Revisión de modelos de predicción de la deserción estudiantil superior en las universidades

Review of models for predicting universities student dropout

DAZA VERGARAY, Alfredo[1]

RESUMEN

Este artículo se propone mostrar los diversos esfuerzos realizado sobre la investigación realizada en la deserción universitaria en el mundo, lo cual amerita hacer una revisión integra. El objetivo es poder identificar las variables y los algoritmos de máquinas de aprendizaje más utilizados así como los modelos propuestos de cada investigador. Se realiza una breve descripción de las máquinas de aprendizajes más utilizados, luego se realiza una revisión de las investigaciones para posteriormente mostrar los modelos propuestos más actuales. Finalmente se concluye que las variables más utilizadas son: Rendimiento académico de la universidad, rendimiento académico del colegio, edad, sexo, deserción y las técnicas de minería de datos más usadas es las redes neuronales y los arboles de decisiones.

Palabras clave: Minería de datos, algoritmos de maquina de aprendizaje, deserción universitaria, predicción.

ABSTRACT

This article aims to show the various efforts made on the research conducted at the University dropout in the world, which warrants a revision integrates. The goal is to identify the variables and machine learning algorithms commonly used and any proposed models of each investigator. A brief description of the most widely used learning machines will be made after a review of the research will be made later to show the most current models proposed. Finally we conclude that the variables most commonly used are: Academic Performance of University College Academic Performance, Age, Sex, Defection and techniques of data mining is most commonly used neural networks and decision trees.

Key words: Data Mining, Machine Learning Algorithms, College Dropout, Prediction.

[1] Facultad de Ingeniería de Sistemas. Universidad Cesar Vallejo. Lima, Perú. adaza@ucv.edu.pe

INTRODUCCIÓN

La minería de datos **[1]** es el proceso de analizar los datos desde diferentes perspectivas y resumir los resultados como información útil. Se define como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles en última instancia en los datos.
En la actualidad, la minería de datos es una tecnología nueva con un gran potencial en el sistema de información, con muchos campos de aplicación como la detención de fraudes, marketing, análisis de ADN, en la industria financiera, en las telecomunicaciones, en la bioinformática, la lucha contra el terrorismo, de igual forma otro dominio de aplicación son las instituciones de educación, dado al aumento histórico en las bases de datos **[2,3]**
Un problema por el cual las Universidades atraviesan es la alta tasa de deserción de los estudiantes, la cual se debe a diferentes factores como por ejemplo los factores económicos, factores de salud, factores vocacionales y factores familiares. Los factores antes mencionados influyen en que un alumno abandone sus estudios en los 3 primeros años de sus estudios por lo cual se hará una revisión de los modelos que se han estudiado, en la cual nos permita analizar cuáles son las variables más utilizadas en los modelos estudiados, las técnicas de minería de datos utilizadas y los resultados obtenidos en la precisión de los estudios.
El objetivo de la investigación es presentar los estudios sobre la deserción de los estudiantes en las universidades de la siguiente manera, la sección dos presenta el marco teórico donde se hace una breve descripción de los conceptos básicos para el entendimiento del trabajo, así como se hace una revisión exhaustiva de las investigación realizadas, y se describe los últimos siete modelos en donde se muestra las técnicas utilizadas en cada trabajo así como los pasos que se realizaron en cada uno de ellos para la obtención de los resultados. Finalmente se realizaran algunas conclusiones importantes para la propuesta de un nuevo modelo hibrido basado en árboles de decisión y redes neuronales.

2. CUERPO DE LA REVISION

2.1 Árboles de clasificación.

Desarrollado por Breiman et al. (1984), **[11]** el objetivo de encontrar que variable independiente (s) puede realizarse sucesivamente una decisión de los datos dividiendo el grupo original de los datos en pares de subgrupos en la variable dependiente. Es importante señalar que a diferencia de la regresión que devuelve un subconjunto de variables, los árboles de clasificación puede clasificar ordenar los factores que afectan a la tasa de retención.

2.2 Redes neuronales.

Las redes neuronales **[11],** como el nombre implica, tratan de imitar las neuronas interconectadas en los cerebros de animales con el fin de hacer que el algoritmo sea capaz de realizar el aprendizaje complejo para la extracción de patrones y detectar tendencias. Se basa en la premisa de que estructuras de datos del mundo real son complejos, y por lo tanto requiere el aprendizaje de sistemas complejos. Una red neuronal entrenada puede ser visto como un "experto" en la categoría de información que ha sido dada a analizar. Este sistema experto puede proporcionar proyecciones dado nuevas soluciones a un problema y la respuesta **"qué pasa si"**. La red neuronal típica se compone de tres tipos de capas, a saber, **la capa de entrada**, **capa oculta y la capa de salida**. Es importante observar que hay tres tipos de capas, no tres capas, en la red puede haber más de una capa oculta y depende de la complejidad del investigador de realizar el modelo.
La capa de entrada contiene los datos de entrada; la capa de salida es el resultado mientras que la capa oculta realiza la transformación y manipulación de datos. Debido a que la entrada y la salida están mediadas por la capa oculta, las redes neuronales son comúnmente visto como un "recuadro negro".

