Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educación Superior Privada.

A model based on decision trees to predict student dropout in Private Higher Education

DAZA VERGARAY, Alfredo

RESUMEN

Las técnicas de minería de datos permiten obtener información útil que se encuentra oculta en grandes base de datos que en su mayoría solo son usados para realizar operaciones transaccionales, así como archivos que aún no han sido ingresado a las base de datos. La información al ser explotado de manera correcta permite mejorar la toma de decisiones así como ofrecer ventajas competitivas con respecto a otras empresas.
Debido a la gran cantidad de datos que tienen las Instituciones de Educación Superior Universitaria en este trabajo de investigación se propone hacer uso de las técnicas de minería de datos para predecir la deserción o el abandono en la Educación Superior Privada. Para el desarrollo de proyecto se usó la metodología CRIPS-DM con la herramienta comercial spss clementine 12.0, para los cuales se hicieron uso de la técnica de minería de datos árboles de decisión, para lo cual se utilizaron 1761 datos de los estudiantes de la Universidad Privada César Vallejo, comprendidos del semestre 2009-I al semestre 2013-II de la Escuela profesional de Ingeniería de Sistemas con 27 atributos para cada uno de ellos que están relacionadas con la deserción del alumno, que fueron extraídos del área de registros académicos, Asuntos Estudiantiles y del área de Informática.
Para el desarrollo del proyecto se hizo uso del algoritmo de árboles de decisión en donde se hizo el entrenamiento, validación y prueba con 100 datos nuevos en donde se obtuvo una precisión de 89%.

Palabras clave: Minería de Datos, Algoritmos de Maquina de Aprendizaje, Deserción Universitaria, Predicción, Árboles de Decisión

ABSTRACT

The data mining techniques allow to obtain useful information that is hidden in large database that are mostly only used to perform transactional operations, as well as files that have not yet been entered into the database. The information to be exploited properly can improve decision-making and deliver competitive advantages over other companies.
Due to the large amount of data with Institutions of Higher Education University in this research it is proposed to use data mining techniques to predict the desertion or abandonment in Private Higher Education. For the development of project CRIPS-DM methodology was used with commercial tool Spss v. 12.0, for which use of mining technique trees decision data were made, for which 1761 data students used Private University Cesar Vallejo, including the semesters 2009-I semester 2013-II of the professional School of Systems Engineering with 27 attributes for each that are related to the defection of the student, which were recovered from the area of academic records, and Student Affairs Computer area.
For the development of the project made use of decision trees algorithm where training, validation and testing with 100 new data where an accuracy of 89% was obtained was made.

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

Facultad de Ingeniería de Sistemas. Universidad César Vallejo-Lima. adaza@ucv.edu.pe

INTRODUCCIÓN

Uno de los problemas principales que se enfrentan las Universidades a nivel nacional y mundial en el año 2013, es la deserción Universitaria [1]la cual ha sido investigada parcialmente, en donde el primer factor importante es la mala selección de los postulantes al momento de ingresar en la Universidad, sin embargo , según Ríos[54] la deserción estudiantil se relación con el rendimiento académico(factores académicos) que se forma en la etapa escolar y por consecuencia el índice de estudiantes que se retira de las Universidades Privadas es en la etapa temprana comprendida entre el primer y cuarto ciclo, además existen otros factores que implican la deserción como los problemas económicos, administrativos, políticos, vocacionales, académicos y institucionales.

En ese sentido, el estudio de los factores e índices que afectan a la deserción ha tomado gran importancia en los últimos años y la necesidad de identificar y predecir la deserción de los estudiantes en los primeros ciclos es indispensable para tomar decisiones pertinentes y poder disminuir este problema que afecta a las Universidades de Educación Superior.

un trabajo de investigación realizado por Aleyda Restrepo[73] Se realizó un estudio descriptivo en 36 estudiantes desertores del programa de enfermería de la Universidad Libre de Pereira y se caracterizaron por variables socio demográficas. Las proporciones calculadas se compararon con las observadas en un estudio anterior en la población general de estudiantes de enfermería y se establecieron las diferencias de proporciones además se indagó sobre las apreciaciones de los estudiantes desertores acerca de los motivos para ingresar al programa y los factores que los llevaron a la deserción mediante escalas tipo Likert.

En donde se pudo determinar que el 27.8% y 16,7% de la deserción ocurrió en primero y segundo semestre, respectivamente. No se observaron diferencias significativas en las variables de sexo, edad, estado civil y estrato con las observadas en la población general de estudiantes de enfermería

Así también el Instituto Nacional de Estadística e Informática del Perú INEI[2], dio a conocer que todo el problema de la deserción estudiantil se incrementa cuando se comienza a estudiar y trabajar, lo contrario a lo que sucede cuando se les exige la dedicación necesaria solo para estudiar.

Las Universidades de Educación Superior durante años han almacenado gran cantidad de información en sus base de datos que, son de gran importancia en el proceso de enseñanza y aprendizaje educativo permitiéndole a los Directores de Escuela, Decanos, coordinadores de Escuelas, docentes y alumnos entre otros, mejorar en cada uno de los procesos que intervienen, la minería de datos ha sido aplicado en diferentes ámbitos de la educación como se muestra en la Figura 1, por lo cual es una herramienta tecnológica que nos permite desarrollar o mejorar cada uno de estos proceso, es la minería de datos para la educación EDM, a través del desarrollo de aplicaciones de métodos de aprendizaje automático usando técnicas como redes neuronales, árboles de decisión, SVM, etc. con el fin de obtener patrones o modelos para ser evaluados e interpretados con el objetivo de generar un nuevo conocimiento que nos servirá de apoyo en las toma de decisiones en las universidades.

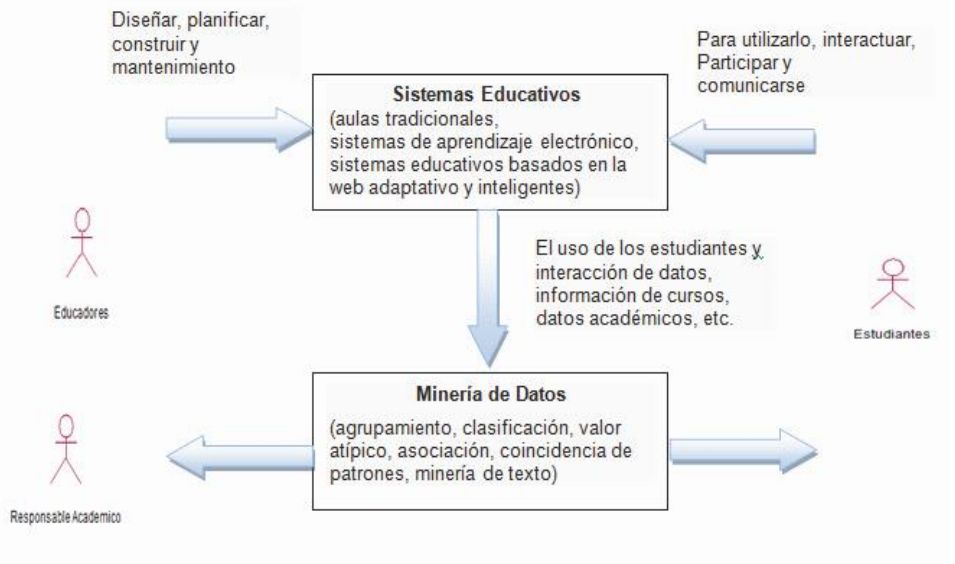

Figura 1: El ciclo de la aplicación de minería de datos en Sistemas Educativos

Fuente: Adaptado a Romero[1]

