DETECCIÓN DE PATRONES DE LOS ALUMNOS DE PRE-GRADO DESAPROBADOS EN EL CURSO DE ESTADÍSTICA GENERAL DE LA UNIVERSIDAD NACIONAL AGRARIA LA MOLINA USANDO TÉCNICAS DE MINERÍA DE DATOS

Salinas Flores, Jesús Walter
jsalinas@lamolina.edu.pe
Universidad Nacional Agraria La Molina (Perú)

RESUMEN

En los últimos semestres el número de estudiantes desaprobados en el curso Estadística General ha correspondido a un 41%. Por ello, en el presente estudio se planteó la hipótesis de que existe dependencia entre el rendimiento académico (aprobado y desaprobado) de los alumnos con las variables socio-demográficas y académicas de dichos alumnos, y que tal dependencia puede expresarse a través de un modelo estadístico. Usando las técnicas estadísticas de minería de datos se estudiaron a los alumnos de pre-grado de la Universidad Nacional Agraria La Molina, que hayan llevado el curso durante tres semestres académicos con un aproximado de 1500 alumnos, y se encontraron las principales variables socio-demográficas y académicas que determinan la situación del rendimiento académico (aprobado y desaprobado). Usando esta información se puede predecir la situación final del alumno (aprobado o desaprobado) apenas el alumno se matricule en el curso sin haber rendido ningún tipo de evaluación.

PALABRAS CLAVE

Rendimiento académico, Modelos predictivos, Técnicas de minería de datos.

INTRODUCCIÓN

El curso de Estadística General es un curso de estudios generales que debe ser cursado por todos los estudiantes de pre-grado de la Universidad Nacional Agraria La Molina (UNALM). Semestralmente se ofrecen de nueve a diez grupos del curso con un total de 60 vacantes por grupos.

En los últimos ocho ciclos académicos, el número de desaprobados en el curso ha fluctuado entre el 34% y el 49% tal como puede apreciarse en el Gráfico 1, con un promedio de 41% de desaprobación lo cual es equivalente a tener aproximadamente cuatro grupos programados solo para alumnos repitentes.

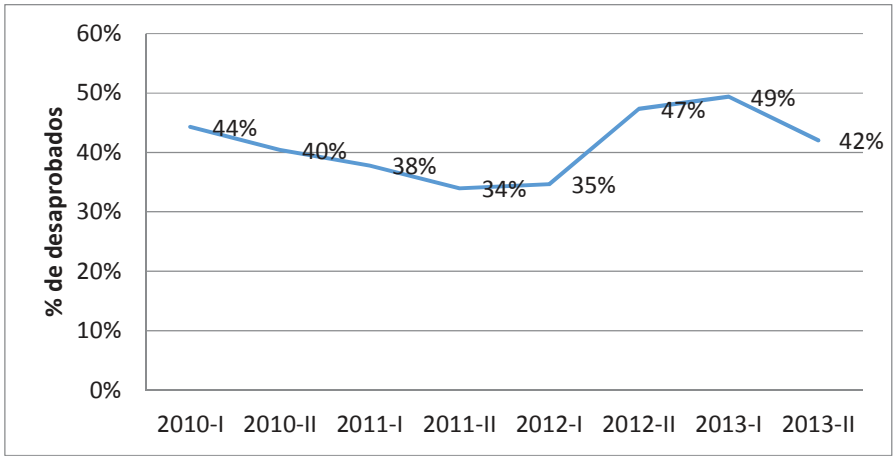

Gráfico 1. Alumnos desaprobados en el curso de Estadística General en la UNALM por semestre
Fuente: Elaboración propia

Los alumnos que están en riesgo de ser desaprobados no reciben algún tipo de asesoría especializada debido a muchos factores, sin embargo, es necesario reconocer que sería mucho más expeditivo brindar algún tipo de asesoría al momento de iniciar el semestre si se pudiera reconocer el patrón de los alumnos que tendrían mayor probabilidad de ser desaprobados para que la asesoría o cualquier otro tipo de ayuda sea más focalizada.

Por ello el objetivo de la presente investigación es reconocer el patrón de los alumnos que resultan desaprobados o aprobados en el curso de Estadística General usando variables socio-demográficas y académicas. Usando este patrón y un modelo estadístico se podría detectar al iniciar un semestre académico aquellos alumnos con mayor probabilidad de ser desaprobados a fin de poder brindarles algún tipo de ayuda o asesoría especializada.

MARCO DE REFERENCIA

Guzmán (2012) usa el análisis de regresión múltiple de toda la generación y en forma específica de cada una de las carreras, con la finalidad de pronosticar el rendimiento final a partir de las variables clasificatorias de rendimiento inicial y rendimiento durante la carrera universitaria. Cobo, Rocha y Álvarez (2011) aplicaron técnicas de minería de datos para identificar aquellos indicadores que pudieran tener mayor valor predictivo, a la hora de medir el rendimiento de los estudiantes en el contexto de una asignatura de grado que combina actividades docentes presenciales con actividades soportadas en aplicaciones de teleformación.

Celis, Moreno, Poblete, Villanueva y Weber (2015) hacen uso de herramientas de *learning analytics* para construir un modelo que predice la caída en causal de eliminación, por motivos académicos, en estudiantes de primer año del Plan Común de Ingeniería y Ciencias de la Universidad de Chile. El modelo clasifica correctamente a más del 86% de los casos, con niveles bajos de error tipo II y una precisión del 38%. El modelo permite desarrollar intervenciones focalizadas en aquellos estudiantes en mayor riesgo.

Vialardi et al. (2011) presentaron la justificación detrás del diseño de un sistema de recomendación para apoyar el proceso de inscripción mediante registro de rendimiento académico de los estudiantes. Para construir este sistema, la metodología CRISP-DM se aplicó a los datos de estudiantes del Departamento de Ciencias de la Computación de la Universidad de Lima, Perú. Los datos fueron modelados usando C4.5, KNN, Naive Bayes, Bagging y Boosting y un conjunto de experimentos fue desarrollado obteniendo que el Bagging es el mejor método con respecto a la exactitud de predicción.

ASPECTOS METODOLÓGICOS

Se seguirá la metodología CRISP (*Cross-Industry Standard Process for Data Mining*) el cual es una metodología probada para trabajos de minería de datos e incluye seis fases que pueden apreciarse en la Figura 1 y que comprende: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelación, evaluación y despliege de resultados.

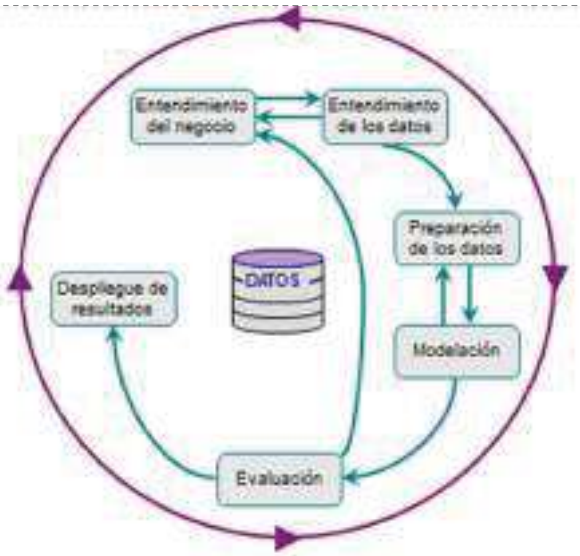

Figura 1. Fases de la Metodología CRISP
Fuente: IBM Corporation, 2012

POBLACIÓN EN ESTUDIO Y METODOLOGÍA ESTADÍSTICA

Se analizaron los datos de una muestra de alumnos de la UNALM de pre-grado que hayan llevado el curso de Estadística General durante los tres ciclos académicos 2013-II, 2014-I y 2014-II con un aproximado de 1500 alumnos. Para la construcción del modelo se usaron los datos obtenidos de una muestra de los ciclos académicos 2013-II y 2014-I. Para validar el modelo se usaron los datos obtenidos de una muestra de alumnos del ciclo académico 2014-II.

Se realizó un análisis descriptivo univariado y bivariado con las variables predictoras y la variable dependiente a predecir (rendimiento académico). Para la etapa de modelamiento se

usaron los modelos de regresión logística y el algoritmo de árbol de clasificación CART. Se usó la validación cruzada 10-folds para la estimación y prueba de los modelos.

DESARROLLO

SELECCIÓN DE VARIABLES PREDICTORAS

Como resultado de las etapas de comprensión del negocio y de los datos se seleccionaron las siguientes variables como principales predictoras del rendimiento académico de un alumno (aprobado/desaprobado) en el curso Estadística General.

Variable	Descripción de la variable	Tipo de variable	Observación
Especialidad	Carrera que está cursando el alumno	Categórica	Datos Generales
Edad	Edad del alumno en años	Cuantitativa	Datos Generales
Sexo	Sexo del alumno	Categórica	Datos Generales
Monto	Aporte semestral del alumno	Cuantitativa	Datos Generales
Modalidad de Ingreso	Modalidad de ingreso a la universidad	Categórica	Datos Generales
Satisfacción	Está estudiando la carrera que eligió como primera opción	Categórica	Datos Generales
Veces_Estadística	Número de veces que está cursando el curso Estadística General	Cuantitativa	Datos antes de iniciar el semestre
Promedio Semestral	Promedio Semestral del último semestre	Cuantitativa	Datos antes de iniciar el semestre
Ratio_Desempeño	Ratio (Número de Créditos Aprobados Acumulados)/(Numero de Créditos Cursados Acumulados)	Cuantitativa	Datos antes de iniciar el semestre
Situación	Situación académica del alumno al iniciar el semestre	Categórica	Datos antes de iniciar el semestre
Retiros	Número de retiros de semestre	Cuantitativa	Datos antes de iniciar el semestre
Veces_Matemática	Número de veces que está cursando el curso Matemática Básica	Cuantitativa	Datos antes de iniciar el semestre
Nota_Matemática	Nota con que aprobó el curso Matemática Básica	Cuantitativa	Datos antes de iniciar el semestre
Veces_Diferencial	Número de veces que está cursando el curso Cálculo Diferencial	Cuantitativa	Datos antes de iniciar el semestre
Nota_Diferencial	Nota con que aprobó el curso Cálculo Diferencial	Cuantitativa	Datos antes de iniciar el semestre
Integral	Situación de aprobación o desaprobación del curso Cálculo Integral antes de llevar el curso Estadística General	Categórica	Datos antes de iniciar el semestre

Tabla 1. Descripción de las variables predictoras
Fuente: Elaboración propia

ANÁLISIS DESCRIPTIVO UNIVARIADO DE LAS VARIABLES PREDICTORAS

- La edad promedio de los matriculados es de 21.5 años.
- El 52% de los matriculados son del sexo masculino.
- El aporte semestral promedio de los matriculados es de S/. 174.2.

- El 44% de los matriculados en el curso están cursando una carrera que no fue su primera opción cuando postularon.
- El número promedio de veces que lleva el curso es de 1.52.
- El promedio ponderado del último semestre es de 13.24
- El ratio de desempeño promedio es de 0.78. Esto es, que de 100 créditos cursados los alumnos han aprobado en promedio 78 créditos.
- El 36% de los matriculados en el curso están en situación no normal.
- El número promedio de veces que llevó el curso Matemática Básica es 1.31.
- La nota promedio con que aprobó el curso Matemática Básica es 13.3.
- El número promedio de veces que llevó el curso Cálculo Diferencial es 1.46.
- La nota promedio con que aprobó el curso Cálculo Diferencial es 12.97.
- El 44% de los matriculados han aprobado el curso Cálculo Integral antes de llevar el curso Estadística General.

ANÁLISIS DESCRIPTIVO BIVARIADO DE LAS VARIABLES PREDICTORAS CON LA VARIABLE DEPENDIENTE

- Las carreras de Biología, Ambiental, Industrias superan el 70% de aprobación en el curso.
- La edad promedio de los desaprobados (22.98 años) supera en más de dos años a la edad promedio de los aprobados (20.93 años).
- El % de aprobación de las mujeres (63%) supera en 7 puntos al % de aprobación de los hombres (56%).
- El número de veces promedio que lleva el curso Estadística General un alumno aprobado es 1.41 y un alumno desaprobado es 1.69.
- El promedio ponderado del último semestre de los aprobados (13.79) es mayor que el de los desaprobados (12.45).
- El ratio de desempeño de los alumnos aprobados (0.84) es mucho mayor que el de los desaprobados (0.69).
- El 60% de los alumnos en situación académica NORMAL aprueban el curso frente a un 24% de aprobación de los alumnos que tienen antecedente de SUSPENSIÓN.
- El número promedio de veces que llevó el curso Matemática Básica los aprobados (1.18) es menor que el de los desaprobados (1.49).La nota promedio con que aprobó el curso Matemática Básica los aprobados (13.68) es mayor que el de los desaprobados (12.74).
- El número promedio de veces que llevó el curso Cálculo Diferencial los aprobados (1.28) es menor que el de los desaprobados (1.71)
- La nota promedio con que aprobó el curso Cálculo Diferencial los aprobados (13.35) es mayor que el de los desaprobados (12.43)

MODELAMIENTO

Se encontró el siguiente árbol de clasificación donde las variables más importantes son: Promedio Semestral, Ratio de Desempeño y Situación de Cálculo Integral tal como puede apreciarse en la Figura 2.

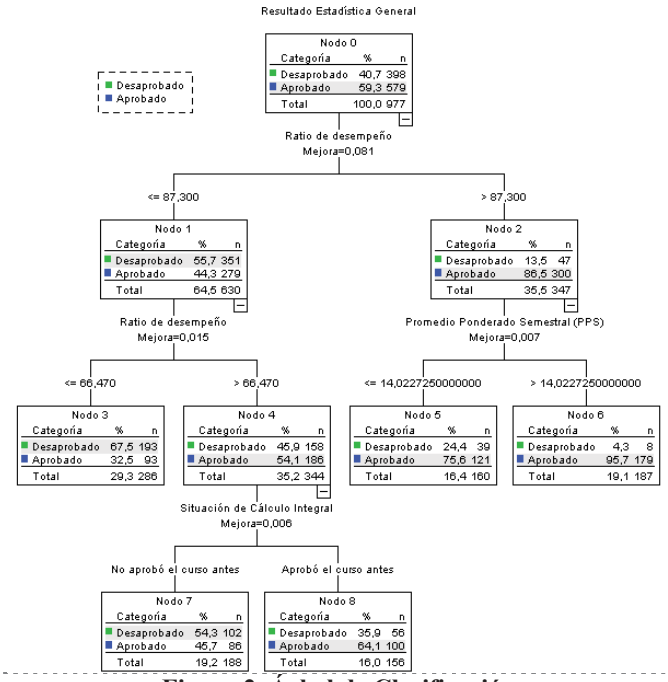

Figura 2. Árbol de Clasificación
Fuente: Elaboración propia

Situación similar se encontró con el análisis de regresión logística que seleccionó a las mismas variables seleccionadas con el árbol de clasificación y se añadieron dos variables predictoras: Edad y Especialidad, tal como puede apreciarse en la Figura 3.

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1[a]	ESP_CATE(1)	,386	,194	3,957	1	,047	1,471	1,006	2,152
	EDAD	-,099	,032	9,833	1	,002	,906	,851	,964
	PROMEDIOSEMESTRAL	,113	,034	10,652	1	,001	1,119	1,046	1,197
	RATIODESEMPEÑO	,046	,005	75,438	1	,000	1,047	1,036	1,058
	INTEGRAL(1)	,559	,153	13,379	1	,000	1,749	1,296	2,359
	Constante	-2,832	,949	8,903	1	,003	,059		

a. Variable(s) introducida(s) en el paso 1: ESP_CATE, EDAD, PROMEDIOSEMESTRAL, RATIODESEMPEÑO, INTEGRAL.

Figura 3. Modelo estadístico usando el Análisis de Regresión Logística
Feunte: Elaboración propia

Con los árboles de clasificación se obtuvo un 71,1% de acierto y con la regresión logística un 69,5% de acierto tal como se aprecia en la Tabla 2.

Observado	Pronosticado	
	% correcto (Árbol de Clasificación)	% correcto (Regresión Logística)
Desaprobado	74,1 %	56, 0%
Aprobado	69,1%	78,8 %
Porcentaje global	71,1 %	69,5 %

Tabla 2. Porcentaje de clasificación correcta e incorrecta obtenido con el Árbol de Clasificación y la Regresión Logística
Fuente: Elaboración propia

EVALUACIÓN Y VALIDACIÓN

Con la finalidad de validar el modelo obtenido con los datos de los semestres 2013-II y 2014-I se trabajó con los datos del semestre 2014-II y se obtuvo en dicho semestre un 67% de predicción usando el modelo obtenido con la regresión logística tal como se aprecia en la Tabla 3.

Observado	Pronosticado		
	Desaprobado	Aprobado	Porcentaje correcto
Desaprobado	86	72	54.5%
Aprobado	67	202	75.1%
Porcentaje global			67.4%

Tabla 3. Porcentaje de clasificación correcta e incorrecta obtenido con la Regresión Logística en la muestra de validación
Fuente: Elaboración propia

CONCLUSIONES

El resultado de que la variable predictora más importante haya sido el Ratio de Desempeño permite comprender que el resultado de Estadística General (aprobado/desaprobado) no es solamente un problema del mencionado curso sino de la situación académica del alumno. Los alumnos con un buen rendimiento académico en sus cursos anteriores tienen más oportunidad de aprobar que los alumnos con un bajo rendimiento académico. De los alumnos cuyo ratio de desempeño es menor de 67 (es decir que de 100 créditos sólo aprobaron 67) más del 67% de ellos desaprobó el curso. En cambio, de los alumnos cuyo ratio de desempeño fue mayor de 86 y cuyo último promedio ponderado semestral es mayor a 14 más, el 95% de ellos aprobó el curso.

Otro resultado importante es el hecho de que un alumno que ha aprobado antes el curso Calculo Integral es más probable que apruebe el curso Estadística General. En la currícula anterior de la UNALM el curso de Cálculo Integral era pre-requisito para matricularse en Estadística General que se cursaba en un cuarto ciclo. Sin embargo, ahora el pre-requisito es Cálculo Diferencial y Estadística General se puede cursar en un tercer ciclo donde aún los alumnos no tienen la suficiente madurez matemática como para afrontar exitosamente el curso.

La edad también es un factor determinante para la situación final del curso. La edad promedio de los desaprobados (22.98 años) supera en más de dos años a la edad promedio de los aprobados (20.93 años).

Es importante continuar el estudio realizado con la información de tres semestres académicos; se recomienda darle mantenimiento al modelo estadístico encontrado con la información de otras variables y semestres. De igual manera, diseñar estrategias para poder

disminuir el porcentaje de aprobados usando los resultados encontrados es de suma importancia; se recomienda empezar a detectar en el semestre 2015-I a los alumnos con una baja probabilidad de aprobar el curso a fin de poder brindarles algún tipo de ayuda mediante consultorías, asesorías, etc.

REFERENCIAS

Celis, S., Moreno, L., Poblete, P., Villanueva, J. y Weber, R. (2015). Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería. *Revista Ingeniería de Sistemas*, *29*, 5-24.

Cobo, A., Rocha, R. & Álvarez, Y. (2011). Selección de atributos predictivos del rendimiento académico de estudiantes en un modelo de b-learning. *Edutec-e. Revista Electrónica de Tecnología Educativa*, *37*. Recuperado de http://edutec.rediris.es/Revelec2/Revelec37/atributos_predictivos_rendimiento_academico_b-learning.html

Guzman, M. (2012). *Modelos predictivos y explicativos del rendimiento académico universitario: caso de una institución privada en México*. [Tesis doctoral]. Universidad Complutense de Madrid, España. Recuperado de http://eprints.ucm.es/15335/1/T33748.pdf

IBM Corporation (2012). *Manual CRISP-DM de IBM SPSS Modeler*. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J. & Ortigosa, A. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *Journal User Modeling and User-Adapted Interaction*, *21*(1-2), 217-248.