2.3 Ashutosh Nandeshwar (2011)[5]. Realizó un trabajo para predecir si los estudiantes se mantendrá durante los tres primeros años de una licenciatura en la universidad, después de haber realizado el estudio, el autor considero para la realización de su estudio 103 variables como se muestra en la Tabla 1, de los cuales indica que los factores que resultaron ser de carácter importante son: el sueldo familiar, la situación socio económica de la familia, el alto promedio escolar y el rendimiento académico de las pruebas en la educación superior.

Tabla 1. Variables identificadas

ATRIBUTO	DESCRIPCIÓN DE AYUDA FINANCIERA	ATRIBUTO	DESCRIPCIÓN DE INDICADORES DE RENDIMIENTO
FinAidAwardType_G	Monto de subvenciones de ayuda financiera	ACT_COMP	ACT puntaje Integral(Antiguo)
FinAidAwardType_J	Monto de ayuda financiera en los puestos de trabajo	ACT_ENGL	ACT puntaje de Ingles(Antiguo)
FinAidAwardType_L	Importe de ayuda financiera de prestamos	ACT_MATH	ACT puntaje de matemática(Antiguo)
FinAidAwardType_S	Importe de ayuda financiera de beca	ACT1_COMP	ACT puntaje Integral(nuevo)
FinAidAwardType_W	Importe de ayuda financiera de renuncia	ACT1_ENGL	ACT puntaje de Ingles(nuevo)
FinAidDEPENDENCY	Estado de dependencia	ACT1_MATH	ACT puntaje de matemática(nuevo)
FinAidFATHER_ED	Nivel de Educación del padre	ACTEQUIV	ACT equivalente al puntaje
FinAidFATHER_WAG	Ingresos del padre	MaxACT	Máximo del puntaje ACT y el equivalente ACT
FinAidMOTHER_ED	Nivel de Educación de la madre	COMP_READ	Leer puntuación de alcance
FinAidMOTHER_WAG	Ingresos de la madre	COMP_WRITE	Escribir puntuación de alcance
FinAidOfferedInd	Indicador de ayuda financiera ofrecida	SAT_TOT	Puntaje total de SAT
offered indicator	Ingreso bruto de los padres	SAT_VERB	Puntaje verbal de SAT
FinAidPARENT_HOU	Tamaño del hogar de los padres	HS_CODE	Código de la escuela secundaria
FinAidPARENT_MAR	Estado Civil de los padres	HS_GPA	Rendimiento Académico del Colegio
FinAidPARENT_TAX	Tipo de formulario impuesto de los padres	HS_PERCENT	Percentil de la Escuela secundaria
FinAidSPOUSE_WAG	Salarios del cónyuge	HS_RANK	Posición en la Escuela Secundaria
FinAidSTUDENT_AG	Ingreso bruto de los estudiantes	HS_SIZE	Tamaño de clase de la escuela secundaria
FinAidSTUDENT_HO	Tamano de familia de los estudiantes	RankHSGPA	Percentil del rendimiento académico de todos los estudiantes del primer año
FinAidSTUDENT_MA	Estado Civil de los Estudiantes	RankMaxACT	Percentil del Act máximo de todos los estudiantes del primer año
FinAidSTUDENT_TA	Tipo de formulario impuesto de los estudiantes	ANTH18	Inscrito en el curso de Antropología
FinAidSTUDENT_WA	Salario del estudiantes	BSCI10	Inscrito en el curso de ciencias biológicas
FirstGenInd	Indicador de la primera Generación	CHEM10	Inscrito en el curso de química
TotalFinAidOffered	Total de ayuda ofrecida	ENG10	Inscrito en el curso de ingles
		ENG11	Inscrito en el curso de ingles
		GEOL11	Inscrito en el curso de geología
		LEST16	Inscrito en cursos de distracción
		MATH10	Inscrito en el curso de nivel 100 de matemática
		MATH11	Inscrito en el curso de nivel 110 de matemática
		MATH12	Inscrito en el curso de nivel 120 de matemática
		MATH14	Inscrito en el curso de nivel 14 de matemática
		PHY11	Inscrito en el curso de nivel 11 de física
		PEP15	Inscrito en el curso de nivel 15 de ed físico

Fuente : Adaptado a Ashutosh Nandeshwar (2011)[5].