Relacionado con la perspectiva de la minería de datos basado en la técnica de redes neuronales, se ha realizado un trabajo de investigación por Ruba Alkhasawneh[2]en(2010)investigo sobre "Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks",en donde realizo una revisión sobre los métodos estadísticos tradicionales aplicados a la deserción de

[1]La deserción estudiantil se refiere a que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella o por demorar más tiempo de lo regular en finalizar, por repetir cursos o por retiros temporales.
[2]INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA (INEI). Compendios Estadísticos 1992,1993, 1994, 1995, 1998. Lima.

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
de redes neuronales 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 la variable rendimiento académico (GPA),obteniendo un precisión correcta de 70.5% en los resultados, lo cual no son suficientes para reducir la alta tasa de deserción.
La mayoría de los trabajos que intentan dar solución al problema de la deserción[72] están enfocados en determinar cuáles son los factores que más afectan al rendimiento de los estudiantes (abandono y fracaso) en los niveles educativos de educación básica, media y superior, haciendo uso de la gran cantidad de información que los actuales sistemas de información almacenan en las bases de datos.
Por consiguiente lo que se propone en esta investigación es dar solución a esta problemática de la deserción Universitaria en la Educación Superior desde la perspectiva de la Minería de Datos en la Educación, por medio del desarrollo de un modelo de árboles de decisión que nos muestre datos correctos, que permita predecir con alta precisión la deserción de los alumnos y con un procesamiento de datos corto.
Para la ejecución del presente proyecto tomaremos como referencia a la Universidad César Vallejo Lima-Este, debido a que se hizo las gestiones para los permisos necesarios en la recopilación de Datos, con la resolución N° 2387-A-2013/VRA.UCV LIMA ESTE.

MARCO TEÓRICO

Deserción universitaria
La graduación especialmente oportuna es una cuestión política cada vez más importante según DesJardins[32]. Los graduados en carreras universitarias ganan el doble que los graduados en la educación secundaria y seis veces más que los desertores de la universidad según Murphy[33]. Además de los beneficios económicos, las esposas de los graduados universitarios son más educados y los niños les van mejor en las escuelas y Universidades. Las tasas de graduación son consideradas como una de la eficacia institucional según Murtaugh[34]. Los estudiantes abandonan debido a diferentes razones, problemas académicos, preferencias académicas, matrimonio, problemas institucionales y su situación económica.

Almacén de datos (Datawarehouse)
Un almacén de datos se define como un conjunto de datos integrados, orientados a un tema de negocio, que varían con el tiempo, y que no son transitorios, los cuales soportan el proceso de toma de decisiones administrativas, de acuerdo a Inmon[38]

Minería de datos
La minería de datos se define como el proceso de extraer conocimiento útil y entendible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Por lo tanto la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos según Hernández.[39]

Modelo
El modelo según Fayyad[36], tiene dos factores importantes: la función del modelo (por ejemplo, clasificación y Clustering) y la forma de representación del modelo (por ejemplo, una función lineal de múltiples variables y una función de probabilidad gaussiana densidad). Un modelo contiene parámetros que se determinan a partir de los datos.

Predicción
En la predicción[56], el objetivo es desarrollar un modelo que se puede inferir un solo aspecto de los datos (variable predicha) a partir de una combinación de otros aspectos de los datos (variables predictoras). Predicción requiere tener etiquetas para la variable de salida para un conjunto de datos limitado, donde una etiqueta representa una información de confianza sobre el valor de la variable de salida en casos específicos. En algunos casos, sin embargo, es importante tener en cuenta el grado en que estas etiquetas pueden de hecho ser aproximado, o incompletamente fiable.
La predicción tiene dos usos principales dentro de la minería de datos en la Educación.
En algunos casos, los métodos de predicción puede ser utilizada para estudiar qué características de un modelo son importantes para la predicción, dando información acerca de la construcción subyacente.
Se trata de un enfoque común en los programas predecir factores de mediación en primer lugar.
En un segundo tipo de uso, los métodos de predicción se utilizan con el fin de predecir lo que el valor de salida podría ser dentro del contexto, en lo que no es deseable para obtener directamente una etiqueta para ese constructo.
Por ejemplo, podemos desear predecir el salario de los graduados de la Universidad Cesar Vallejo con 10 años de experiencia laboral, predecir el estilo de aprendizaje más adecuado en la enseñanza del curso de matemática I en la escuela profesional de Ingeniería de Sistemas, o el potencial de ventas en el mercado de un nuevo producto por su precio

Árboles de clasificación.
El árbol de decisión desarrollado por Breiman[48], trata de encontrar que variable independiente(s) puede hacer sucesivamente una decisión de los datos dividiendo el grupo de datos original en pares de subgrupos en la variable dependiente.
Es importante tener en cuenta que a diferencia de regresión que devuelve un subconjunto de las variables, los árboles de clasificación puede clasificar los factores que afectan a la tasa de retención.

REVISIÓN LITERARIA

Se han desarrollado varios trabajos de investigación con respecto a la deserción Universitaria Superior con estadística y minería de datos usando diferente algoritmos, para el presente artículo mencionares los que están más relacionados con el proyecto:

Ruba Alkhasawneh[2]. 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).

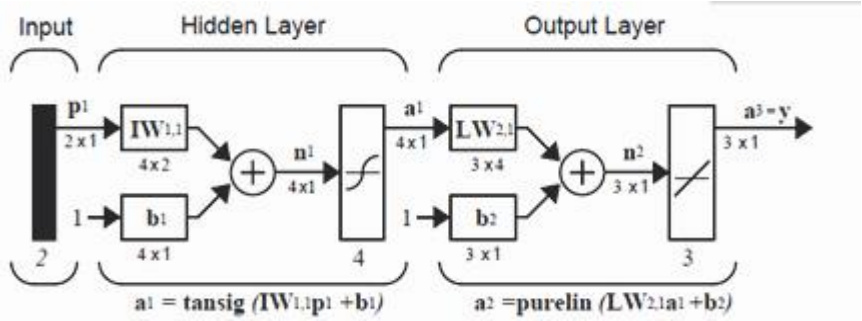

Figura 2: multilayer feed forward back propagation network

El primer modelo que plantea el trabajo de investigación predice la retención de estudiantes de primer año de ingreso e 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 riego 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 1 y 2 mostradas en la parte inferior se muestra los resultados obtenidos relacionados con la precisión del modelo.

