_Dif > 12: Aprobo (34.0/9.0)
|   X1_Sit_Curso = Nuevo: Desaprobo (108.0/26.0)
X3_Prom_Acum = Mayor_a_11.5
|   X7_Veces_Dif = Má_de_una_vez_Dif
|   |   X8_Nota_Dif <= 11: Desaprobo (44.0/13.0)
|   |   X8_Nota_Dif > 11
|   |   |   X1_Sit_Curso = Repitente
|   |   |   |   X2_Sexo = Femenino: Aprobo (26.0/11.0)
|   |   |   |   X2_Sexo = Masculino
|   |   |   |   |   X8_Nota_Dif <= 16: Desaprobo (26.0/10.0)
|   |   |   |   |   X8_Nota_Dif > 16: Aprobo (2.0)
|   |   |   X1_Sit_Curso = Nuevo: Aprobo (49.0/17.0)
|   X7_Veces_Dif = Una_vez_Dif: Aprobo (541.0/121.0)
Number of Leaves  :         9
Size of the tree :  17
```

Figura 1. Reglas obtenidas con árbol de decisión J48

En la Figura 1 presenta el conjunto de reglas generadas con el árbol J48, con un tamaño de 17 nodos y 9 hojas. Se observa que las variables más importantes para predecir el resultado de un estudiante en el curso de EG son en orden de jerárquico el promedio acumulado, situación del curso, número de veces que lleva el curso de diferencial, nota en el curso diferencial y sexo. Las variables que no influyen son número de veces que llevo matemáticas y situación en el curso de integral. Las tres reglas para que un estudiante Desaprueba el curso de EG son:

- Si X3_Prom_Acum <= 11,5 & X1_Sit_Curso = Repitente & X8_Nota_Dif <= 12
- Si X3_Prom_Acum <= 11,5 & X1_Sit_Curso = Nuevo
- Si X3_Prom_Acum >= 11,5 & X7_Veces_Dif >una_ vez_Dif & X8_Nota_Dif <= 11

Redes neuronales

Se aplica el perceptron multicapa con función de activación (entrada) y transferencia (salida) la logística. El aprendizaje con la red neuronal resultó con una capa oculta con cuatro nodos y los pesos se presentan en la Tabla 6.

Tabla 6. Pesos de red neuronal Perceptrón Multicapa

Variable	Intercepto	Nodo 2	Nodo 3	Nodo 4	Nodo 5
Nodo 0	2,09	-2,06	-1,75	-1,82	-1,16
Nodo 1	-2,09	2,06	1,75	1,82	1,16
Intercepto		-5,70	-8,05	9,39	-10,1
X1_Sit_Curso=Nuevo		-11,11	4,18	6,09	3,53
X2_Sexo=Masculino		2,14	-1,85	-1,62	-3,59
X3_Prom_Acum=Mayor a 11.5		-9,77	8,81	8,69	11,21
X5_Veces_Mat=Una vez Mat		-2,64	-5,82	9,32	2,84
X7_Veces_Dif=Una vez Dif		-1,85	4,67	-0,65	3,84
X8_Nota_Dif		9,52	1,35	13,53	5,27
X9_Sit_Integral=Llevo Integral		-2,59	-0,55	5,29	6,95

Redes bayesianas

Se considera el algoritmo para la red bayesiana Naive de Bayes con el algoritmo K2. En la Tabla 7 se muestra las probabilidades a priori de cada clase, las condicionales de cada variable dada su clase y total de cada valor de la variable. Las variables más importantes para predecir el resultado de los estudiantes son llevó cálculo integral, nota en diferencial, situación del curso, sexo y número de veces que llevó diferencial.

Tabla 7. Probabilidades a priori, condicionales y total para RB Naive Bayes

		Variable clase		Probabilidad
		Desaprobado	Aprobado	Total
Resultado del curso de EG		36,7	63,3	
Situación curso	Repitente	47,8	30,8	37,0
	Nuevo	52,2	69,2	63,0
Sexo	Masculino	57,3	48,9	52,0
	Femenino	42,7	51,1	48,0
Promedio ponderado	Menor = a 11.5	41,5	15,1	24,8
	Mayor 115.	58,5	84,9	75,2
Matemática	Más 1 vez	36,2	14,9	22,7
	Una vez	63,8	85,1	77,3
Cálculo diferencial	Más 1 vez	45,4	22,0	30,6
	Una vez	54,6	78,0	69,4
Nota diferencial	Menor = a 12.5	64,1	39,7	48,7
	Mayor a 12.5	35,9	60,3	51,3
Cálculo integral	No llevo	65,9	49,4	55,5
	Si llevo	34,1	50,6	44,5

Técnicas para evaluar clasificadores

En la Tabla 8 se presenta las métricas calculadas a partir de la matriz de confusión con la finalidad de comparar y evaluar las TMD propuestas.

