Modelo neuronal para la estimación del riesgo de deserción en alumnos de grado

Cecilia Balestieri[1], Araceli E. Martin[1] y Cecilia S. Romitti[1]

[1] Facultad de Ingeniería y Ciencias Hídricas, UNL
{ceciliabalestieri, ara.martin.14, csromitti }@gmail.com

Carrera: Ingeniería Informática
Cátedra: Inteligencia Computacional
Profesor: Diego H. Milone

Modelo neuronal para la estimación del riesgo de deserción en alumnos de grado

Resumen. El objetivo de este trabajo es la producción de un modelo que permita estimar el riesgo de deserción en alumnos de grado en la carrera de Ingeniería Informática. Para alcanzar este fin, se propone un método de extracción de características y se emplearon redes neuronales como herramienta de clasificación, en particular se implementó un perceptrón multicapa. La base de datos de alumnos empleada en estos casos se elaboró considerando como fuente la información académica registrada en el Sistema de SIU-Guaraní. Fueron considerados diferentes casos de estudio para evaluar el desempeño del modelo construido. Se analizaron los resultados a partir de los valores del porcentaje de error alcanzado y los indicadores de sensibilidad y especificidad. Se logró un buen nivel de detección de desertores ya que en general el error se mantuvo en un rango entre 7% y 9%, la sensibilidad entre 82% y 92%, y la especificidad superior al 93% para todos los casos.

Palabras clave: Estimación de riesgo, deserción, minería de datos, redes neuronales.

1 Introducción

La deserción estudiantil universitaria es un problema complejo que enfrentan las universidades en el ámbito nacional y mundial. Para combatir este problema se realizan programas de tutorías, talleres, asistencia pedagógica, entre otras actividades, pero aun así no se logra mejorar la situación. La motivación de este trabajo tiene como fin poder brindar una herramienta en respuesta a esta problemática, encontrando un modelo que permita identificar, de manera temprana, a aquellos alumnos en riesgo de abandonar sus estudios de modo que la institución pueda llevar adelante las acciones que considere necesarias.

Sobre esta temática son muchos los análisis que han sido realizados y publicados. Se pueden observar dos líneas de estudio diferenciadas: la primera de ellas apunta a un análisis esencialmente descriptivo de las características de los alumnos desertores, y la segunda hace hincapié en la construcción de un modelo que permita determinar el riesgo de deserción a tiempo. Dentro del primer grupo de publicaciones se encuentra [1], donde los autores focalizan en la presentación de proporciones y cifras basadas en elementos representativos del desempeño académico. Se utiliza la estadística para argumentar pero no concluye en estimaciones precisas que permitan identificar desertores dentro de un intervalo de confianza. Dentro del segundo grupo, [2] evalúa datos académicos en conjunto con aspectos socio-económicos, geográficos, laborales, tanto personales del alumno como de su grupo, para identificar perfiles de deserción y poder predecir qué estudiantes son los que abandonan sus estudios.

Para el segundo tipo de análisis las técnicas empleadas son diversas pero la minería de datos (MD) suele ser el método más frecuente, en particular con herramientas de

predicción como árboles de decisión [2], y regresión logística [3]. En [4] y [5] se han realizado comparaciones en respecto a los resultados de estas dos herramientas junto a los de redes neuronales (RN). Respeto a las RN, en [6] y [7] se utilizan perceptrones multicapa (MLP, por sus siglas en inglés) para predecir el éxito o fracaso de estudiantes. Finalmente cabe mencionar investigaciones donde se ha implementado la combinación de técnicas de aprendizaje de máquina (como RN, máquinas de vectores de soporte (SVM), fuzzy ARTMAP) con esquemas de decisión [8].

El enfoque de este trabajo es, en primer lugar, realizar un análisis de cuáles son datos referidos al desempeño académico que pueden ser indicadores de la deserción estudiantil. Luego de esta etapa de selección de variables y extracción de características, se entrena un modelo MLP para estimar el riesgo de que cada alumno deserte.

En la siguiente sección se menciona el procedimiento de recolección y selección de los datos con los que se genera la base de datos (BD), y se describe el proceso de MD basado en la construcción de un MLP que satisfaga las necesidades de este problema, así como también los indicadores que evaluarán el desempeño del modelo. En la Sección III se presentan los resultados numéricos y comparativos, alcanzados mediante cuatro casos de estudio y sus correspondientes ensayos. Finalmente en la Sección IV, se exponen las conclusiones obtenidas del trabajo realizado junto con propuestas para trabajos futuros.

2 Materiales y Métodos

En el desarrollo de este trabajo se realizó un estudio específico de la situación en los estudiantes de la carrera de Ingeniería en Informática (II), ya que es ésta la carrera de la Facultad que cuenta con el mayor índice de desertores. A fin de obtener los datos necesarios y representativos de esta población, se generó una BD a partir del acceso y relevamiento de la información que se registra en el Sistema de Gestión Académica SIU-Guaraní. En particular, dentro del Sistema de Análisis O3Portal [9] que se crea a través del módulo de generación de cubos[1], se utilizó el Cubo 02 que corresponde al "Rendimiento Académico". Cabe aclarar que podrían utilizarse otros Cubos para obtener datos más específicos respecto a características socio-económicas geográficas y laborales, pero el presente trabajo debió limitarse sólo al Cubo 02 ya que en los demás la cantidad de datos incompletos es extremadamente alta.

Para poder continuar con el análisis de los datos relevados es necesario considerar los siguientes conceptos que se emplearán a lo largo de este informe:

- Cohorte: grupo de alumnos que inician al mismo tiempo sus estudios en un programa educativo, es decir, quienes se inscriben e ingresan en un mismo año (la misma generación).
- Criterio de deserción: se considera que un alumno ha desertado cuando, en los dos últimos años del período bajo análisis, no se han registrado asignaturas cursadas ni exámenes finales rendidos.

2.1 Recopilación y Selección de Datos

[1] Cubos: Son estructuras multidimensionales que contienen datos resumidos de grandes bases de datos o sistemas transaccionales.

Fig. 1 Datos de Entrada y Referencia Cohorte 2006

Los datos que se extrajeron corresponden a la cohorte 2006 dentro del período 2006-2010, y a la cohorte 2007 dentro del período 2007-2011. Este relevamiento permitió identificar 235 estudiantes para la cohorte 2006 y 255 para la cohorte 2007, con los siguientes atributos:

- Nro. Legajo
- Carrera – Sede
- Cohorte
- Año académico
- Intentos de cursadas y resultado
- Intentos de rendir cada asignatura
- Equivalencias otorgadas
- Asignaturas rendidas y calificación obtenida

Luego, con el propósito de mejorar la calidad de los datos recolectados se eliminaron los que refieren a estudiantes que se encuentran de intercambio en esta Facultad. De esta manera el registro de alumnos para la cohorte 2006 se reduce a 232 y a 254 para la cohorte 2007. Posteriormente se completaron los atributos de aquellos estudiantes que, luego de matricularse en la carrera, no presentan ningún registro de actividad en todo el período evaluado pero que deben ser considerados.

2.2 Identificación de Atributos y Etiquetado

El siguiente paso en la construcción de la BD fue buscar aquellos atributos del alumno que se consideran relevantes para la problemática bajo análisis, eliminando los que no son necesarios. Los atributos seleccionados como característicos del desempeño académico se recopilaron para cada año del período evaluado y se listan a continuación:

- Cantidad de cursadas positivas: cantidad de asignaturas cursadas cuyo resultado es regular o promovido.
- Cantidad de rendidas: cantidad de exámenes finales a los que el alumno se presentó (aprobados o no aprobados), es decir que se exceptúan exámenes finales cuyo resultado es "ausente".
- Avance en la carrera: porcentaje de asignaturas aprobadas respecto al total. En este caso, la carrera de II tiene un total de 42 asignaturas, considerando las correspondientes al Plan de estudio y los requisitos de asignaturas Optativas y Electivas.
- Promedio académico: media aritmética de todas las calificaciones obtenidas en exámenes finales (o por equivalencia), considerando aplazos.

Contando con todos estos datos disponibles, se procedió al etiquetado de los alumnos registrados según el criterio de deserción establecido. Para ambas cohortes se

Tabla 1 Datos de Entrada y Referencia Cohorte 2006

Cursadas positivas			Cantidad de rendidas			Avance en la carrera			Promedio académico			Salida
2006	2007	2008	2006	2007	2008	2006	2007	2008	2006	2007	2008	Estado
Nro.	Nro.	Nro.	Nro.	Nro.	Nro.	%	%	%	Nro.	Nro.	Nro.	-1, 1

realiza el mismo procedimiento, considerando los dos últimos años para determinar si un alumno desertó. Por ejemplo, para la cohorte 2006 se tomaron como referencia los datos de la actividad académica del período 2009-2010 (Fig. 1), y se etiqueta a los patrones con "-1" si identificó un desertor o "1" en caso contrario. De esta forma, los datos de entrenamiento quedaron conformados por los atributos correspondientes al período 2006-2008 y la etiqueta de deserción obtenida en base al período 2009-2010. En la Tabla 1 se puede observar un ejemplo simplificado de esta estructura para la cohorte 2006.

2.3 Construcción de la Red

Esta fase tiene como propósito la construcción del modelo predictivo en sí utilizando como herramienta de clasificación un MLP. El entrenamiento se realizó con el algoritmo supervisado de backpropagation, aplicando la regla de aprendizaje "error-corrección" [10]. Los indicadores antes mencionados tendrán influencia en la elección de una estructura adecuada para este MPL ya que, se pretende se pretende mantener lograr un bajo error, lo cual implica minimizar los FN, pero a su vez mantener un bajo nivel de complejidad para maximizar la capacidad de generalización. La complejidad de la red dependerá de la cantidad de neuronas en su capa oculta y el número de interconexiones que contiene, es por esto que se emplearon diferentes variaciones, alterando la topología del MLP, tanto en cantidad de capas ocultas, como de neuronas por capa, y realizando evaluaciones de validación cruzada sólo con los datos pertenecientes a la cohorte 2006.

2.4 Entrenamiento y Prueba

Para entrenar la red se emplearon los patrones correspondientes a la cohorte 2006, con los datos referidos al período 2006-2008, es decir que se trabaja con 3 años de desempeño académico para cada estudiante. A su vez, se realizaron 10 particiones de este conjunto de entrenamiento, donde cada una de éstas contiene una selección aleatoria del 80% de los patrones (185 patrones por partición). Se consideró esta estrategia de fraccionado sobre los patrones de entrenamiento ya que en ocasiones permite conseguir mejor capacidad de generalización en comparación a un entrenamiento con todos los patrones. Además en todos los ensayos se se probaron varias tasas de aprendizaje y épocas de entrenamiento y se obtuvieron mejores resultados con $\eta = 0.01$ y 50 épocas de entrenamiento. Con esta cantidad de épocas el algoritmo de retropropagación alcanza un error cuadrático medio (ECM) que da lugar a predicciones con una tasa de acierto satisfactoria, y al incrementar el número de épocas aunque se logre un ECM menor, éste no mejora la tasa de aciertos sobre el 20% de datos no utilizados para el entrenamiento.

Para la etapa de pruebas, se tomaron los patrones correspondientes a la cohorte 2007, cuyos datos refieren al período 2007-2009, de igual manera que con los datos de entrenamiento, se cuenta con la información de 3 años académicos por alumno. El uso de la cohorte 2007 para realizar la etapa de test, responde a la necesidad de asegurar la capacidad de generalización por parte del estimador, ya que este modelo tiene como fin evaluar alumnos que pertenecen a otras cohortes y no a la población específica con la que se entrena (cohorte 2006) (Fig. 2). Se pretende entonces, que éste arroje un resultado que siga siendo válido independientemente del grupo de alumnos a los que se aplique el estimador y la cohorte a la que pertenecen los datos. Así, se realizaron 10 entrenamientos con las particiones de 80% de la cohorte 2006 y se probó en todos los casos con el 100% de los patrones de la cohorte 2007. Cada una de estas pruebas pone a disposición las medidas e indicadores que nos permiten realizar el análisis de los resultados.

2.5 Indicadores y Medidas de Validación

En la evaluación del desempeño del modelo predictor se tendrán en cuenta medidas e indicadores que determinen su capacidad de estimación y la bondad de los resultados obtenidos. Dentro de estas medidas se considerarán:

- Tasa de error (e): porcentaje de predicciones incorrectas realizadas por la red.
- Verdaderos Positivos (VP): cantidad de alumnos para los cuales la red determinó que desertaron, y así fue.
- Verdaderos Negativos (VN): cantidad de alumnos para los cuales la red determinó que no desertaron, y así fue.
- Falsos Positivos (FP): cantidad de alumnos para los cuales la red determinó desertaron, cuando en la realidad no sucedió.
- Falsos Negativos (FN): cantidad de alumnos para los cuales la red determinó que no desertaron, cuando en la realidad sí lo hicieron.

Siendo el objetivo de los procesos de diagnóstico o predicción, como el que se trata en este informe, el de minimizar la cantidad de FN, es posible emplear estas medidas en el cálculo de la capacidad de la red para clasificar correctamente a los alumnos que desertan y a los que no. Esto se consigue mediante el uso de dos indicadores [11]:

- Sensibilidad: es la probabilidad de clasificar correctamente a un estudiante "desertor", es decir, la capacidad del test para detectar la deserción:

$$S = \frac{VP}{VP + FN}$$

Esta medida varía de 0 a 1 [0 a 100%]. Cuanto más alto el valor numérico, hay mejor capacidad de detectar la deserción.

- Especificidad: es la probabilidad de clasificar correctamente a un estudiante "no desertor", es decir, la capacidad de test para detectar a los que no abandonan la carrera:

$$E = \frac{VN}{VN + FP}$$

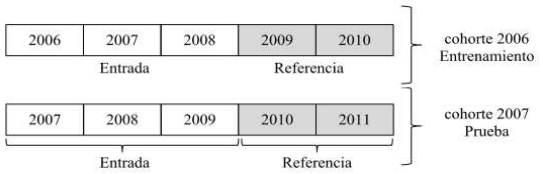

Fig. 2 Patrones de Entrenamiento y Patrones de Prueba

Esta medida varía de 0 a 1 [0 a 100%]. Cuanto más alto el valor numérico, hay mejor capacidad de detectar a dichos alumnos.

3 Resultados

Se consideraron cuatro casos que abarcan distintas selecciones de atributos según criterios establecidos y realizando variaciones en la estructura de la red. En la Tabla 2 se presentan los resultados para cada uno de los casos y sus correspondientes particiones, junto con el promedio de las variables consideradas: error, sensibilidad y especificidad. Como el objetivo es minimizar la cantidad de FN se resaltan las particiones, para cada caso, donde se obtuvieron altos valores en el indicador de Sensibilidad, y el mejor caso respecto al promedio de todas las variables consideradas.

Inicialmente se consideró trabajar con una estructura de 1 capa oculta con n neuronas, y 1 capa de salida con 1 neurona. Para esto se asignaron diferentes valores a n, comenzando por 2 e incrementando en una unidad dicho valor en cada ensayo. Se concluyó que el incremento de neuronas en la capa oculta no producía efectos considerables en los resultados cuando n pertenece a $[2,5]$, ya que la tasa de error arrojada por las pruebas no presentaba diferencias significativas. Y cuando $n > 5$, el valor de sensibilidad comienza a disminuir proporcionalmente al aumento del porcentaje de error, sin mencionar el incremento en la complejidad de la red y el costo computacional. Luego se consideró implementar una estructura de 2 capas ocultas con n y m neuronas respectivamente, y 1 capa de salida, a fin de obtener un mejor rendimiento de la red, pero los distintos ensayos efectuados variando los valores de n y m, no mostraron mejoras en tasa de error, ni en la medida de sensibilidad, por lo tanto se descartó esta configuración topológica.

3.1 Casos de Estudio:

Caso 1. En primer lugar se trabajó con los atributos presentados en la Tabla 1, donde 12 variables son representativas de los atributos académicos y 1 variable representa la salida deseada, haciendo un total de 13 valores para cada patrón. Como vemos, en este caso se consideraron todos los datos desplegados año a año en el período considerado para cada cohorte.

Caso 2. En este caso se pensó como alternativa la selección de los atributos referidos a Cantidad de rendidas, Avance en la carrera, y Promedio académico, desplegados año a año, sin considerar las Cursadas positivas como un factor decisivo a para la deserción. De este modo la cantidad de atributos por patrón es de 10 en total.

Tabla 2 Resultados y promedios de los casos de estudio

		PARTICIONES										
		P 1	**P2**	**P3**	**P 4**	**P 5**	**P 6**	**P 7**	**P 8**	**P 9**	**P 10**	**PROMEDIO**
CASO 1	**Error**	8,70%	7,11%	8,70%	6,72%	6,32%	6,72%	6,72%	6,72%	6,32%	6,32%	**7,04%**
	Sensibilidad	90,20%	92,45%	91,84%	92,59%	94,34%	92,59%	92,59%	89,66%	92,73%	94,34%	**92,33%**
	Especificidad	91,58%	93,00%	91,18%	93,47%	93,50%	93,47%	93,47%	94,36%	93,94%	93,50%	**93,15%**
CASO 2	**Error**	6,32%	5,93%	7,91%	7,51%	7,11%	6,72%	7,11%	7,51%	7,91%	7,11%	**7,11%**
	Sensibilidad	94,34%	94,44%	89,09%	90,74%	92,45%	94,23%	92,45%	90,74%	89,09%	92,45%	**92,00%**
	Especificidad	93,50%	93,97%	92,93%	92,96%	93,00%	93,03%	93,00%	92,96%	92,93%	93,00%	**93,13%**
CASO 3	**Error**	9,09%	7,51%	9,09%	8,30%	7,91%	9,49%	7,91%	9,88%	10,28%	10,67%	**9,01%**
	Sensibilidad	85,71%	90,74%	85,71%	87,50%	89,09%	83,05%	89,09%	81,67%	80,33%	82,14%	**85,50%**
	Especificidad	92,39%	92,96%	92,39%	92,89%	92,93%	92,78%	92,93%	92,75%	92,71%	91,37%	**92,61%**
CASO 4	**Error**	9,88%	10,67%	9,09%	9,49%	9,09%	9,88%	9,88%	9,09%	9,88%	9,88%	**9,68%**
	Sensibilidad	80,65%	79,03%	83,33%	80,95%	83,33%	80,65%	80,65%	83,33%	79,69%	80,65%	**81,23%**
	Especificidad	93,19%	92,67%	93,26%	93,68%	93,26%	93,19%	93,19%	93,26%	93,65%	93,19%	**93,26%**

Caso 3. Se estableció como criterio considerar atributos globales y el desglose de una sola de las variables. De este modo se obtuvieron valores referidos a Totales de Cursadas positivas y rendidas, Promedio académico resultante del período bajo análisis y un despliegue del Avance en la carrera año a año. Como consecuencia se cuenta con 7 atributos que componen el patrón.
Caso 4. Para este caso se creyó conveniente analizar los resultados que se consiguen concentrando todos los datos acumulados en el período bajo análisis, es decir: Total de cursadas positivas, Total de rendidas, Avance total en la carrera y Promedio académico del período. La cantidad de valores por patrón se redujo a 5.

En general se puede observar que el valor de especificidad se encuentra estable en un rango entre 90% y 93% para todos los casos, lo cual permite considerar que el estimador posee una buena capacidad para detectar estudiantes que no desertan. Además, los valores de la Tabla 2 permiten notar que la calidad de los resultados obtenidos también se relaciona con la elección acertada de los parámetros de entrada a la red, que deben ser lo suficientemente representativos del problema que se está tratando, pero sin incluir información redundante o irrelevante. Respecto a esto, en cada uno de los casos analizados se observa que se obtuvieron mejores resultados cuando se dispuso de todos los atributos detallándolos para cada año del período (Caso 1), e incluso, eliminando una de las variables, pero considerando el resto desplegado (Caso 2). Fueron estos casos en dónde se lograron los mejores porcentajes de sensibilidad, con un promedio en torno al 93%, en correspondencia con bajos porcentajes de error, entre el 6% y 7%. En general en estos casos se mantuvo una topología de red sencilla donde se observa un nivel de complejidad reducido. Finalmente se puede concluir que el modelo para estos casos tiene un buen

desempeño en la predicción de la deserción de un alumno y una verificada capacidad de generalización.

Al reducir la cantidad de variables de entrada condensando los atributos (Caso 4), la tasa de error supera el 9% para todas las particiones de entrenamiento, y la Sensibilidad alcanza en promedio el 81%, si bien este indicador ronda el 80% en todas las particiones, es menor a lo que pudo a obtenerse en los demás casos. De todos modos, es un valor aceptable, y dada una situación particular en la que no se tiene disponibilidad de datos detallados del desempeño académico del alumno, podría emplearse este caso como último recurso. Cabe destacar que cuando se trabajó en conjunto con atributos condensados por período y atributos descriptos año por año (Caso 3), los resultados obtenidos, en cuanto a la Sensibilidad, pueden considerarse como una situación intermedia entre las dos mencionadas anteriormente ya que este indicador alcanza un promedio de 85%, pero el nivel de error se mantiene superior al 7%.

4 Conclusiones

En este trabajo se propusieron varias alternativas para la extracción de características y un modelo neuronal para la estimación del riesgo de deserción de alumnos en carreras de grado. Los resultados permiten concluir que el trabajo cumple con los objetivos planteados, ya que la red MLP propuesta funciona como estimador del riesgo de deserción en alumnos con un porcentaje de acierto satisfactorio, y consiguiendo buenos valores del indicador de sensibilidad. El impacto de la implementación de esta herramienta en la universidad permitirá identificar de manera rápida y efectiva a la población en riesgo de manera de poder tener un contacto personal con cada uno de los alumnos en esta situación y tomar las medidas acordes a cada caso particular a fin de evitar que el alumno deserte.

Como trabajo futuro se podría considerar la incorporación de nuevos atributos referidos a aspectos socio-económicos, geográficos y laborales, que al momento de realizar esta investigación se encontraban incompletos en el Sistema de Análisis O3Portal. En una segunda etapa sería interesante utilizar algoritmos no supervisados de minería de datos para descubrir conocimiento respecto a las características propias de la población que ha desertado, esto permitiría tomar decisiones generales para prevención y, lograr, a largo plazo, que ésta disminuya. Para este fin se deberían considerar los nuevos atributos señalados anteriormente y entrenar mapas auto-organizativos que permitan visualizar el comportamiento y las tendencias comunes del grupo bajo estudio.

Referencias

1. Porto Alberto, Di Gresia Luciano: “Rendimiento de estudiantes universitarios y sus determinantes”, Asociación Argentina de Economía Política, (2001).
2. Timarán Pereira Ricardo: “Una Lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos”, Revista Científica Guillermo de Ockham, vol. 8, núm. 1, pp. 121-130, (2010).

3. Ponsot B. Ernesto, Varela Leonardo, Sinha Surendra, Valera Jorge: "Un modelo de regresión logística del rendimiento en los estudios universitarios: Caso FACES-ULA", Actualidad Contable Faces, Vol. 12, Núm. 18, pp. 81-102, (2009).
4. M. Jadric, Z. Garaca, and M. Cukusic: "Student Dropout Analysis with Application of Data Mining Methods", Management, Vol. 15, No. 1, pp. 31-46, (2010).
5. Porcel Eduardo, Dapozo Gladys, López María V: "Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios", XI Workshop de Investigadores en Ciencias de la Computación, p. 635-639, (2009).
6. Pinninghoff M. A., Salcedo P., Contreras R.: "Neural Networks to Predict Schooling Failure/Sucess", Lecture Notes Computer Science. Vol. 4528, (2007).
7. Stamos T. Karamouzis and Andreas Vrettos: "An Artificial Neural Network for Predicting Student Graduation Outcomes", Proceedings of the World Congress on Engineering and Computer Science, (2008).
8. Lykourentzou, Ioanna et al: Dropout prediction in e-learning courses through the combination of machine learning techniques - Computers & Education Volume 53, Issue 3, Pages 950–965, (2009).
9. https://www.ideasoft.biz/wiki/display/o3man/O3+Portal
10. Haykin Simon, Neural Networks: A Comprehensive Foundation, Macmillan College Publishing Company (1994).
11. Wen Zhu, Nancy Zeng, Ning Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", Health Care and Life Sciences, (2010).