Tabla 1: 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 2: Resumen de resultados de análisis de errores

Variable	S&E	Ciencia	Ingeniería
Mínimo	0.002808	0.000519	8.06E-05
Máximo	2.623909	1.652878	2.772855
Promedio	0.41657	0.408178	0.410695

Jadric[4]. Realizó un estudio de la deserción de estudiantes usando la metodología SEMMA para luego aplicar las 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 3.

Tabla 3: 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

Realizó el experimento con cada uno de las técnicas antes mencionadas usando 286 estudiantes, después de haber realizado el entrenamiento se puede observar que 98 estudiantes desertan mientras que 188 estudiantes continúan sus estudios después del segundo año como se muestra en la Figura 3.

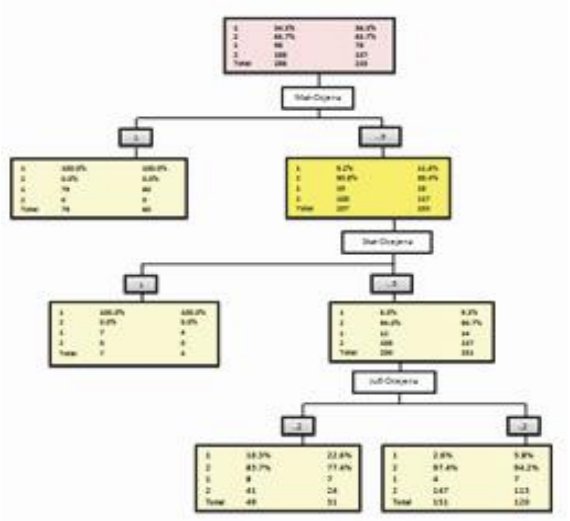

Figura 3: Análisis por árboles de decisión.

Fuente: Adaptado a Mario Jadric

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 modelo de aprendizaje debido a que el proceso es relativamente lenta y exigente, (optimización de los factores de peso) más trabajos de investigación relacionados con el tema se encuentran en la bibliografía.

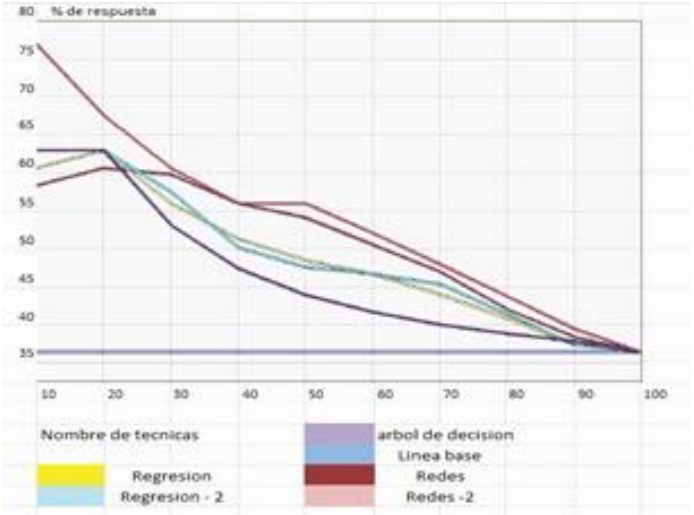

Figura 4: Evaluación y comparación de modelo.

Fuente: Adaptado a Mario Jadric (2009)[4]

METODOLOGÍA Y RESULTADOS

Para el presente trabajo de investigación se realizo de la siguiente manera:

a) Se identificó todas las variables de entradas que se han utilizado en los modelos estudiados en otras investigaciones. **(A)**

b) Se hizó análisis de las base de datos de la Universidad César Vallejo – Lima Este, para luego hacer la extracción de las variables aplicando la **metodología CRISP**

c) Se realizó la propuesta de nuevas variables que afectan la deserción, que no han sido usados por los modelos propuestos.(C) Las variables de entrada (V.E) están dadas por la siguiente ecuación:

$$V.E = (A \cap B) \cup (B \cap C) = (A \cup C) \cap B$$

$$X_i, i = 1, 2, 3,4,5,6,\ldots\ldots n$$

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

d) Se realizó 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) Como variables de entrada , se utilizó aquella variables que influyen en la deserción y han sido utilizados con mayor frecuencia por los investigadores antes mencionados , se considero como primeras variables para el estudio de la investigación las siguientes :

X1 = Código de identificación
X2= Rendimiento académico de la Universidad.
X3 = Rendimiento académico del Colegio
X4 = Edad.
X5 = Sexo.
Y = Deserción

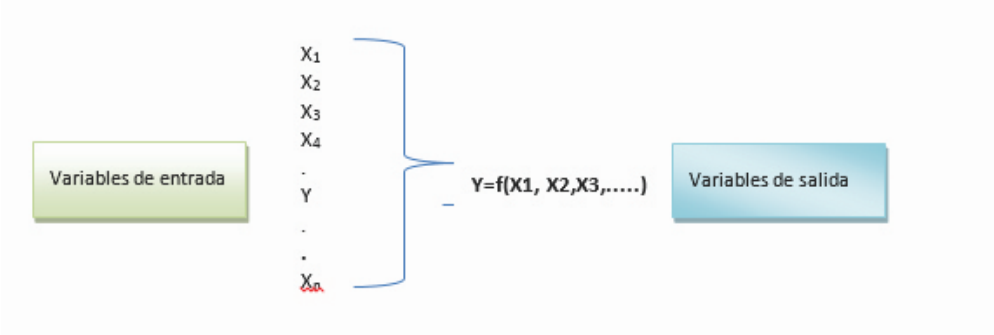

Figura 5: Relación de las variables

Fuente: Elaboración propia.

En el trascurso del desarrollo del presente trabajo de investigación se colocaron, más variables que permitan mejorar la precisión de los resultados del modelo de árboles de decisión para la construcción del proyecto se hizo en primer lugar el vaciado de los certificados de estudios informacion a un archivo excel llamado registros_academicos_tesis_2013_1_notas.xls como se encuentra en la figura 6

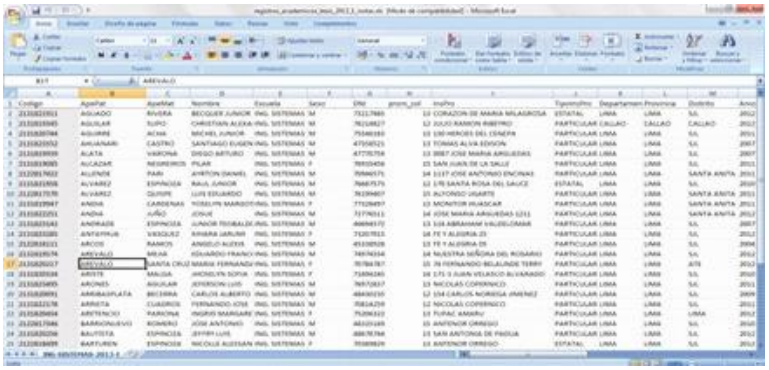

Figura 6: datos de certificado de estudios

Fuente: Elaboración propia.

Se realizó luego la selección de los campos que se encontraban en el sistema Gestor de Bases de datos sqlserver 2000 así como la limpieza de los datos, en la edad, la cantidad de tutoría llevadas obteniendo el archivo llamado obtenidos_de_motor_selecionado_limpiados.xls como se muestra en la figura 7.

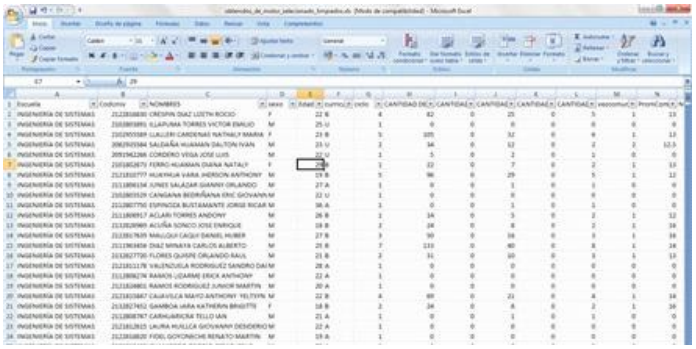

Figura 7: datos procesados y limpiados

Fuente: Elaboración propia.

El siguiente paso fue realizar la integración de las notas de secundaria de los alumnos, con los datos que han sido extraidos del sistema gestor de base de datos sqlserver 2000, que se encuentran en el archivo integrado_notas_de_colegio_m.xls como se muestra en la figura 8

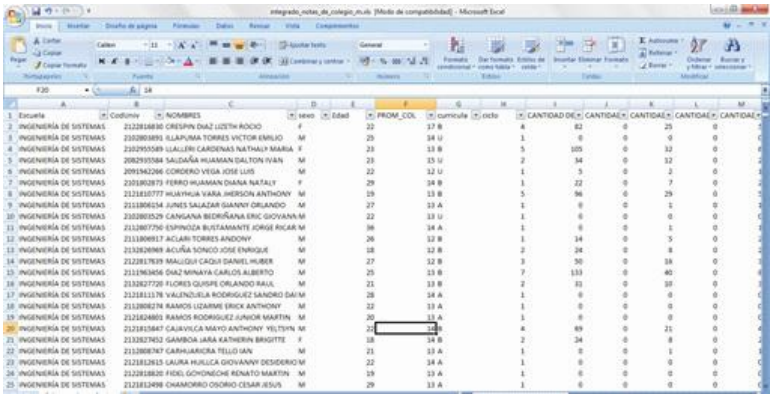

Figura 8: datos integrados

Fuente: Elaboración propia.

Por último se realizó la transformación de algunos datos que se encuentran ubicados en el archivo integrado_limpiado_transformados_final.xls como se muestra en la figura 9 (que nos permitió hacer uso de una manera adecuada en los algoritmos de redes neuronales , árboles de decisión, y el modelo hibrido propuesto con los algoritmos antes mencionados.

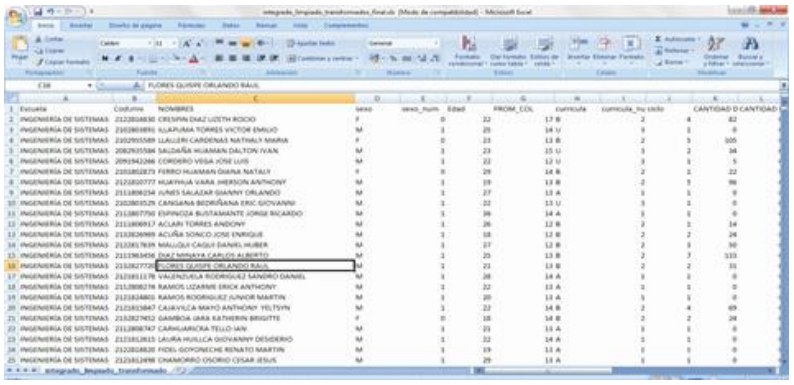

Figura 9: integración de datos

Fuente: Elaboración propia.

Entendiendo los datos

En este apartado se explica cada uno de los campos que se uso en el modelo basado en árboles de decisión, como se detalla en la tabla 4

Tabla 4: Descripción de los campos de la vista minable

N°	CAMPOS	DESCRIPCIÓN
1	Escuela	Este campo almacena la información de la escuela a la que pertenece el alumno en este caso será Ingeniería de Sistemas.
2	Codalumno	En este campo se almacena el código del alumno que representa la identificación única dentro de la Universidad.
3	Nombres	Se almacenara los nombres y apellidos de cada uno de los alumnos de la Escuela Académica Profesional de Ingeniería de Sistemas.
4	Sexo	Almacena el sexo del alumno en este caso toma dos valores F (femenino) y M (masculino).
5	Edad	Representa la edad del alumno en este campo se almacena valores numéricos, para nuestro caso está comprendido entre 15 y 56
6	Prom_col	Representa la nota que obtuvo el alumno en la educación secundaria estos valores están comprendidos de 11 a 18
7	Currícula	Representa la currícula a la cual estudio el alumno en este caso puede ser A, B y U
8	Ciclo	Representa en que ciclo se encuentra actualmente el alumno o en que ciclo dejó de estudiar los valores están comprendido de 1 al 10.

Continuación Tabla 4: Descripción de los campos de la vista minable

9	Cantidad de créditos aprobados	Representa la cantidad de créditos que ha probado el alumno hasta el ciclo que se encuentra actualmente y también va a depender de la currícula.(el dato que se almacena es numérico)
10	Cantidad de créditos desaprobados	Representa la cantidad de créditos que ah desaprobado el alumno hasta el ciclo que se encuentra actualmente y también va a depender de la currícula. .(el dato que se almacena es numérico)
11	Cantidad de cursos aprobados	Representa la cantidad de cursos que ha aprobado el alumno hasta el ciclo que se encuentra actualmente y también va a depender de la currícula. .(el dato que se almacena es numérico)
12	Cantidad de cursos desaprobados	Representa la cantidad de cursos que ah desaprobado el alumno hasta el ciclo que se encuentra actualmente y también va a depender de la currícula. . (El dato que se almacena es numérico).
13	Cantidad de tutorías llevadas	En este campo se almacena la cantidad de tutorías que ha llevado el alumno hasta el ciclo que se encuentra. Los valores que se almacena están comprendidos entre 0 y 8.
14	vezcomunica	Representa la cantidad de veces que el alumno llevó el curso de comunicación y también va a depender de la currícula el valor que se almacena será 0, 1,2.
15	PromCom	En este campo se almacena el promedio que el alumno ha obtenido en el curso de comunicación.
16	NNotalog	Representa la nota que obtuvo el alumno en el curso de lógica cuyo valor está comprendido entre 0 y 20.
17	NNotamatuno	Representa la nota que obtuvo el alumno en el curso de matemática uno, cuyo valores está comprendido entre 0 y 20.
18	NNotamatdos	Representa la nota que obtuvo el alumno en el curso de matemática dos, cuyos valores está comprendido entre 0 y 20.
19	NNotamattres	Representa la nota que obtuvo el alumno en el curso de matemática tres, cuyo valores está comprendido entre 0 y 20
20	NNotamatTot	Representa la cantidad de cursos de matemática que ha llevado el alumno los valores van de 1 al 3.
21	PromMat	Representa el promedio que ha obtenido el alumno de los curso de matemática que ha llevado, los valores está comprendido entre 0 y 20.
22	TotalCursos	Representa la cantidad de cursos que ha llevado el alumno y también va a depender de la currícula A,B y U.
23	Nppa	Promedio ponderado de los alumnos hasta el ciclo en que se encuentran y los valores están comprendido entre o y 20.
24	Modalidad	Este campo representa la modalidad en que los alumnos ingresaron en la universidad.
25	NivelIng	Este campo representa el nivel de inglés en que se encuentra el alumno y los valores están comprendidos entre 0 y 6
26	AcIngle	Este campo representa la nota promedio de los cursos de inglés que ha llevado el alumno. el valor está comprendido entre 0 y 20
27	NivelCom	Representa el nivel de computación en la cual se encuentra el alumno y los valores están comprendido entre 1 y 3 y que se refiere a computación 1, computación 2 y computación 3
28	AcCompu	Representa la nota promedio de los cursos de computación y los valores están comprendido entre 0 y 20
29	Categoria	En este campo se almacena la categoría del alumno y esto va depender de su estado socioeconómico
30	ActESTUDIA	En este campo se almacena el estado del alumno si estudia o no estudia los valores que se almacenan son SI , NO, o también pueden ser 1 , 0 debido a que son valores nominales

Fuente: Elaboración propia.

para la verificación de los datos que se han procesado, se hizo uso de la fiabilidad que se encuentra en el paquete estadístico spss stadistic 21, para lo cual se uso el archivo integrado_limpiado_transformados_final_numerico.xls y se exportó al software estadístico como se muestra en la figura 10, en donde nos muestra los campos asi mismo los 1761 datos. luego se realizó el análisis de fiabilidad obteniendo un alfa de crombach de 0,825 como se muestra en la Figura 11, eso nos indicó que los datos que hemos usado son confiables.

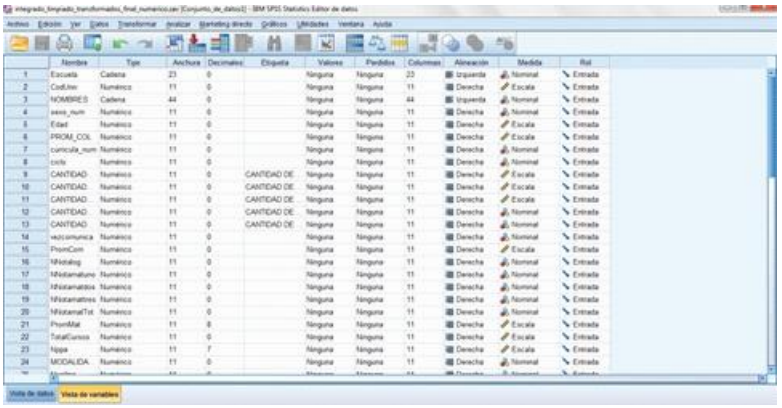

Figura 10: datos exportados a spss stadistic

Fuente: Elaboración propia.

Figura 11: Alfa de crombach

Fuente: Elaboración propia.

Se construyó un modelo con los algoritmos de árboles de decisión (C5.0, CRT, CHAID, QUEST) como se muestra en la figura 12 y así mismo se hizo la validación de los modelos con todos los datos para cada uno de ellos.

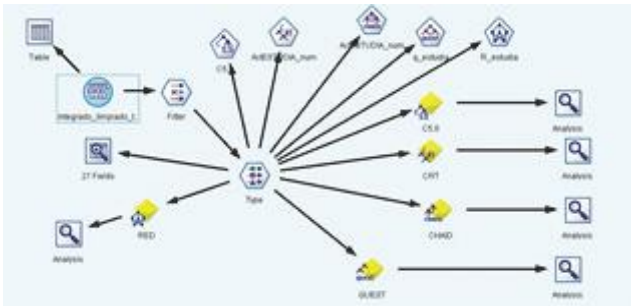

Figura 12: árboles de decisión

Fuente: Elaboración propia.

En relación con los árboles de decisión el que obtuvo mejores resultados fue el algoritmo C5.0 en donde se realizó el entrenamiento con 1761 datos obtenido una precisión de 90,52% como se muestra en la figura 12.1, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 801 y de manera incorrecta 46 es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 793 y de manera incorrecta 121, es decir han sido predichos como alumnos que han desertado.

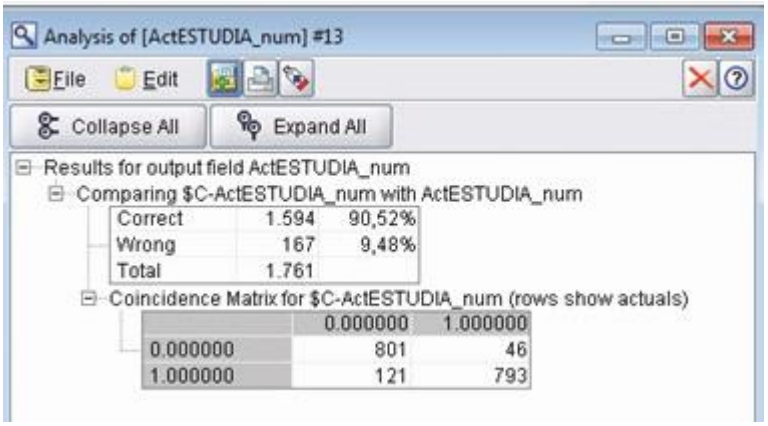

Figura 12.1: Precisión del árbol de decisión C5.0 durante el entrenamiento

Fuente: Elaboración propia.

y luego se realizó la prueba con 100 (Figura 13) datos teniendo una precisión de 89% como se muestra en la figura 14, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 51 y de manera incorrecta dos, es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 38 y de manera incorrecta 9 es decir han sido predichos como alumnos que han desertado.

Figura 13: Modelo del árbol de decisión C5.0 durante la prueba

Fuente: Elaboración propia.

Figura 14: Precisión del árbol de decisión C5.0 durante la prueba

Fuente: Elaboración propia.

En relación con los árboles de decisión también se aplicó el algoritmo CRT en donde se realizó el entrenamiento con 1761 datos obtenido una precisión de 87,45% como se muestra en la figura 15, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 760 y de manera incorrecta 87 es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 780y de manera incorrecta 134, es decir han sido predichos como alumnos que han desertado

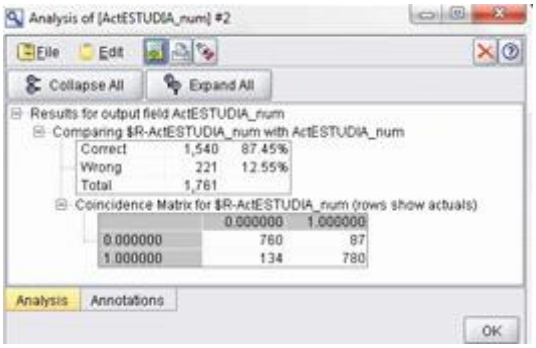

Figura 15: Precisión del árbol de decisión CRT durante el entrenamiento

Fuente: Elaboración propia.

y luego se realizó la prueba con 100 (Figura 16) datos teniendo una precisión de 84% como se muestra en la figura 17, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 47 y de manera incorrecta seis, es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 37 y de manera incorrecta diez es decir han sido predichos como alumnos que han desertado.

Figura 16: Modelo del árbol de decisión CRT durante la prueba

Fuente: Elaboración propia.

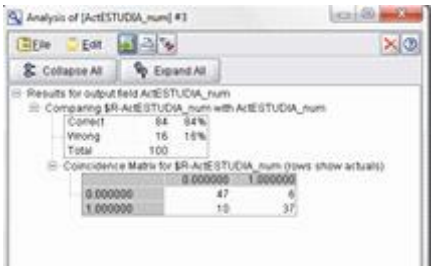

Figura 17: Precisión del árbol de decisión CRT durante la prueba

Fuente: Elaboración propia.

En relación con los árboles de decisión también se aplicó el algoritmo CHAID en donde se realizó el entrenamiento con 1761 datos obtenido una precisión de 87,05% como se muestra en la figura 18, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 757 y de manera incorrecta 90 es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 776 y de manera incorrecta 138, es decir han sido predichos como alumnos que han desertado

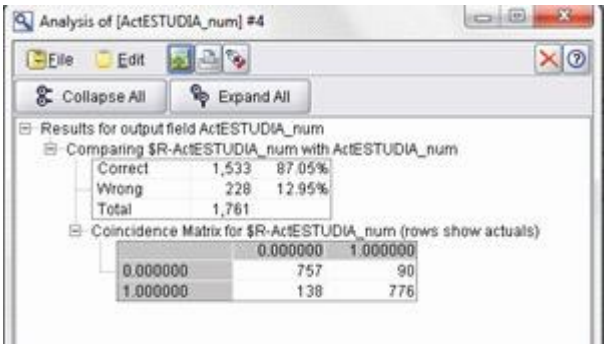

Figura 18: Precisión del árbol de decisión CHAID durante entrenamiento

Fuente: Elaboración propia.

y luego se realizó la prueba con 100 (Figura 19) datos teniendo una precisión de 89% como se muestra en la figura 20, en donde se puede apreciar que de los alumnos desertores han sido predichos de manera correcta 48 y de manera incorrecta cinco, es decir han sido predicho como alumnos que aún siguen estudiando, en relación a los alumnos que si estudian han sido predichos de manera correcta 41 y de manera incorrecta seis es decir han sido predichos como alumnos que han desertado.

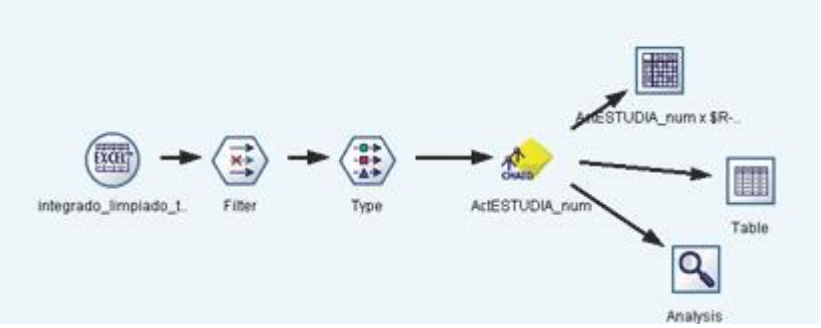

Figura 19: Modelo del árbol de decisión CHAID durante la prueba

Fuente: Elaboración propia.

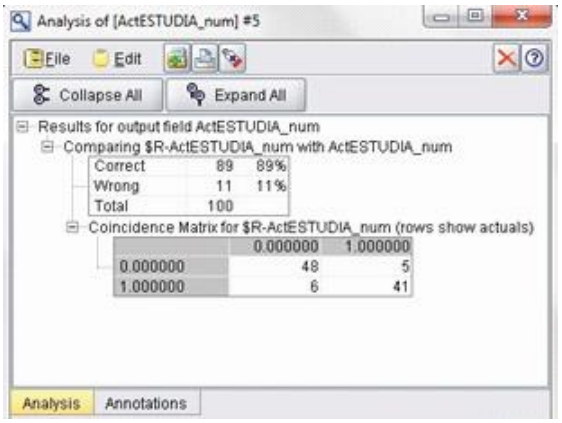

Figura 20: Precisión del árbol de decisión CHAID durante la prueba

Fuente: Elaboración propia.

En la tabla 5 podemos apreciar que el árbol de decisión que nos da mejores resultados es el algoritmo C5.0 con un apreciación de 90.52% con respecto a los demás algoritmos.

Tabla 5 : Comparación de resultados de los algoritmos de minería de datos

Nº	MODELO	DATOS DE ENTRENAMIENTO	PRECISIÓN	DATOS DE PRUEBA	PRECISIÓN
1	Árbol de decisión C5.0	1761	90.52%	100	89%
2	CRT	1761	87.45%	100	84%
3	CHAID	1761	87.05&	100	89%

Fuente: Elaboración propia.

CONCLUSIONES

Se ha realizado la recopilación de los datos de las diferentes áreas involucradas en el trabajo de investigación y luego se realizó el procesado de datos(ETL) así como la limpieza y la selección de los datos, ya que la calidad y fiabilidad de la información afecta de manera directa en los resultados obtenidos de las técnicas de árboles de decisión como para cualquier otro modelo de minería de datos. Se ha demostrado que los algoritmos de clasificación como los árboles de decisión en este caso el algoritmo C5.0, dan buenos resultados para predecir la deserción de un alumno en la Educación Superior, Se pudo demostrar en base a las investigación hechas en otros autores mencionados en el estado del arte, así con las pruebas realizadas en el laboratorio haciendo uso de software académicos como empresariales, que las variables base a tomar en cuenta, o las que aportan más en este tipo de investigación: son código de Universidad, rendimiento académico de la Universidad, rendimiento académico del colegio, edad, sexo y la deserción, pero además de ellos se pudo encontrar otras variables que influyen de manera importante en el trabajo de investigación las cuales son : curricula, ciclo, cantidad de créditos aprobados, cantidad de créditos desaprobados, etc. La idea básica y central fue crear varios modelos de árboles de decisión, que permita analizar y identificar si es necesario mejorar algún asunto del proceso educativo, como por ejemplo el nivel académico, eran las causa del alto índice de deserción en nuestra institución, al proveerle este conocimiento a profesores, fue posible empezar a darle el tratamiento y seguimiento adecuado a cada uno de nuestros alumnos ingresantes y a los que actualmente cursan los primeros tres semestres (Ciclos) en donde se muestra el alto índice de deserción en el estudio hecho.

El establecimiento de políticas encaminadas para el seguimiento y la corrección de situaciones de deserción superior, a partir de las conclusiones de este trabajo no fue un propósito de la investigación. Sin embargo, esperamos que dichas políticas sean establecidas por las autoridades correspondientes de la Universidad Cesar Vallejo- Lima Este, tomando como base a los resultados que ya están viendo como parte del trabajo de investigación.

REFERENCIAS BIBLIOGRÁFICAS

1. Romero C, Ventura S. 'Educational data Mining: A Survey from 1995 to 2005', Expert Systems with Applications. 2007. pp. 135-146.
2. Ruba R. Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks, IEEE Global Engineering Education Conference (EDUCON)–"Learning Environments and Ecosystems in Engineering Education" 2011.
3. Ashutosh T, Adam N. Learning patterns of university student retention, Expert Systems with Applications 38 (2011) 14984–14996.
4. Jadrić M, Garača Z, Ćukušić M. Student Dropout Analysis with Application of data Mining Methods, Management, Vol. 15, 2010, 1, pp. 31-46.
5. Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques, Computers & Education.
6. Dekker G, Pechenizkiy M, Vleeshouwers J. Predicting Students Drop Out: A Case Study, Educational Data Mining 2009.
7. Lin J, Imbrie P, Reid K. Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results, Engineering Education Symposium 2009.
8. Yathongchai W, YAthongchaI C, Kerdprasop K, Kerdprasop N. Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out, Latest Advances in Educational Technologies.
9. Levine J, Wycokoff J. Predicting persistence and success in baccalaurate engineering. Education, 1991. 111(4),461-468.
10. Schaeffers KG, Epperson DL, Nauta, MM. Women's Career Development: Can Theoretically Derived Variables Predict Persistence in Engineering Majors Journal of Counseling Psychology, 1997. V. 44, pp. 173-183.
11. Zhang Z, RiCharde RS. Prediction and Analysis of Freshman Retention. Paper presented at the Annual Forum of the Association for Institutional Research (AIR). 1998.
12. Besterfield-Sacre M, Shuman L, Wolfe H, Scalise A, Larpkiattaworn S, Muogboh OS, et al. Modeling for Educational Enhancement and Assessment. Paper presented at the Annual Conference of American Society for Engineering Education. 2002.
13. French BF, Immekus C, OakesWC. An Examination of Indicators of Engineering Students Success and Persistence. Journal of Engineering Education, .(2005). p.419-425
14. Schaeffers KG, Epperson DL, Nauta MM. Women's Career Development: Can Theoretically Derived Variables Predict Persistence in Engineering Majors Journal of Counseling Psychology, 1997. V.44, pp.173-183.
15. Pascarella ET, Terenzini PT. Predicting Voluntary Freshman Year Persistence/Withdrawal Behaviorina Residential University : A Path Analytic Validation of Tinto's Model. Journal of Educational Psychology, .1983. V.75(2),p.215-226.

16. Fuertes J, Sedlacek W. Using the SAT and Non cognitive Variables to Predict the Grades and Retention of Asian American University Students. Measurement and Evaluation in Counseling & Development, 1994. V.27,p.74-84.
17. Burtner J. The Use of Discriminant Analysis to Investigate the Influence of Non-Cognitive Factors on Engineering School Persistence. Journal of Engineering Education, 2005. July2005.
18. Aitken ND. College Student Performance , Satisfaction and Retention: Specification and Estimation of a Structural Model. Journal of Higher Education. 1982. v53(n1), p32-50.
19. Nora A, Attinasi LC, Matonak A. Testing Qualitative Indicators of Precollege Factors in Tinto's Attrition Model: A Community College Student Population. Review of Higher Education, 1990. V.13(3),P.337.
20. Cabrera A, Nora A, Castañeda M. College Persistence: Structural Equation Modeling Test of an Integrated Model of Student Retention. Journal of Higher Education, 1993.vol. 64, pp. 123-129.
21. French B F, Immekus JC, Oakes W. A structural model of engineering students success and persistence. Paper presented at the Frontiers in Education, 2003
22. Kukar M, Kononenko I,Groselj C,Kralj K, Fettich J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intelligence Medicine, 1999. 16(1),25-50.
23. Coit DW, Jackson BT, Smith AE. Static neural network process models: considerations and case studies. International Journal of Production Research, 1998. 36(11), 2953-2967.
24. Imbrie PK, Lin JJ, Malyscheff A. Artificial Intelligence Methods to Forecast Engineering Students' Retention based on Cognitive and Non-cognitive Factors. Paper presented at the Annual Conference of American Society for Engineering Education, 2008.
25. Gaskins B. A Ten-Year Study of the Conditional Effects on Student Success in the First Year of College, Bowling Green State University, 2009.
26. Lin J. et al., Student Retention Modelling : An Evaluation of Different Methods and their Impact on Prediction Results, in Proc. of the Research in Engineering Education Symposium Palm Cove, QLD, 2009.
27. Nghe N. et al., A comparative analysis of techniques for predicting academic performance, 2007. ,37th ASEE/IEEE Frontiers in Education Conference, octubre 2010
28. Mendez A. et al., Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests, JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-, vol. 97, p. 57, 2008.
29. Ayesha S, Mustafa T, Sattar AR, Khan MI. Data Mining Model for Higher Education System, European Journal of Scientific Research, 2010, Vol.43, No.1, pp.24-29.
30. Sembiring S, Zarlis M, Hartama D, Wani E. Prediction of Student Academic Performance by an Application of Data Mining Techniques. Proceedings of International Conference on Management and Artificial Intelligence, 2011, pp.110-114.
31. Wu X, H. Zhang y H. Zhang, Study of Comprehensive Evaluation Method of Undergraduates Based on Data Mining, Proceedings of International Conference on Intelligent Computing and Integrated Systems, pp 541-543.
32. DesJardins SL, DA. Ahlburg, McCall BP. A temporal investigation of factors related to timely degree completion. J. Higher Education, 2002. vol 73: pag 555-581.
33. Murphy, K and F. Welch, 1993. Inequality and relative wages. Americ. Economic review, Vol 83: 104-109.
34. Murtaugh PA, Burns LD, Schuster J. Predicting the retention of university students. Higher Education, 4: 355-357.
35. Guzmán C, Diana Duran M, Jorge Franco G, Deserción Estudiantil en la Educación Superior Colombiana, 2009 pag (23,27), Bogotá – Colombia.
36. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Usama Fayyad, Gregory Piatetsky - Shapiro , y Padhraic Smyth, COMMUNICATIONS OF THE ACM Vol. 39, No. 11 pag 31
37. Goddard JC et al.," Redes Neuronales y Árboles de Decisión: Un Enfoque Híbrido",Memorias del Simposium Internacional de Computación organizado por el Instituto Politécnico Nacional - November 1995, pp 1-7.
38. Inmon W. (2005). Building the Data Warehouse.(4th Ed).Indianapolis, Indiana: Wiley Publishing.
39. Hernández J., Ferrari C. y Ramírez M. (2004). Introducción a la minería de datos. España: Pearson Educación.
40. Hand J, Kanmber M. Data Minig Concepts and Tecniques, Edit 2006 by Elsevier.
41. González L. Una arquitectura para el análisis de información que integra procesamiento analítico en línea con minería de datos, Maestría en Ciencias con Especialidad en Ingeniería en Sistemas Computacionales. Puebla, México, Universidad de las Américas Puebla, 2005, pp 170.
42. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Usama Fayyad, Gregory Piatetsky - Shapiro, and Padhraic Smyth, COMMUNICATIONS OF THE ACM Vol. 39, No. 11 29-30
43. Ryan SJ, Baker D, Yacef K. The State of Educational Data Mining in 2009: A Review and Future Visions, International Educational Data Minig Society
44. Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert Systems with Application, 33:135–146, July 2007.
45. Romero C, Ventura S. Pechenizkiy, and R. Baker. Handbook of Educational Data Mining. Boca Raton, FL: CRC Press, Taylor y Francis, 2010
46. Sumathi S, Sivanandam S. Introduction to Data Mining and its Applications. Studies in Computational Intelligence, 29, editado por Springer-Verlag, pp. 828, 2006. 3-540-34350-4, Heidelberg, Alemania.
47. Larose D. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley y Sons, Inc., pp. 222, 2005. ISBN: 0-471-66657-2, New Jersey, Estados Unidos.
48. Ho C, DiGangi S, 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 (2010), 307-325.
49. Blanco R. Extracción y contextualización de reglas comprensibles a partir de modelos de "caja negra",Valencia, España, Universidad Politécnica de Valencia 2006, pp. 257
50. Moreno B. "Minería Sobre Grandes Cantidades de Datos" ,México DF, Universidad Autónoma Metropolitana, 2009,166.
51. Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organizationin the Brain. Cornell Aeronautical Laboratory, Psychological Review. Vol.65:386–407. 1958.
52.McClelland J. Rumelhart, D. Learning, Representations by backpropagation. Nature.1986.
53. Jang J, Sun C, Mizutani P. Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence. IEEE TRANSACTIONS ON AUTOMATIC CONTROL, VOL. 42, NO. 10, OCTOBER 1997

54. Rios R, Pineda L. Factores Relacionados con Deserción Temprana en Estudiantes de Medicina, IV Cuarta Conferencia Latinoamericana sobre el abandono en la Educación Superior,2014, pag 1-9
55. Besterfield-Sacre, M, Atman CJ, Shuman LJ. Characteristics of freshman engineering students: Models for determining student attrition in engineering. Journal of Engineering Education, 1997.86(2), 139-149
56. Ryan S, Baker J. Data Mining for Education, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
57. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, 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
58. Spady W. Dropouts from Higher Education: An Interdisciplinary Review and Synthesis. Interchange, (1970). 1, 64-65
59. Tinto V. Colleges as Communities: Taking Research on Student Persistence Seriously. (1998). The Review of Higher Education, 21 (2), 167-177.
60. Tinto V. Limits of Theory and Practice in Student Attrition. Journal of Higher Education, (1982). 53 (6), 687-700.
61. Giovagnoli P. Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración, Documento de Trabajo 37, (2002). Universidad Nacional de la Plata.
62. Castaño E, Gallón S, Gómez K, Vásquez, J. (2004). Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de Economía, 60*, 41-65.
63. Tinto V. Definir la deserción: una cuestión de perspectiva. Revista de Educación Superior Nº 71, 1989. ANUIES, México
64. Tinto V. Principles of Effective Retention. Journal of the Freshmen Year Experience, 1990. 2 (1), 35-48.
65. Bean J. Dropouts and Turnover: The Synthesis and Test of a Casual Model of Student Attrition. Research in Higher Education, 1980. 12, 155-187.
66. Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers. 2nd edition: 560. 2005
67. Tinto V. Dropouts from Higher Education: A Theoretical Synthesis of the Recent Literature. A Review of Educational Research, 1975. 45, 89-125.
68. Cabrera A, Nora A. Castañeda M. Collage Persistence: Structural Equations Modelling Tests of an Integrated Models Student Retention. The Journal of Human Resources, 1993.64, 123-139
69. Porto A, Di Gresia. Rendimiento de estudiantes universitarios y sus determinantes. Asociación Argentina de Economía Política. 2001.
70. DesJardins S, Ahlburg D, McCall B. An Event History Model of Student Departure. Economics of Education Review, 1999. 18, 375-390.
71. Montoya M. Extended Stay at University: An Application of Multinomial Logit and Duration Models. , Applied Economics, 1999. Vol. 31, No. 11, 1411-1422.
72. Araque F, Roldán C, Salguero A. Factors Influencing University Drop Out Rates, Computers & Education, vol. 53, pp. 563–574, 2009.
73. Aleyda R. Factores Relacionadas con la Deserción Estudiantil en el Programa de Enfermería de la Universidad Libre de Pereira, Universidad Libre - Seccional Pereira, 2010, pag 1-10.
74. Garcia R, Minería de Datos en Encuestas de Profesores al fin del Semestre de la Facultad de Ingeniería, Universidad Nacional Autónoma de México, 2011, pag 118.
75. Pérez C, Santin D. Minería de datos: técnicas y herramientas, 2ª ED., España: Thomson Ediciones paraninfo S.A,2008, 775.
76. RapidMiner Studio - RapidMiner, RapidMiner Studio, Disponible en : https://rapidminer.com/products/studio/, Última consulta el 03-02-2015

Recibido: 14 marzo 2016 | **Aceptado:** 19 junio 2016