En el presente trabajo se muestra un resumen (Tabla 2) de la literatura que revisó el autor, en la cual detalla la técnica usada y la precisión que se obtuvieron en cada uno de los estudios realizados.

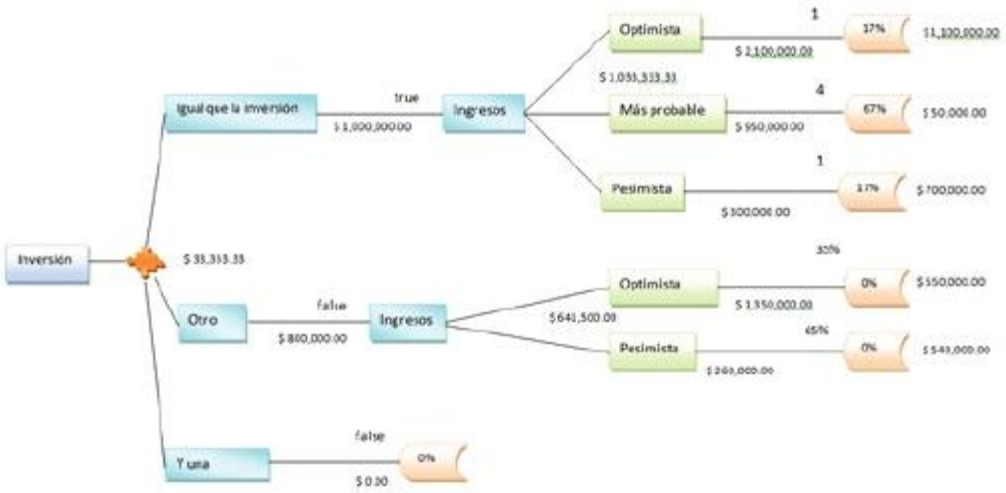
Tabla 2. Trabajos revisados

AUTOR (AÑOS)	NOTAS	TAMAÑO DEL GRUPO	SE DEBE MANTENER #	SE DEBE MANTENER %	MEDIDA EXACTA DEL GRUPO
Spady (1971)		683	615	90.04	R^2 de 0.3132 de hombres y 0.3879 de mujeres
Bean (1980)		906	769	84.88	R^2 de 0.22 de mujeres y 0.09 de hombres
Terenzini y Pascarella(1980)	estudio 1	379	60	15.80	R^2 de 0.246
	estudio 3	518	428	82.63	R^2 de 0.256
	estudio 5	763	673	88.20	R^2 de 0.309
	estudio 6	763	673	88.20	R^2 de 0.476 de hombres y 0.553 de mujeres
Stage(1989)		323	264	91.00	
Dey y Austin(1993)		947	152	16.00	Multiplique R 0.354,0.351 y 0.323

Fuente: Adaptado a Ashutosh Nandeshwar (2011)[5].

Para la realización del estudio se analizó 6 técnicas que son: one-R, C4.5, Adtree, Reyes bayesinas, Networks y radial biasnetkwords de las cuales uso para su experimento : arboles de decisión figura 1 y redes bayesianas obteniendo una precisión del 90%.

Figura 1. Árboles de decisión

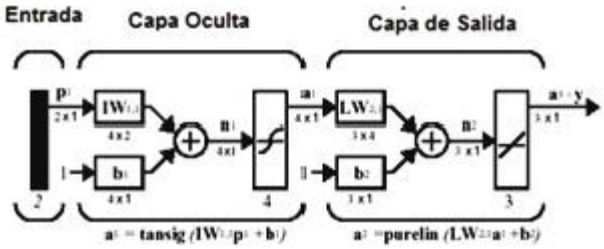


Fuente: Adaptado a Ashutosh Nandeshwar (2011)[5]

2.4 Ruba Alkhasawneh (2010)[4].

Realizó una revisión sobre métodos estadísticos tradicionales aplicados a la deserción de estudiantes y además técnicas cualitativas para identificar los factores que afectan la retención de los estudiantes, en donde el autor critica que los métodos estadísticos muestran menor precisión que los métodos de minería de datos por lo cual desarrolla dos modelos de redes neuronales (Figura 2) que utilizan una red de propagación de alimentación hacia adelante para predecir la retención de estudiantes en los campos de la ciencia y la ingeniería utilizando como variable principal el rendimiento académico (GPA)

Figura 2. multilayer feed forward back propagation network



Fuente: Ruba Alkhasawneh (2010)[4].

El primer modelo que plantea el trabajo de investigación predice la retención de estudiantes de primer año de ingreso y identifica factores correlaciónales entre los factores pre-universitarios. El segundo modelo clasifica a los grupos de primer año en tres clases: en situación de riesgo si el GPA es Menor que 2.7, intermedio si el GPA está entre 2.7 y 3.4, y el riesgo es alto si el GPA mayor a 3.4. El experimento se realizó con un total de 338 estudiantes de los cuales 44% representa a Ingeniería y el 56% corresponde a los alumnos de ciencias. En las tablas mostradas en la parte inferior (Tabla 3 y Tabla 4) se muestra los resultados obtenidos relacionados con la precisión del modelo.

Tabla 3. Los resultados del valor r y la mejor precisión

VARIABLE	S&E	CIENCIA	INGENIERÍA
Valor R	0.54	0.57	0.59
Precisión	68%	70.5%	68.9%
Total	338	190	148

Tabla 4. Resumen de resultados de análisis de errores

VARIABLE	S&E	CIENCIA	INGENIERIA
Mínimo	0.002808	0.000519	8.06E-05
Máximo	2.623909	1.652878	2.772855
media	0.41657	0.408178	0.410695

2.5 Mario Jadric (2009) [6] realizó un estudio de la deserción de estudiantes usando la **metodología SEMMA** creando un modelo en el software SAS Enterprise Miner, para luego aplicarlas en técnicas de minería de datos como: regresión logística, árboles de decisión y redes neuronales en la cual utilizó las variables que se muestran en la Tabla 5.

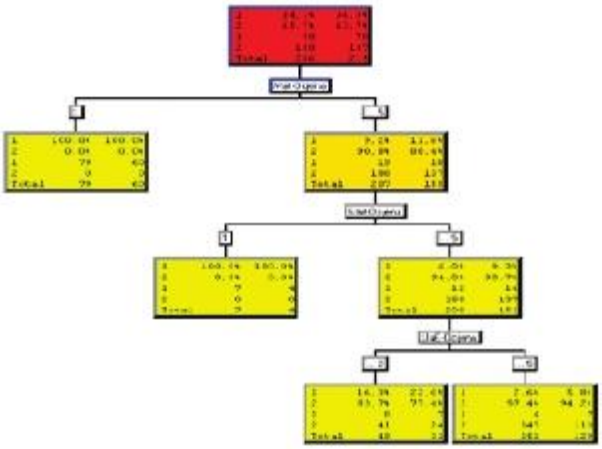
Tabla 5. Variables identificadas

VARIABLES		
ID	Sexo	Estado
Programa de Estudios	Calificaciones del Padre	Calificaciones de la Madre
Condición Social	Indicador de la Vivienda	Agrupación del examen de entrada

Fuente: Adaptado a Mario Jadric (2009) [6]

La realización del experimento lo realizó con cada uno de las técnicas antes mencionadas tomando a 286 estudiantes y después de haber realizado el entrenamiento con la técnica de árbol de decisiones se puede observar que el 98 estudiantes desertan mientras que 188 estudiantes continúan sus estudios después del segundo año como se muestra en la Figura 3.

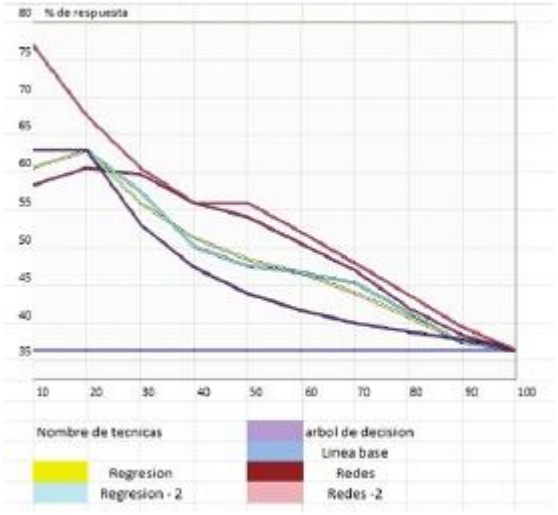
Figura 3. Análisis por Arboles de Decisión



Fuente: Adaptado a Mario Jadric (2009)[6]

Después de realizar las comparaciones de los métodos experimentados se determinó que las redes neuronales se comportan muy bien en problemas de clasificación más complejos según la Figura 4.
Su desventaja, en comparación con los métodos más sencillos, es el proceso que es relativamente lento y exigente del modelo de "aprendizaje" (optimización de los factores de peso)

Figura 4. Evaluación y comparación de modelos

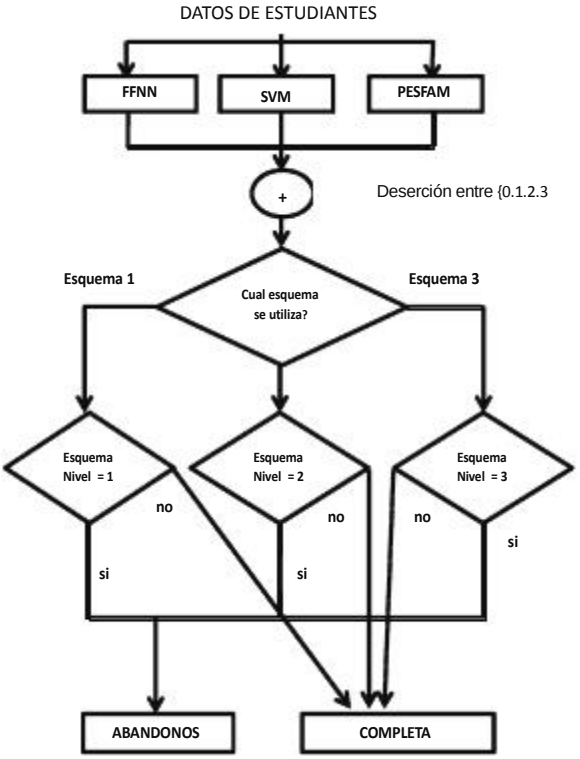


Fuente: Adaptado a Mario Jadric (2009)[6]

En la figura 4 mostrada en la parte superior se puede observar que los primeros 30% de puntuación del modelo, la red neuronal y el árbol de decisión proporcionan los mismos resultados, es decir, 100% de los estudiantes que abandonan sus estudios.

2.6 Loanna Lykourentzou (2009) [7] desarrollo un método de predicción de deserción en los cursos de e-learning, basado en tres técnicas populares de aprendizaje automático. Las técnicas de aprendizaje automático utilizadas son redes neuronales con aprendizaje hacia adelante, máquinas de soporte de vectores y conjunto probabilístico simplificada ARTMAP difusa.
También indica que una sola técnica puede fallar para clasificar con precisión a algunos estudiantes de e-learning, mientras que otro puede tener éxito, en donde realizó tres sistemas de toma de decisiones basados en el esquema mostrado en la Figura 5, que se combinan para obtener diferente resultados de las tres técnicas de máquinas de aprendizaje.

Figura 5. Esquema de decisión usada por el método propuesto



Fuente: Adaptado Loanna Lykourentzou (2009) [7]

Para la realización del experimento se utilizó las variables que no son variantes en el tiempo (demográficas) y las variables que son cambiantes en el tiempo (sesiones de aprendizaje) mostradas en la tabla 6.

Tabla 6. Atributos de estudiantes usadas para el entrenamiento y testeo de redes de aprendizaje automático

	CATEGORÍA RELACIONADA CON LA LITERATURA	ATRIBUTO	RANGO DE VALORES
Atributos invariantes en el tiempo	Demográfico	Genero	Masculino, Femenino
		Residencia	Capital, Provincia
		Experiencia de trabajo	>=0 años
	Rendimiento Académico	Nivel de Educación	Básico, intermedio, alto, grado de master , Grado PhD
		Idioma Ingles	Elemental, básico, alto,
Atributos variables en el tiempo	Calificación del examen con opciones múltiples		completo 0–20
	Calificación de Proyecto		0–100
	Fecha de Presentación del Proyecto (Días contados a partir de la sección de plazo)		>= 0
	Actividad de la sección		>= 0

Fuente: Loanna Lykourentzou (2009) [7]

El método propuesto fue examinado en términos de precisión general y la sensibilidad, la precisión que se obtuvo se encontraba en un rango del 75 – 85% y sus resultados obtenidos son significativamente mejor a los métodos usados de manera independiente.

2.7 Dekker (2009) [8] realizó un trabajo de minería de datos aplicado a la educación en los alumnos de Ingeniería Eléctrica de la Universidad Tecnológica de Eindhoven (en donde la deserción es de 40%) después del primer semestre de sus estudios o incluso antes de entrar en el programa de estudio, el objetivo del trabajo es determinar qué datos (variables) son los predictores de la deserción para poder de esta manera determinar cuando la predicción es mejor, las variables utilizadas en el desarrollo de trabajo de investigación, esta basadas en datos **pre-universitarios** como se muestra en la Tabla 7.

Tabla 7. Atributos de estudiantes usadas para el estudio

ATRIBUTOS	TIPO	DESCRIPCIÓN
IDNR	Numérico	Solo para chequear los datos
Año VWO	Nominal	Principales cambios en el sistema educativo Holandés {1..4, 'n/a'}
Curricuclo VWO	Nominal	Currículo de educación preuniversitaria ,{1..5, 'n/a'}
Numero de cursos VWO	numérico	Número de cursos tomados
Promedio VWO	Nominal	{ n/a, pobre, promedio, sobre el promedio, bueno, excelente }
Número de cursos de ciencias VWO	Nominal	{ n/a, < 3, 3, >3 }
Promedio en ciencia VWO	Nominal	As VWO mean
Número de cursos de matemáticas VWO	Nominal	{n/a, 0,1,2}
Promedio de matematica VWO	Nominal	As VWO mean
Educacion		otro}
Año HO	Nominal	Igual categoría VWO año
Grado HO	Nominal	As VWO mean
Año Gap	Nominal	{n/a, < -1, -1, 0, 1, >1 }
Clasificación	Nominal	{-1, 1}

Fuente: Dekker (2009) [8]

El experimento se realizó con la participación de 648 estudiantes del programa de Ingeniería Eléctrica, en donde los resultados que se obtuvieron muestran que los clasificadores más sencillos e intuitivos (**árboles de decisión**) dan como resultado significativo útil una precisión entre 75 y 80%.

2.8 Joe J.J [9] realizó un trabajo en la cual desarrollo una combinación de cinco modelos de retención y hace uso de cuatro metodologías de modelamiento destacados en las cuales se encuentran las redes neuronales, regresión logística, análisis discriminante y modelo de ecuaciones estructurales. En los modelos de retención de estudiantes que propuso consideró diferentes conjuntos de datos que van desde 9 hasta 71 variables de entrada, entre ellos variables de **factores cognitivos y / o no cognitivos**, las cuales se pueden observar en la Figura 6.

Figura 6. Predicción de retención de estudiantes en ingeniería

	VARIABLES PARA LA RETENCIÓN UN AÑOS DESPUÉS		
Factores No cognitivos	atributos	**Factores cognitivos**	Puntuación de la Escuela Secundaria
	Liderazgo		Rendimiento Académico de Escuela Secundaria
	Mayor decisión		Promedio de Escuela Secundaria en Matem ática, Ciencia Ingles
	auto eficacia		Número de veces que le toma matemática
	Equipo		
	Motivación		

Fuente: Joe J.J [9]

El experimento lo realizó con 1508 estudiantes entre los cuales 289 eran mujeres y 1219 hombres, los resultados del experimento de los cinco modelos propuestos muestran que el **método de red neuronal** produce los mejores resultados de predicción con respecto a los otros tres métodos de manera consistente dando una precisión de **71.9% en el modelo C usando variables cognitivas y no cognitivas.**

2.9 Wilairat Yathongchai[10] realizó un estudio en la que considera que existen tres factores importantes que afectan la tasa de deserción de los estudiantes. Estos factores son las condiciones relacionadas con los estudiantes antes de **su ingreso**, los factores relacionados con los estudiantes durante los **períodos de estudio en la universidad**, y todos los factores que incluyen el valor del objetivo que se predicen para el análisis de factores.
El estudio lo realizó en la Universidad Buriram Rajabhat, con 731 estudiantes de los cuales 251 estudiantes desertaron, la información fue obtenida de diferentes tablas de la base de datos académica MIS y las variables que se consideraron para el estudio se muestra en la Tabla 8. Para realizar las pruebas utilizó la técnica de árboles de decisión, basado en la clasificación, J48 o C4.5 y Naive Bayes, como herramienta de desarrollo se utilizó el software weka con 513 casos para realizar el entrenamiento y 218 casos para realizar la validación del modelo y se obtuvieron los resultados que se muestra en la Tabla 9.

Tabla 8. Variables relacionadas con los estudiantes.

VARIABLE	DESCRIPCIÓN	POSIBLES VALORES
Programa	Programa para estudiar en la facultad de ciencias	{230, 240, 241, 243,247, 249, 264, 265,284, 285, 286}
GPA1-GPA4	GPA entre el term1-term4(dentro del año académico 2008-2009)	débil, Medio, bueno, mejor}? débil =GPA< 1.6 Medio=GPA 1.6-1.99 bueno=GPA 2.0-2.5 mejor=GPA>2.5
GPAX del colegio	GPAX de la educación secundaria	numero
Programa del colegio	Programa de estudio en la educación secundaria	{1, 2, 3}? 1 = ciencia + matemática 2 = lenguaje + matemática 3 = otro.
Tamaño del colegio	Tamaño del colegio	{Pequeño, Mediano ,Grande }
Préstamo	Préstamo del estudiante	{Si, No}? si = el estudiante tiene préstamo No = el estudiante no tiene préstamo
causa	Causa de abandono	Estudio, jubilado, finanza, cambio de programa
Termino de la		
Deserto	Estado de abandono	{Si, No}

Fuente: Wilairat Yathongchai[10]

Tabla 9. Comparación de los resultados de dos algoritmos de clasificador sobre todos los factores.

CLASIFICACIÓN	J48		REYES BAYESIANAS	
	Conjunto de Validación	Conjunto de prueba	Conjunto de Validación	Conjunto de prueba
Precisión	87.00%	84.86%	85.08%	82.11%
Tasa TP	0.87	0.849	0.851	0.821
Tasa FP	0.073	0.066	0.033	0.033
Tasa TN	0.843	0.831	0.864	0.872
Tasa FN	0.851	0.849	0.851	0.521

RESULTADOS

Para obtener las variables más usadas en los trabajos de investigación, así como proponer las nuevas variables según sea el caso de cada Universidad se deben seguir los siguientes pasos:

a) Se identifica todas las variables de entradas que se han utilizado en los modelos estudiados en otras investigaciones a nivel mundial según la frecuencia de uso (Tabla 11). (A) las cuales se describen a continuación :

X1 = Código de Identificación.

X2 = Rendimiento Académico de la Universidad.

X3 = Rendimiento Académico del Colegio.

X4 = Edad.

X5 = Sexo.

Y = Deserción.

b) Se realiza el análisis de las base de datos de las universidades, para luego hacer la extracción de las variables aplicando la metodología CRISP (B) como se muestra en la Figura 6.

c) Se realiza la propuesta de nuevas variables que afectan la deserción, que no han sido usados por los modelos propuestos.(C)

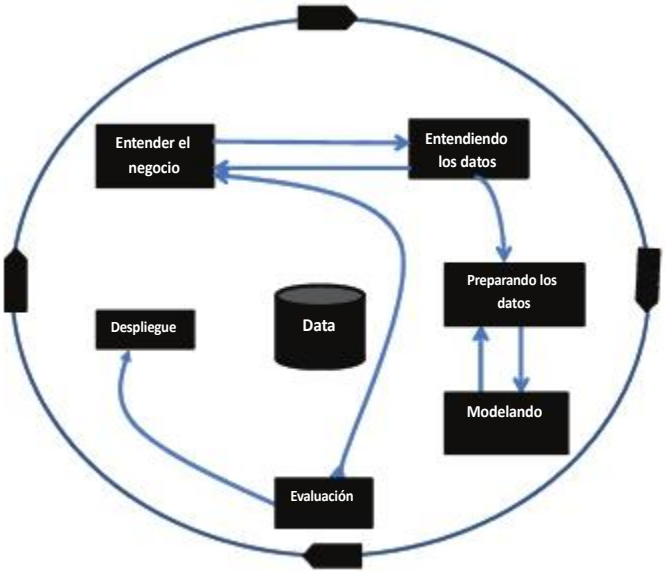
Para llegar a deducir que las variables de entrada **(V.E)** estarán dadas por la siguiente ecuación:

V.E = (A ∩ B) U (B ∩ C) = (A U C) ∩ B

X_i , i = 1, 2, 3,4,5,6,.....n

Variables de salida: predecir la deserción **V.S**= Y

Figura 7. Fases del proceso de minería de datos, basado en la metodología CRISP –DM.



Fuente: adaptado Chapman et., 2000[12]

d) Luego se realiza la correlación de cada una de las variables de **entrada (X_i)** con la variable de **salida (Y)**, para determinar el grado de relación entre las variables.

e) Finalmente se aplica las variables de entrada obtenidas a las técnicas de minería más usadas (redes neuronales o arboles de decisión) y que han dado buenos resultados para la obtención de un modelo de predicción.

CONCLUSIONES

En esta investigación, se ha analizado los diversos modelos estudiados de la deserción de estudiantes de las universidades a nivel mundial.

Las técnicas utilizadas para el desarrollo en la deserción de estudiantes se muestran en la siguiente tabla (Tabla 10)

Tabla 10. Técnicas de minería de Datos

N°	TÉCNICAS	AUTORES	AÑO
1	Redes neuronales hacia adelante	Ruba Ikhasawneh	2010
		Ioanna Lykourentzou	2009
2	Árboles de decisión	Mario Jadri´c	2009
		Laurence G. Moseley	2007
		Dekker	2009
		WILAIRAT YATHONGCHAI	2003
		Ashutosh Nandeshwar	2011
3	Máquinas de vectores soporte	Ioanna Lykourentzou	2009
4	Análisis discriminante	Joe J.J. Li	2009
5	Modelo de ecuaciones estructurales	Joe J.J. Li	
6	Regresión logística	Joe J.J. Li	2009
		Mario Jadri´c	2009
7	Redes neuronales	Mario Jadri´c	2009
		Joe J.J. Li	
		Ashutosh Nandeshwar	2011
		Ioanna Lykourentzou	2009

Fuente: Elaboración propia

En la Tabla 10 se puede observar que las técnicas más utilizadas son árboles de decisión y redes neuronales. Según el documento en estudio realizado se identificó las variables más utilizadas en los estudios (Tabla 11) de los investigadores las cuales son:

Tabla 11. Variables identificadas más usada

N°	VARIABLE	AUTOR	AÑO
1	GPA	Ruba Alkhasawneh	2010
		Mario Jadri´c	2009
		Joe J.J. Li	2009
		WILAIRAT YATHONGCHAI	2003
2	Sexo	Mario Jadri´c	2009
		Ioanna Lykourentzou	2009
		Laurence G. Moseley	2007
3	Fecha de Nacimiento	Mario Jadri´c	2009
		Laurence G. Moseley	2007
4	Rendimiento de los estudiantes en cada semestre en diferentes módulos	Laurence G. Moseley	2007
		WILAIRAT YATHONGCHAI	2003
5	ID	Mario Jadri´c	2009
		Dekker	2009
6	Programa de estudios	Mario Jadri´c	2009
		WILAIRAT YATHONGCHAI	2003

Fuente: Elaboración propia

En la Tabla 12 se puede observar la precisión más alta obtenida en los modelos antes estudiados fue de Laurence G. Moseley con una precisión de 94% con un análisis de sensibilidad de 84% y una especificidad de 70%

Tabla 12. Precisión obtenida por las investigación realizadas por otros autores

N°	PRECISIÓN	AUTORES	AÑO
1	90%	Ashutosh Nandeshwar	2011
2	70.5%	Ruba Alkhasawneh	2010
3	Percentiles	Mario Jadri´c	2009
4	75.855	Ioanna Lykourentzou	2009
5	81%	Dekker	2009
6	71.9	Joe J.J. Li	2009
7	94% (análisis de sensibilidad = 84,especificidad 70)	Laurence G. Moseley	2007
8	87.90	Wilairat Yathongchai	2003

Fuente: Elaboración propia

Además podemos concluir que de todos los modelos revisados la técnica estadística de análisis discriminante no son adecuados para predecir la deserción estudiantil superior.

REFERENCIAS BIBLIOGRÁFICAS

1. Mr. M. N. Quadri1 and Dr. N.V. Kalyankar,Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques, Global Journal of Computer Science and Technology
2. Han, J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition.
3. Chang, W.H.T. and Lee, Y.H. (2000). Telecommunications data mining for target marketing. Journal of Computers, 12(4), 60–74.
4. Ruba Alkhasawneh and Rosalyn Hobson, Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks, IEEE Global Engineering Education Conference (EDUCON) 2011.
5. Ashutosh Nandeshwar, TimMenzies, and Adam Nelson, Learning patterns of university student retention, Expert Systems with Applications 38 (2011) 14984–14996.
6. Mario Jadrić, Željko Garača and Maja Ćukušić, Student Dropout Analysis with Application of data Mining Methods, Management, Vol. 15, 2010, 1, pp. 31-46.
7. Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis and Vassili Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques, Computers & Education.
8. Gerben W. Dekker, Mykola Pechenizkiy2 and Jan M. Vleeshouwers, Predicting Students Drop Out: A Case Study, Educational Data Mining 2009.
9. Joe J.J. Lin, P.K. Imbrie and Kenneth J. Reid, Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results, Engineering Education Symposium 2009.
10. Wilairat Yathongchai, Chusak YAthongchay, Kittisak Kerdprasop and Nittaya Kerdprasop, Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out, Latest Advances in Educational Technologies.
11. Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell and Charles Kaprolet A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year, Journal of Data Science 8(2010), 307-325
12. Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C. y Wirth R. (2000). CRISP-DM 1.0 Step-by-step Data Mining Guide. Disponible en :<http://www.crisp-dm.org/CRISPWP-0800.pdf>. Última consulta el 28.04.2011

Recibido: 08 abril 2014 | **Aceptado:** 05 junio 2014