Tabla 8. Métricas para evaluar a los clasificadores

TMD	Tasa de buena clasificación	Área bajo la curva ROC	Coeficiente Kappa
Regresión Logística	68,4	0,59	0,28
Árboles de Decisión J48	68,3	0,58	0,29
Redes Neuronales	67,9	0,57	0,29
Red Naive de Bayes	71,0	0,62	0,36

Según la Tabla 8, la red Naive de Bayes muestra ligeramente una mayor precisión, indicando que el 71,0% de las instancias las está clasificando correctamente. Los resultados para los valores del área bajo la curva ROC muestran un nivel de satisfacción aceptable para las cuatro técnicas al superar el valor del 0,5. Respecto a la concordancia de la matriz de confusión, las cuatro técnicas presentan un grado discreto al evaluar el coeficiente Kappa.

Inferencia de las TMD

Con la finalidad de realizar la predicción de la clasificación (Aprobado o Desaprobado), se ha considerado una muestra de 10 nuevos estudiantes que se matricularon en el curso de EG. Los resultados para las cuatro TMD con sus respectivas probabilidades se presenta en la Tabla 9.

Tabla 9. Predicción de la clasificación para una muestra de 10 nuevos estudiantes

Estudiante	RL	Prob.	AD	Prob.	RN	Prob.	RB	Prob.
E1	D	0,76	A	0,75	D	0,51	D	0,96
E2	D	0,79	A	0,74	D	0,51	D	0,92
E3	D	0,66	D	0,71	D	0,65	D	0,71
E4	D	0,55	D	0,71	D	0,52	D	0,73
E5	A	0,79	A	0,65	A	0,71	A	0,9
E6	A	0,5	A	0,65	A	0,75	D	0,73
E7	D	0,59	D	0,71	D	0,89	D	0,8
E8	A	0,63	A	0,58	A	0,7	A	0,54
E9	A	0,51	D	0,62	A	0,55	D	0,74
E10	D	0,69	D	0,62	D	0,87	D	0,89
Precisión	70		60		70		70	

Considerando la predicción del algoritmo de red Naive de Bayes, existen ocho estudiantes que posiblemente podrían tener problemas académicos con un desaprobación en el curso de EG, y por lo tanto necesitarán un seguimiento y asesoramientos en el curso.

4. Conclusiones

Las TMD demuestran ser herramientas eficaces para obtener modelos que permitan predecir el resultado de los estudiantes matriculados en el curso de Estadística General. La técnica de la red Naive de Bayes resultó con una la mayor precisión, al obtener un 71,0% de correcta clasificación. Así mismo, se aprecia que en las cuatro técnicas respecto a la precisión de cada clase, resultó con mayores porcentajes de correcta clasificación para la clase Aprobado y menores para la clase Desaprobado. Las variables que influyen en el resultado del curso de EG, fueron el promedio ponderado, situación del curso, nota en diferencial y número de veces que llevó diferencial. Se recomienda aplicar las TMD con información socio económica, de los estudiantes a fin de mejorar el modelo predictivo.

5. Literatura citada

Alpaydin. (2010). Introduction to Machine Learning. Second Edition. Massachusetts Intitute of Tecnology.

Baker, J. R. (2008). Educational Data Mining 2008. The 1st International Conference on Educational Data Mining. Montreal.

Cocea, M., & Weibelzahl, S. (2007). Cross-System Validation of Engagement Prediction from Log Files. pp. 14-25.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational Psychol, pp. 37-46.

Dole, L., & Rajurkar, J. (2014). A Decision Support System for Predicting Student Performance. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2.

El Din Ahmed, A., & Sayed, I. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*. Vol 2(2): 43-47.

Gart, S., & Sharma, A. (2013). Comparative Analysis of Data Mining Techniques on Educational Dataset. *International Journal of Computer Applications*. Vol. 74 (5):1-5.

Goyal, M., & Vohra, R. (2012). Applications of Data Mining in Higher Education.

Heiner, C., Baker, R., & Yacef, K. (2006). Proceedings of Educational Data Mining workshop. 8th International Conference on Intelligent Tutoring Systems.

Johnson, L. S. (2011). The 2011 Horizon Report. Austin, Texas: The New Media Consortium.

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. pp. 61-72.

Kabra, R., & Bichkar, R.S. (2011). Performance Prediction of Engineering Students using Decision Trees. *International Journal of Computer Applications*. Volume 36(11):8-12.

Kumar, V., & Chadha, A. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *International Journal of Advanced Computer Science and Applications*, 2 (3):80-84.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33(1):159-174.

Pimpa, C. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I.

Quinlan, J. (1993). C4.5: Programs for Machine Learning.

Ramaswami, M. a. (2009). A Study on Feature Selection Techniques in Educational Data Mining. INTERNATIONAL WORKING GROUP ON EDUCATIONAL DATA MINING, Vol. 1, Issue 1.

Rubyl, J., & David K. (2015). Analysis of Influencing Factors in Predicting Students Performance Using MLP - A Comparative Study. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 2.

Thai, N. (2007). A comparative analysis of techniques for predicting academic performance. 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge.