Factores de éxito en los alumnos SUAyED (modalidad a distancia): Aplicación de técnicas en minería de datos

José Gerardo Moreno Salinas
[gerardo_moreno@cuaed.unam.mx]
Coordinación de Universidad Abierta y Educación a Distancia – UNAM

Resumen

Cuando hablamos de estudios a nivel superior, sin duda la educación a distancia es un componente clave. Sin embargo, las tasas de deserción son altas y dada la importancia en la inversión de recursos tanto de las universidades y personas que deciden llevar una educación formal en la modalidad a distancia, se vuelve trascendental determinar los factores clave para el éxito y el fracaso. En este trabajo se analizan con técnicas de minería de datos los perfiles de ingreso, antecedentes académicos y matrícula de los alumnos del Sistema Universidad Abierta y Educación a Distancia (SUAyED) de la Universidad Nacional Autónoma de México (UNAM), con el propósito de determinar los elementos clave que impulsan el éxito y el fracaso de los alumnos, así como la creación de su respectivo modelo predictivo usando el algoritmo de clasificación Naive Bayes.

Palabras clave: Educación a distancia, factores de éxito, minería de datos, clasificador Naive Bayes.

Introducción

Las tasas de abandono en los programas a distancia son conocidas por ser significativas. El costo de la pérdida de un alumno es muy alto en términos de menoscabo de tiempo, esfuerzo y dinero por parte de todos los involucrados (alumnos, maestros y universidades). Tan pronto un alumno a distancia deja el programa, se pierde casi toda conexión con él y las instituciones por lo general no hacen nada para determinar los factores que motivaron la deserción del alumno. Yolanda Gayol (2016) presentó un comparativo entre las tasas de deserción de las principales universidades con programas en la modalidad a distancia en América Latina, donde hace evidente el gran problema que enfrentan las instituciones educativas, además de una interesante crítica sobre la utilización de la palabra deserción y sus implicaciones.

El abandono escolar y la identificación de los factores de éxito de los alumnos en educación superior ha sido un tema preocupante para todos los involucrados en el sistema educativo. En particular, para la modalidad a distancia Yukselturk (2014), investigó lo que varios autores dijeron estar de acuerdo *"las tasas de abandono para la educación a distancia son generalmente más altas que para la educación convencional. Muchos estudiantes fácilmente están abandonando sus cursos y programas en modalidades a distancia o en su defecto, los terminan insatisfechos*".

Por lo anterior, en tanto más y mejor conozcamos sobre los factores de éxito de nuestros alumnos, directamente estaremos abonando hacía mejorar la retención en la modalidad a distancia. Los datos pueden contener información esencial para mejorar la calidad de la educación, con el gran potencial de personalizar las experiencias de aprendizaje de los alumnos (Zhao y Luan, 2006).

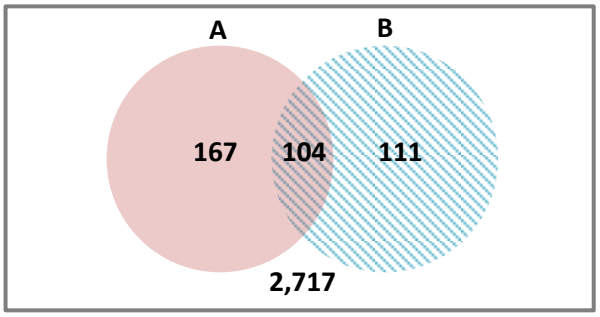
En la modalidad a distancia del Sistema Universidad Abierta y Educación a Distancia (SUAyED) de la Universidad Nacional Autónoma de México (UNAM), se han matriculado del 2005 a 2015 alrededor de 27,000 alumnos en 20 licenciaturas. Además de estar creciendo constantemente la cantidad de alumnos en el sistema, también se están sumado los datos que obtenemos de ellos, por ejemplo: perfiles de ingreso, antecedentes académicos y matrícula. Resultado del constante incremento en los volúmenes de datos, ha sido necesario recurrir a formas más eficaces y eficientes de analizar los datos y así, posibilitar la identificación de los principales factores que determinan el éxito de los alumnos en la modalidad a distancia.

La educación a distancia es más que una realidad, mucho de los aspirantes que están en posibilidades de estudiar una licenciatura, elijen la modalidad a distancia como su primera opción; tal es el caso de las tres últimas generaciones de nuevo ingreso al SUAyED, donde el 80% de alumnos seleccionaron la modalidad a distancia como su primera opción, el 15% como su segunda opción y el 5% como su tercera opción[1]. Por otro lado, las razones de porqué los aspirantes seleccionan ésta modalidad son variadas, entre las más representativas están: Mantener su trabajo, la distancia de los planteles de estudio y el cuidado de sus hijos[1]. Este par de variables aportan información de carácter descriptiva, pero asilada una de la otra.

[1] *Datos obtenidos del cuestionario de perfil de ingreso (2014-2015), CUAED, UNAM.*

Cuando se tienen dos variables y se quiere analizar qué tanto una está determinada por otra y viceversa, se pueden realizar a través de análisis condicionales. Por ejemplo, cuando hablamos del éxito que pueden alcanzar los alumnos lo podemos medir desde diferentes perspectivas, siendo las principales variables del alumno: "calificación promedio" y "duración de sus estudios", mismas que se pueden analizar de manera aislada o en su conjunto, tal es el caso de medir qué tanto la duración de los estudios está determinado por la calificación promedio y/o qué tanto la calificación promedio está determinada por la duración de sus estudios. Para hacer un mayor énfasis en dichas medidas y de su importancia, a continuación se presenta el diagrama de Venn y las probabilidades para una muestra de datos de los alumnos SUAyED en la modalidad a distancia.

Diagrama de Venn



N = 3,099

Dónde:

A) Alumnos con un promedio de calificaciones ≥ 9
B) Alumnos que cumplen con el 100% de créditos en 10 semestres

Probabilidades independientes

P(A) = 167/3,099 = *0.053%*
P(B) = 111/3,099 = *0.035%*
P(A∩B) = 104/3,099 = *0.033%*

Probabilidad condicional

$P(A|B) = \frac{\mathbf{P(A \cap B)}}{\mathbf{P(B)}}$ = *94%*

∴ La probabilidad de que un alumno tenga un promedio de calificaciones ≥ 9 dado que cubrió el 100% de créditos en 10 semestres es del **94%**

[Teorema de Bayes]

$P(B|A) = \mathbf{P(A|B)} \frac{\mathbf{P(B)}}{\mathbf{P(A)}}$ = *62%*

∴ La probabilidad de que un alumno termine el 100% de créditos en 10 semestres dado que tiene un promedio de calificaciones ≥ 9 es del **62%**

Figura 1. Diagrama de Venn y probabilidades

Del ejercicio anterior se puede precisar que un alumno tendrá mayores probabilidades de obtener una calificación promedio mayor o igual a nueve dado que cubrió el 100% de sus créditos en 10 semestres, es decir, será un factor mayor de éxito en

comparación con la probabilidad de cubrir el 100% de créditos en diez semestre dado que obtiene una calificación promedio mayor o igual a nueve.

Con el anterior ejemplo presentamos las relaciones de dependencia de tan sólo un pequeño subconjunto de variables, de las muchas posibles que existen y que requieren un análisis holístico, donde integre a la mayor cantidad de variables y muestre sus posibles relaciones. De modo que permita no sólo tener información sino conocimiento de los alumnos y sus variables, de manera que podamos notar cuáles son sus factores de éxito.

Para lograr lo anterior, se hizo una revisión en la literatura especializada donde se encontraron muchos análisis y casos de estudio que describen la pertinencia de realizar minería de datos sobre datos académicos. Incluso ya es común el término “Educational Data Mining” (Lile, 2011). Un claro ejemplo de la potencialidad que tiene hacer estudios de minería de datos en educación es el caso que presentaron Zang y Lin (2003), utilizaron un cuestionario de 37 preguntas en cinco secciones donde incluyeron datos demográficos y perspectivas de los estudiantes de un curso en línea. En ese estudio, los investigadores utilizaron algoritmos de minería de datos para predecir el éxito académico de los estudiantes, y encontraron que los estudiantes que publican preguntas y respuestas, y navegan en Internet tienden a obtener puntuaciones más altas.

Tal como lo comenta Romero (2008), la minería de datos incluye varias técnicas. Una de ellas es la clasificación, la cual se encarga de categorizar las entidades ya conocidas (datos de prueba) para después aplicarse a nuevos datos (entrenamiento). Los algoritmos de clasificación más populares, son los árboles de decisión, redes neuronales y Naive Bayes basado en el teorema de Bayes y el algoritmo genético (Yukselturk, 2014). Kotsiantis (2003) examinó con múltiples algoritmos de minería de datos los perfiles (sexo, edad, ocupación y asignaturas) de 510 estudiantes en línea, obteniendo los mejores resultados con el algoritmo Naive Bayes.

Condiciones previas

En análisis de esta naturaleza es imprescindible contar con la mayor cantidad de datos posibles, de modo que no se pierda el propósito del estudio y con el cumplimiento de la clase o condición buscada. No hay que perder de vista que utilizaremos técnicas de minería de datos centradas en algoritmos de clasificación, por lo que se vuelve vital preguntarnos qué es lo que queremos encontrar y definirla como la (sí clase).

Así como también es muy importante conocer las llamadas “reglas del negocio”, las cuales establecen las posibles políticas, operaciones, definiciones y restricciones presentes en una organización, es este caso de la universidad.

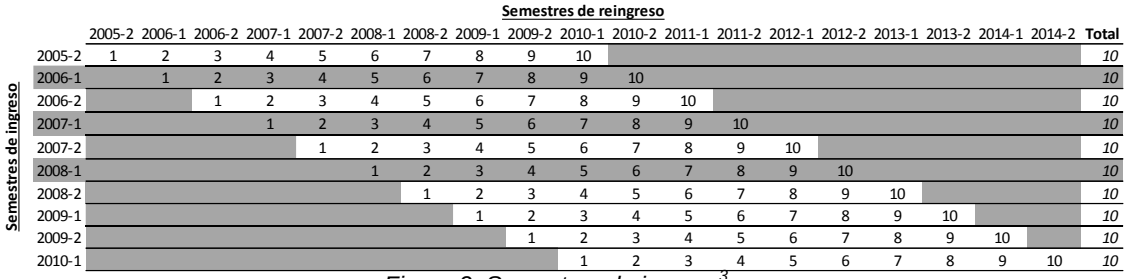
Por lo anterior, es importante mencionar que los alumnos del SUAyED y de acuerdo con el Reglamento de Estatuto General del SUAyED[2], tienen dos veces la duración señalada en el plan de estudios para el cumplimiento de la totalidad de los requisitos de estudios, al término del cual se causará baja en la institución. A este tiempo se le conoce como reglamentario. Por otro lado, el tiempo curricular es la duración que marca los planes y programas de estudios en la modalidad presencial de la UNAM, siendo en promedio nueve y un máximo diez semestres.

Lo anterior está muy relacionado con la segunda perspectiva del factor de éxito (duración de estudios) de los alumnos, analizado anteriormente en el diagrama de Venn. Desde el año 2009 el SUAyED en la modalidad a distancia comenzó a tener los primeros alumnos que cumplieron con el 100% de sus créditos en al menos nueve semestres y a partir de este año, cada semestre se han venido incrementado los alumnos que prefieren no agotar el tiempo reglamentario al que tienen derecho y deciden terminar sus estudios en un tiempo curricular. Dicha revelación ha motivado en descubrir cuáles fueron los factores determinantes para que los alumnos cumplieran con el 100% de créditos en un tiempo curricular, convirtiéndose así en nuestra (sí clase).

[2] *Disponible en http://bit.ly/1L8foQZ, consultado el 19 de abril de 2016.*

Recolección de datos

Del total de alumnos inscritos en la modalidad a distancia, se considerarán aquellos que tengan una duración de diez semestres en el SUAyED. En la siguiente figura se muestran los semestres de ingreso que cumplen con dicha condición.



Semestres de ingreso \ Semestres de reingreso	2005-2	2006-1	2006-2	2007-1	2007-2	2008-1	2008-2	2009-1	2009-2	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2	Total
2005-2	1	2	3	4	5	6	7	8	9	10										*10*
2006-1		1	2	3	4	5	6	7	8	9	10									*10*
2006-2			1	2	3	4	5	6	7	8	9	10								*10*
2007-1				1	2	3	4	5	6	7	8	9	10							*10*
2007-2					1	2	3	4	5	6	7	8	9	10						*10*
2008-1						1	2	3	4	5	6	7	8	9	10					*10*
2008-2							1	2	3	4	5	6	7	8	9	10				*10*
2009-1								1	2	3	4	5	6	7	8	9	10			*10*
2009-2									1	2	3	4	5	6	7	8	9	10		*10*
2010-1										1	2	3	4	5	6	7	8	9	10	*10*

Figura 2. Semestres de ingreso[3]

Con base en lo anterior, resulta una población de estudio igual a 2,889 alumnos, de los cuales el 6% (177) ha cumplido con la totalidad de sus créditos en tiempo curricular. A continuación se muestra la distribución por semestre de ingreso y clase.

Semestre de Ingreso	Clase		Total
	Sí	No	
2005-2	26	209	235
2006-2	12	240	252
2007-2	20	349	369
2008-2	21	353	374
2009-1	0	45	45
2009-2	26	313	339
2010-1	72	1,203	1,275
Total	**177**	**2,712**	**2,889**

Cuadro 1. Alumnos por semestres de ingreso y clase

La clase es un identificador que nos permite clasificar a los alumnos en dos grupos, los que al término de diez semestres sí cumplieron el 100% de créditos (sí clase) o en su defecto, los que al término de diez semestres no cumplieron el 100% de sus créditos (no clase).

Para cada uno de los alumnos se tienen los siguientes datos, los cuales son el principal recurso para el análisis y representan algunos de sus posibles factores, a continuación se muestran de acuerdo a su fuente de datos, descriptor y categoría.

[3] *En los semestres de ingreso 2006-1, 2007-1 y 2008-1 no hubo convocatorias para la modalidad a distancia.*

Fuente	Descriptor	Categoría
Matrícula (1er. semestre)	Semestre de ingreso	{Semestre Par, Semestre Non}
	Estado sede	{Chiapas, DF, Edomex, Hidalgo, Oaxaca, Querétaro, Tlaxcala}
	Asignaturas inscritas	{0, 1, 2, 3, 4, 5, 6, 7}
	Asignaturas aprobadas	{0, 1, 2, 3, 4, 5, 6, 7}
	Promedio a 1er semestre	{0, 5 - 5.4, 5.5 - 5.9, 6 - 6.4, 6.5 - 6.9, 7 - 7.4, 7.5 – 7.9, 8 - 8.4, 8.5 - 8.9, 9 - 9.4, 9.5 - 9.9, 10}
	Causa ingreso[4]	{56, 58, 64, 67, 74, 76, 97}
	Sexo	{Femenino, Masculino}
	Edad de ingreso	{16 a 19, 19 a 22, 22 a 24, 24 a 26, 26 a 28, 28 a 31, 31 a 34, 34 a 38, 38 a 43, 43 a 63}. Categorías con base a “coarse graining”
	Nacionalidad	{1 (Mexicana), 2 (Extranjero), N.D.*}
	Entidad de residencia	{Aguascalientes, Chiapas, Chihuahua, Coahuila, Distrito Federal, Durango, Guanajuato… Zacatecas}
	Carrera	{Administración, Bibliotecología, C. Políticas, Contaduría, Derecho, Economía, Español como LE., Informática, Inglés como LE., Pedagogía, Periodismo, Psicología, R. Internacionales, Sociología}
Perfil de ingreso	Estado civil	{Casado, Divorciado, Viudo, Separado, N.D., Soltero, Unión libre}
	Descendencia	{N.D., No, Si}
	Condición laboral	{N.D., No, Si}
	Relación laboral	{Eventualmente, Permanentemente, No contesto, N.D.}
	Horas de trabajo a la semana	{10 o menos, De 11 a 20, De 21 a 30, De 31 a 40, Más de 40, No contesto, N.D.}
	Relación del trabajo con sus estudios	{No, Si, No contesto, N.D.}
	Personas que dependen del sostén económico	{1 a 2, 3 a 4, 5 a 6, 7 o más, N.D.}
	Personas que contribuyen al ingreso familiar	{1, 2 a 3, 4 o más, N.D.}
	Escolaridad máxima de la madre	{Carrera técnica, Licenciatura, Media superior o normal, N.D., Posgrado, Primaria, Secundaria, Sin instrucción}
	Escolaridad máxima del padre	{Carrera técnica, Licenciatura, Media superior o normal, N.D., Posgrado, Primaria, Secundaria, Sin instrucción}
	Acceso a equipo de cómputo	{N.D., No, Si}
	Entiende inglés	{Bien, Con dificultad, Nada, N.D.}
	Habla inglés	{Bien, Con dificultad, Nada, N.D.}
	Escribe inglés	{Bien, Con dificultad, Nada, N.D.}
Historial académico	Nivel máximo de estudios	{Bachillerato, Licenciatura, Licenciatura sin concluir, N.D., Posgrado, Posgrado sin concluir}
	Promedio obtenido último nivel de estudios	{7 a 7.9, 8 a 8.9, 9 a 10, N.D.}
	Tiempo sin estudiar	{Menos de 2 años, De 2 a 4 años, De 5 a 8 años, Mas de 8 años, N.D}
Aciertos en el examen de Ingreso		{33 a 40, 41 a 48, 49 a 56, 57 a 64, 65 a 72, 73 a 80, 81 a 88, 89 a 96, 97 a 104, N.D.}

Cuadro 2. Datos (fuente, descriptor y categoría)

[4] 56 (concurso de selección); 58 (2da carrera); 64 (2da carrera y concurso de selección); 67 (cambio de carrera y/o plantel); 74 (procedente de nivel técnico); 76 (cambio de sistema); 97 (cambio de modalidad).
** N.D. (No Disponible).*

Metodología

Al tener seleccionados los datos con los que trabajará el análisis, así como la identificación de las clases (sí/no), es momento presentar la metodología que se utilizó con técnicas de minería de datos y la implementación del algoritmo de clasificación Naive Bayes con el cual se logró identificar los factores de éxito de los alumnos inscritos en la modalidad a distancia del SUAyED. Así como presentar el mapeo de todas las variables analizadas.

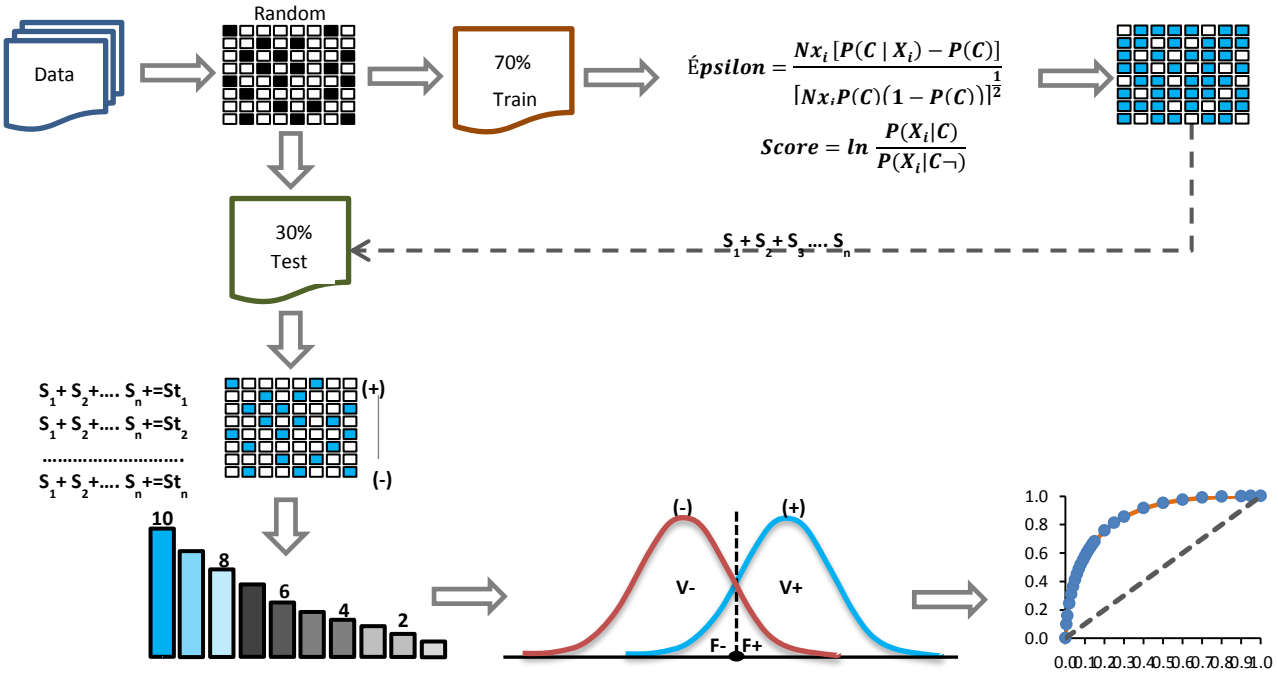


Figura 3. Metodología para la minería de datos

La metodología describe las fases a seguir para la selección de datos, cálculos, análisis e interpretación de resultados, así como la visualización y medición del rendimiento del modelo clasificador.

Del total de alumnos a distancia del SUAyED se seleccionaron aquellos que por su semestre de ingreso han cumplido con la duración de diez semestres en sus estudios. Se definió la clase para el análisis y se identificaron los perfiles de ingreso, antecedentes académicos y matrícula en el primer semestre. Para después, integrarlos en una base de datos relacional. Se seleccionaron aleatoriamente dos conjuntos de datos; el 70% se asignó para el entrenamiento del modelo y el 30% para los datos de prueba. Sobre los datos de entrenamiento se calcularon las medidas de Épsilon y Score, los cuales permitieron tener otra dimensión de análisis para cada descriptor de los alumnos y sus categorías, haciendo así, más fácil el proceso de identificación de variables a considerar para el modelo de clasificación.

Al tener los Score seleccionados por categoría, se asignaron al 30% de los datos de prueba. Después, se suman todos los Score por instancia y los resultados finales se ordenan de forma descendente para graficar por deciles la distribución que tiene el modelo de clasificación y así ver la discriminación que hace entre las clases. Para evaluar el rendimiento del modelo de clasificación se utiliza una matriz de confusión y

sus medidas de rendimiento. Por último, se grafica el modelo en una curva ROC (Receiver Operating Characteristic) para examinar la capacidad que tiene el modelo clasificador para identificar y predecir correctamente los casos positivos y el número de casos negativos que se clasifican incorrectamente.

Algoritmo de clasificación

El algoritmo Naive Bayes es un algoritmo de clasificación basado en el teorema de Bayes, el cual asume predictores condicionalmente independientes que suman hacia el resultado objetivo C_k. Es decir, adoptando el teorema de Bayes para calcular la probabilidad posterior de la variable C_k considerando la independencia de los predictores X= $\{X_1, X_2 \dots X_m\}$, se obtiene la siguiente distribución de probabilidad conjunta.

$$P\ (C_k, X) = \ P(C_k|\ X)\ P(X)$$

Aplicando el teorema de Bayes, resulta:

$$P\ (C_k, X) = \ \frac{P\ (X|C_k)\ P\ (C_k)}{P(X)} = \ \frac{\prod_{i=1}^{N} P(X_i|\ C_k)}{P(X)}$$

En la segunda igualdad se asume que las variables X_i son independientes y el producto es sobre el total de variables (N) consideradas como factores condicionantes para C_k. Stephens (2009) presenta la siguiente prueba estadística la cual mide la dependencia estadística de C_k en X_i.

$$\varepsilon\ (C_k\ |\ X_i) = \frac{Nx_j[P(C_k\ |\ X_i) - P(C_k)]}{\left[Nx_j P(C_k)\big(1 - P(C_k)\big)\right]^{\frac{1}{2}}}$$

Dónde N es el total de observaciones, Nx_j es el número total de observaciones con $X_i = j$, NC_k es el total de observaciones que cumplen con C_k, $P(C_k)$ es la probabilidad de C_k y $P\big(C_k\ |\ X_i\big)$ es la probabilidad de C_k dado X_i. Dicha ecuación se le conoce como Épsilon y considerando una aproximación normal a la distribución binomial, $\varepsilon\ (C_k\ |\ X_i) = 2$ representa el 95% del nivel de confianza y es estadísticamente significante, Stephens (2009).

Relacionado al indicador de Épsilon, el Score suma al modelo clasificador hacía el mayor cumplimiento de la clase, es decir, el Score mide el grado en que una instancia es miembro de una clase, por lo que a mayores valores de Score, mayor será las probabilidad de cumplir con C_k. La función Score permite calcular las predicciones dado una serie de predictores y se utiliza como *Proxy* del algoritmo Naive Bayes, para $P\ (C_k, X)$.

$$S\ (C_k|\ X) = \sum_{i=1}^{N} S\ (C_k|X_i\ ) = \ \sum_{i=1}^{N} ln\frac{P(X_i|C_k)}{P(X_i|C_k\neg)}$$

Donde $C_k\neg$ es el complemento del conjunto C_k. Cuando el valor de Score es igual a cero, significa que la probabilidad de encontrar C_k es aleatoria. Un valor positivo en Score indica un alto factor de ocurrencia de la clase C_k.

Resultados e interpretación

Al programar el algoritmo de clasificación y al ejecutarlo con la base de datos definida anteriormente (recolección de datos), procesó los siguientes resultados para cada una de las variables (descriptor, con su respectiva categoría).

Descriptor	Categoría	Nx	Nxc=Ncx	N	Nc	Pc	P(x|c)	P(c|x)	Épsilon	Score
Semestre de ingreso	Semestre Par	1,107	70	2,022	114	0.06	0.61	0.06	0.989	0.119
Semestre de ingreso	Semestre Nor	915	44	2,022	114	0.06	0.39	0.05	-1.088	-0.163
Estado sede	Chiapas	22	0	2,022	114	0.06	0.01	0.00	-1.147	-0.334
Estado sede	DF	497	27	2,022	114	0.06	0.24	0.05	-0.199	-0.021
Estado sede	Edomex	197	13	2,022	114	0.06	0.12	0.07	0.585	0.220
Estado sede	Hidalgo	197	12	2,022	114	0.06	0.11	0.06	0.276	0.140
Estado sede	Oaxaca	411	23	2,022	114	0.06	0.21	0.06	-0.037	0.016
Estado sede	Querétaro	9	0	2,022	114	0.06	0.01	0.00	-0.733	0.499
Estado sede	Tlaxcala	689	39	2,022	114	0.06	0.34	0.06	0.025	0.012
Asignaturas inscritas a 1er. semestre	0	45	1	2,022	114	0.06	0.02	0.02	-0.993	-0.312
Asignaturas inscritas a 1er. semestre	1	13	0	2,022	114	0.06	0.01	0.00	-0.881	0.162
Asignaturas inscritas a 1er. semestre	2	10	0	2,022	114	0.06	0.01	0.00	-0.773	0.403
Asignaturas inscritas a 1er. semestre	3	46	0	2,022	114	0.06	0.01	0.00	-1.658	-1.049
Asignaturas inscritas a 1er. semestre	4	82	1	2,022	114	0.06	0.02	0.01	-1.735	-0.912
Asignaturas inscritas a 1er. semestre	5	126	3	2,022	114	0.06	0.03	0.02	-1.585	-0.633
Asignaturas inscritas a 1er. semestre	6	980	59	2,022	114	0.06	0.52	0.06	0.519	0.069
Asignaturas inscritas a 1er. semestre	7	720	50	2,022	114	0.06	0.44	0.07	1.520	0.224
Asignaturas aprobadas a 1er. semestre	0	896	4	2,022	114	0.06	0.04	0.00	-6.737	-2.384
Asignaturas aprobadas a 1er. semestre	1	167	1	2,022	114	0.06	0.02	0.01	-2.823	-1.624
Asignaturas aprobadas a 1er. semestre	2	118	1	2,022	114	0.06	0.02	0.01	-2.256	-1.276
Asignaturas aprobadas a 1er. semestre	3	139	1	2,022	114	0.06	0.02	0.01	-2.514	-1.440
Asignaturas aprobadas a 1er. semestre	4	107	1	2,022	114	0.06	0.02	0.01	-2.109	-1.178
Asignaturas aprobadas a 1er. semestre	5	164	17	2,022	114	0.06	0.16	0.10	2.625	0.694
Asignaturas aprobadas a 1er. semestre	6	289	58	2,022	114	0.06	0.51	0.20	10.636	1.432
Asignaturas aprobadas a 1er. semestre	7	142	31	2,022	114	0.06	0.28	0.22	8.366	1.549
Promedio 1er. semestre	0	542	2	2022	114	0.06	0.03	0.00	-5.318	-2.394
Promedio 1er. semestre	5 a 5.4	394	2	2,022	114	0.06	0.03	0.01	-4.415	-2.074
Promedio 1er. semestre	5.5 a 5.9	65	1	2,022	114	0.06	0.02	0.02	-1.433	-0.680
Promedio 1er. semestre	6 a 6.4	109	1	2,022	114	0.06	0.02	0.01	-2.137	-1.197
Promedio 1er. semestre	6.5 a 6.9	88	2	2,022	114	0.06	0.03	0.02	-1.369	-0.566
Promedio 1er. semestre	7 a 7.4	168	7	2,022	114	0.06	0.07	0.04	-0.827	-0.207
Promedio 1er. semestre	7.5 a 7.9	125	8	2,022	114	0.06	0.08	0.06	0.369	0.228
Promedio 1er. semestre	8 a 8.4	163	24	2,022	114	0.06	0.22	0.15	5.029	1.079
Promedio 1er. semestre	8.5 a 8.9	118	20	2,022	114	0.06	0.18	0.17	5.327	1.251
Promedio 1er. semestre	9 a 9.4	174	34	2,022	114	0.06	0.30	0.20	7.951	1.408
Promedio 1er. semestre	9.5 a 9.9	61	10	2,022	114	0.06	0.09	0.16	3.642	1.248
Promedio 1er. semestre	10	15	3	2,022	114	0.06	0.03	0.20	2.412	1.623
Causa de ingreso	54	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Causa de ingreso	56	1,809	106	2,022	114	0.06	0.92	0.06	0.409	0.033
Causa de ingreso	58	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Causa de ingreso	64	62	4	2,022	114	0.06	0.04	0.06	0.278	0.333
Causa de ingreso	67	146	4	2,022	114	0.06	0.04	0.03	-1.518	-0.552
Causa de ingreso	69	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Causa de ingreso	74	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Causa de ingreso	76	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Causa de ingreso	97	2	0	2,022	114	0.06	0.01	0.00	-0.346	1.703
Sexo	F	1,013	61	2,022	114	0.06	0.53	0.06	0.530	0.069
Sexo	M	1,009	53	2,022	114	0.06	0.47	0.05	-0.531	-0.074

Descriptor	Categoría	Nx	Nxc=Ncx	N	Nc	Pc	P(x|c)	P(c|x)	Épsilon	Score
Edad de ingreso	16 a 19	192	10	2,022	114	0.06	0.09	0.05	-0.258	-0.010
Edad de ingreso	19 a 22	208	6	2,022	114	0.06	0.06	0.03	-1.722	-0.566
Edad de ingreso	22 a 24	201	12	2,022	114	0.06	0.11	0.06	0.204	0.119
Edad de ingreso	24 a 26	211	7	2,022	114	0.06	0.07	0.03	-1.461	-0.442
Edad de ingreso	26 a 28	208	12	2,022	114	0.06	0.11	0.06	0.082	0.083
Edad de ingreso	28 a 31	209	5	2,022	114	0.06	0.05	0.02	-2.034	-0.730
Edad de ingreso	31 a 34	196	10	2,022	114	0.06	0.09	0.05	-0.325	-0.032
Edad de ingreso	34 a 38	201	16	2,022	114	0.06	0.15	0.08	1.427	0.409
Edad de ingreso	38 a 43	202	17	2,022	114	0.06	0.16	0.08	1.712	0.466
Edad de ingreso	43 a 63	194	19	2,022	114	0.06	0.17	0.10	2.510	0.627
Nacionalidad	1	2,018	114	2,022	114	0.06	0.99	0.06	0.022	-0.006
Nacionalidad	2	3	0	2,022	114	0.06	0.01	0.00	-0.423	1.415
Nacionalidad	N.D.	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Entidad de residencia	Aguascaliente	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Entidad de residencia	Chiapas	23	0	2,022	114	0.06	0.01	0.00	-1.172	-0.377
Entidad de residencia	Chihuahua	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Entidad de residencia	Coahuila	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Entidad de residencia	Distrito Feder	336	16	2,022	114	0.06	0.15	0.05	-0.696	-0.137
Entidad de residencia	Durango	3	1	2,022	114	0.06	0.02	0.33	2.080	2.396
Entidad de residencia	Guanajuato	4	0	2,022	114	0.06	0.01	0.00	-0.489	1.192
Entidad de residencia	Guerrero	3	0	2,022	114	0.06	0.01	0.00	-0.423	1.415
Entidad de residencia	Hidalgo	201	11	2,022	114	0.06	0.10	0.05	-0.102	0.034
Entidad de residencia	Jalisco	3	0	2,022	114	0.06	0.01	0.00	-0.423	1.415
Entidad de residencia	México	347	23	2,022	114	0.06	0.21	0.07	0.800	0.195
Entidad de residencia	Michoacán	7	1	2,022	114	0.06	0.02	0.14	0.992	1.549
Entidad de residencia	Morelos	15	2	2,022	114	0.06	0.03	0.13	1.292	1.261
Entidad de residencia	Oaxaca	404	23	2,022	114	0.06	0.21	0.06	0.048	0.034
Entidad de residencia	Puebla	60	3	2,022	114	0.06	0.03	0.05	-0.214	0.127
Entidad de residencia	Querétaro	12	1	2,022	114	0.06	0.02	0.08	0.405	1.010
Entidad de residencia	Quintana Roo	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Entidad de residencia	San Luis Potos	3	0	2,022	114	0.06	0.01	0.00	-0.423	1.415
Entidad de residencia	Sonora	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Entidad de residencia	Tabasco	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Entidad de residencia	Tamaulipas	3	0	2,022	114	0.06	0.01	0.00	-0.423	1.415
Entidad de residencia	Tlaxcala	583	31	2,022	114	0.06	0.28	0.05	-0.336	-0.048
Entidad de residencia	Veracruz	10	2	2,022	114	0.06	0.03	0.20	1.969	1.703
Entidad de residencia	Yucatán	1	0	2,022	114	0.06	0.01	0.00	-0.244	2.108
Entidad de residencia	Zacatecas	2	0	2,022	114	0.06	0.01	0.00	-0.346	1.703
Carrera	Administració	98	4	2,022	114	0.06	0.04	0.04	-0.668	-0.143
Carrera	Bibliotecologí	29	1	2,022	114	0.06	0.02	0.03	-0.511	0.127
Carrera	C. Políticas	153	6	2,022	114	0.06	0.06	0.04	-0.920	-0.250
Carrera	Contaduría	224	18	2,022	114	0.06	0.16	0.08	1.556	0.413
Carrera	Derecho	521	47	2,022	114	0.06	0.41	0.09	3.348	0.509
Carrera	Economía	75	2	2,022	114	0.06	0.03	0.03	-1.116	-0.404
Carrera	Español como	2	0	2,022	114	0.06	0.01	0.00	-0.346	1.703
Carrera	Informática	29	1	2,022	114	0.06	0.02	0.03	-0.511	0.127
Carrera	Inglés como L	0	0	2,022	114	0.06	0.01	0.00	0.000	0.000
Carrera	Pedagogía	187	1	2,022	114	0.06	0.02	0.01	-3.026	-1.737
Carrera	Periodismo	120	6	2,022	114	0.06	0.06	0.05	-0.303	0.002
Carrera	Psicología	518	26	2,022	114	0.06	0.23	0.05	-0.610	-0.103
Carrera	R. Internacior	47	1	2,022	114	0.06	0.02	0.02	-1.043	-0.356
Carrera	Sociología	19	1	2,022	114	0.06	0.02	0.05	-0.071	0.550
Estado civil	Casado	554	45	2,022	114	0.06	0.40	0.08	2.536	0.395
Estado civil	Divorciado, vi	111	8	2,022	114	0.06	0.08	0.07	0.717	0.354
Estado civil	N.D.	367	14	2,022	114	0.06	0.13	0.04	-1.514	-0.360
Estado civil	Soltero	846	43	2,022	114	0.06	0.38	0.05	-0.700	-0.104
Estado civil	Unión libre	144	4	2,022	114	0.06	0.04	0.03	-1.488	-0.538

Descriptor	Categoría	Nx	Nxc=Ncx	N	Nc	Pc	P(x|c)	P(c|x)	Épsilon	Score
Descendencia	N.D.	479	21	2,022	114	0.06	0.19	0.04	-1.190	-0.237
Descendencia	No	748	35	2,022	114	0.06	0.31	0.05	-1.137	-0.186
Descendencia	Si	795	58	2,022	114	0.06	0.51	0.07	2.026	0.275
Condición laboral	N.D.	362	12	2,022	114	0.06	0.11	0.03	-1.916	-0.495
Condición laboral	No	263	20	2,022	114	0.06	0.18	0.08	1.383	0.349
Condición laboral	Si	1,397	82	2,022	114	0.06	0.72	0.06	0.376	0.038
Relación laboral	Eventualment	476	18	2,022	114	0.06	0.16	0.04	-1.756	-0.383
Relación laboral	Permanentem	884	60	2,022	114	0.06	0.53	0.07	1.482	0.197
Relación laboral	No contesto	273	20	2,022	114	0.06	0.18	0.07	1.209	0.308
Relación laboral	N.D.	389	16	2,022	114	0.06	0.15	0.04	-1.304	-0.290
Horas de trabajo a la semana	10 o menos	150	7	2,022	114	0.06	0.07	0.05	-0.516	-0.089
Horas de trabajo a la semana	De 11 a 20	118	12	2,022	114	0.06	0.11	0.10	2.134	0.693
Horas de trabajo a la semana	De 21 a 30	163	8	2,022	114	0.06	0.08	0.05	-0.404	-0.051
Horas de trabajo a la semana	De 31 a 40	360	22	2,022	114	0.06	0.20	0.06	0.389	0.111
Horas de trabajo a la semana	Más de 40	576	29	2,022	114	0.06	0.26	0.05	-0.628	-0.104
Horas de trabajo a la semana	No contesto	266	20	2,022	114	0.06	0.18	0.08	1.330	0.336
Horas de trabajo a la semana	N.D.	389	16	2,022	114	0.06	0.15	0.04	-1.304	-0.290
Relación del trabajo con sus estudios	No	637	35	2,022	114	0.06	0.31	0.05	-0.157	-0.017
Relación del trabajo con sus estudios	Si	721	42	2,022	114	0.06	0.37	0.06	0.218	0.040
Relación del trabajo con sus estudios	No contesto	275	21	2,022	114	0.06	0.19	0.08	1.437	0.351
Relación del trabajo con sus estudios	N.D.	389	16	2,022	114	0.06	0.15	0.04	-1.304	-0.290
Personas que dependen del sostén económ	1 a 2	737	41	2,022	114	0.06	0.36	0.06	-0.088	-0.008
Personas que dependen del sostén económ	3 a 4	675	40	2,022	114	0.06	0.35	0.06	0.324	0.060
Personas que dependen del sostén económ	5 a 6	174	14	2,022	114	0.06	0.13	0.08	1.377	0.428
Personas que dependen del sostén económ	7 o más	27	2	2,022	114	0.06	0.03	0.07	0.399	0.642
Personas que dependen del sostén económ	N.D.	409	17	2,022	114	0.06	0.16	0.04	-1.299	-0.282
Personas que contribuyen al ingreso familia	1	680	41	2,022	114	0.06	0.36	0.06	0.443	0.077
Personas que contribuyen al ingreso familia	2 a 3	883	54	2,022	114	0.06	0.47	0.06	0.615	0.087
Personas que contribuyen al ingreso familia	4 o más	52	0	2,022	114	0.06	0.01	0.00	-1.763	-1.169
Personas que contribuyen al ingreso familia	N.D.	407	19	2,022	114	0.06	0.17	0.05	-0.848	-0.167
Escolaridad máxima de la madre	Sin instrucción	144	13	2,022	114	0.06	0.12	0.09	1.764	0.558
Escolaridad máxima de la madre	Primaria	598	42	2,022	114	0.06	0.37	0.07	1.469	0.240
Escolaridad máxima de la madre	Secundaria	347	18	2,022	114	0.06	0.16	0.05	-0.364	-0.053
Escolaridad máxima de la madre	Media superic	176	5	2,022	114	0.06	0.05	0.03	-1.609	-0.554
Escolaridad máxima de la madre	Carrera técnic	259	20	2,022	114	0.06	0.18	0.08	1.454	0.365
Escolaridad máxima de la madre	Licenciatura	110	4	2,022	114	0.06	0.04	0.04	-0.910	-0.262
Escolaridad máxima de la madre	Posgrado	19	0	2,022	114	0.06	0.01	0.00	-1.065	-0.194
Escolaridad máxima de la madre	N.D.	369	12	2,022	114	0.06	0.11	0.03	-1.987	-0.514
Escolaridad máxima del padre	Sin instrucción	119	10	2,022	114	0.06	0.09	0.08	1.308	0.499
Escolaridad máxima del padre	Primaria	521	34	2,022	114	0.06	0.30	0.07	0.879	0.166
Escolaridad máxima del padre	Secundaria	331	22	2,022	114	0.06	0.20	0.07	0.796	0.200
Escolaridad máxima del padre	Media superic	197	7	2,022	114	0.06	0.07	0.04	-1.269	-0.372
Escolaridad máxima del padre	Carrera técnic	151	10	2,022	114	0.06	0.09	0.07	0.525	0.243
Escolaridad máxima del padre	Licenciatura	274	15	2,022	114	0.06	0.14	0.05	-0.117	0.013
Escolaridad máxima del padre	Posgrado	52	3	2,022	114	0.06	0.03	0.06	0.041	0.276
Escolaridad máxima del padre	N.D.	377	13	2,022	114	0.06	0.12	0.03	-1.843	-0.460
Acceso a equipo de cómputo	N.D.	393	16	2,022	114	0.06	0.15	0.04	-1.347	-0.300
Acceso a equipo de cómputo	No	95	10	2,022	114	0.06	0.09	0.11	2.066	0.745
Acceso a equipo de cómputo	Si	1,534	88	2,022	114	0.06	0.77	0.06	0.168	0.013
Entiende inglés	Bien	448	23	2,022	114	0.06	0.21	0.05	-0.463	-0.075
Entiende inglés	Con dificultad	1,036	67	2,022	114	0.06	0.59	0.06	1.157	0.143
Entiende inglés	Nada	172	11	2,022	114	0.06	0.10	0.06	0.431	0.199
Entiende inglés	N.D.	366	13	2,022	114	0.06	0.12	0.04	-1.730	-0.429
Habla inglés	Bien	251	15	2,022	114	0.06	0.14	0.06	0.232	0.106
Habla inglés	Con dificultad	1,009	58	2,022	114	0.06	0.51	0.06	0.152	0.020
Habla inglés	Nada	378	25	2,022	114	0.06	0.22	0.07	0.822	0.190
Habla inglés	N.D.	384	16	2,022	114	0.06	0.15	0.04	-1.250	-0.276

Descriptor	Categoría	Nx	Nxc=Ncx	N	Nc	Pc	P(x|c)	P(c|x)	Épsilon	Score
Escribe inglés	Bien	371	22	2,022	114	0.06	0.20	0.06	0.244	0.079
Escribe inglés	Con dificultad	858	52	2,022	114	0.06	0.46	0.06	0.537	0.078
Escribe inglés	Nada	381	21	2,022	114	0.06	0.19	0.06	-0.107	0.003
Escribe inglés	N.D.	412	19	2,022	114	0.06	0.17	0.05	-0.903	-0.179
Nivel máximo de estudios	Bachillerato	756	54	2,022	114	0.06	0.47	0.07	1.794	0.253
Nivel máximo de estudios	Licenciatura	302	12	2,022	114	0.06	0.11	0.04	-1.254	-0.307
Nivel máximo de estudios	Licenciatura s	446	21	2,022	114	0.06	0.19	0.05	-0.851	-0.162
Nivel máximo de estudios	N.D.	435	20	2,022	114	0.06	0.18	0.05	-0.941	-0.185
Nivel máximo de estudios	Posgrado	51	6	2,022	114	0.06	0.06	0.12	1.897	0.919
Nivel máximo de estudios	Posgrado sin c	32	1	2,022	114	0.06	0.02	0.03	-0.616	0.029
Promedio obtenido último nivel	7 a 7.9	650	40	2,022	114	0.06	0.35	0.06	0.570	0.100
Promedio obtenido último nivel	8 a 8.9	743	40	2,022	114	0.06	0.35	0.05	-0.301	-0.042
Promedio obtenido último nivel	9 a 10	252	20	2,022	114	0.06	0.18	0.08	1.582	0.395
Promedio obtenido último nivel	N.D.	377	14	2,022	114	0.06	0.13	0.04	-1.620	-0.388
Tiempo sin estudiar	Menos de 2 ai	457	29	2,022	114	0.06	0.26	0.06	0.656	0.141
Tiempo sin estudiar	De 2 a 4 años	432	20	2,022	114	0.06	0.18	0.05	-0.909	-0.178
Tiempo sin estudiar	De 5 a 8 años	282	11	2,022	114	0.06	0.10	0.04	-1.265	-0.320
Tiempo sin estudiar	Mas de 8 años	386	37	2,022	114	0.06	0.33	0.10	3.362	0.581
Tiempo sin estudiar	N.D.	465	17	2,022	114	0.06	0.16	0.04	-1.853	-0.415
Examen de ingreso	33 a 40	68	4	2,022	114	0.06	0.04	0.06	0.087	0.236
Examen de ingreso	41 a 48	422	20	2,022	114	0.06	0.18	0.05	-0.800	-0.153
Examen de ingreso	49 a 56	553	32	2,022	114	0.06	0.28	0.06	0.152	0.040
Examen de ingreso	57 a 64	378	20	2,022	114	0.06	0.18	0.05	-0.292	-0.038
Examen de ingreso	65 a 72	269	17	2,022	114	0.06	0.16	0.06	0.485	0.158
Examen de ingreso	73 a 80	129	7	2,022	114	0.06	0.07	0.05	-0.104	0.069
Examen de ingreso	81 a 88	72	5	2,022	114	0.06	0.05	0.07	0.481	0.374
Examen de ingreso	89 a 96	40	3	2,022	114	0.06	0.03	0.08	0.511	0.550
Examen de ingreso	97 a 104	5	0	2,022	114	0.06	0.01	0.00	-0.547	1.010
Examen de ingreso	N.D.	86	6	2,022	114	0.06	0.06	0.07	0.538	0.353

Cuadro 3. Cálculo de variables (Épsilon y Score)

No hay que olvidar que los resultados anteriormente presentados refieren exclusivamente a los datos de entrenamiento (70%), es decir, son datos que se utilizaron para construir el modelo de clasificación, para ello se calcularon las medidas de Épsilon y Score. Al tener todos los valores de Épsilon posibilitó realizar un análisis de selección de variables o mejor conocido por su nombre en inglés “Feature Selection” y así, decidir qué variables (categorías por descriptor) y sus respectivos Score se incluirán en el modelo de clasificación. El criterio de selección fue para todos aquellos Épsilon mayores o igual a dos desviaciones estándar (95% de nivel de confianza). A continuación sólo se presentan los descriptores en los que alguna de sus categorías cumple con $\varepsilon\ (C_k \mid X_i) = 2$

En la columna Épsilon se aplicó un formato condicional por cada descriptor, de modo que permita identificar aquellas categorías que tienen una mayor representación para el modelo, las barras de color rojo identifican a los valores negativos y las de color azul a los Épsilon positivos. A continuación por cada una de las fuentes de datos, se hace la interpretación con base a los resultados calculados de Épsilon.

Matrícula

- El semestre de ingreso por tener un sesgo no se considerará en el modelo, ya que son cinco semestres pares en comparación con dos semestres impares.
- El estado de México e Hidalgo son las únicas sedes que tiene un Épsilon positivo.

- Para el caso de las asignaturas inscritas y aprobadas en el primer semestre, los alumnos que inscriben de seis a siete asignaturas obtienen un Épsilon positivo que no supera el 1.5 de desviaciones estándar, mientras tanto, los alumnos que aprueban en su primer semestre cinco o más asignaturas, tiene un Épsilon mayor de 2.6, en particular, 10.6 y 8.4 para los que aprueban seis y siete asignaturas, respectivamente.
- Del promedio de calificaciones en el primer semestre, están por arriba de dos desviaciones estándar aquellos que promedian en sus asignaturas una calificación igual o mayor a ocho. En particular resalta el valor de Épsilon de 7.95 para quienes obtienen un promedio entre 9 a 9.4 en sus asignaturas al primer semestre.
- La únicas causas de ingreso con valores positivos en Épsilon, son la 56 (ingreso a la licenciatura por concurso de selección) y la 64 (ingreso a la licenciatura por segunda carrera), pero ninguna de éstas es representativa.
- Las mujeres tienen un Épsilon positivo versus los hombres, Épsilon negativo.
- Los alumnos con una edad mayor a 34 años al momento de ingresar a la carrera, tienen Épsilon por encima a 1.4, en particular los que están dentro de los 43 a 63 años tienen 2.5 desviaciones estándar.
- La nacionalidad de los alumnos al tener casi la totalidad de los alumnos (Nx) y en una misma categoría se considera una variable tautológica.
- Los estados de residencia de los alumnos con valores más altos de Épsilon, son: Durango y Veracruz con 2.08 y 1.96, respectivamente.
- Las carreras con Épsilon más grandes son contaduría y derecho con 1.5 y 3.3, respectivamente.

Perfiles de ingreso

- Los alumnos que su estado civil es casado, sí tienen hijos y no trabajan tienen un Épsilon promedio igual a dos.
- Dependiendo de la respuesta que se dé a la pregunta “condición laboral” se derivan tres posibles preguntas. Para el particular caso de 11 a 20 horas de trabajo a la semana, resulta 2.13 desviaciones estándar.
- De acuerdo al número de personas que dependen del principal sostén económico, las familias de los alumnos con cinco a seis integrantes tienen el mayor Épsilon.
- De uno a tres integrantes que contribuyen al ingreso familiar tienen Épsilon positivo, pero no son significativos.
- Las madres y padres de los alumnos que no tienen instrucción académica presentan el mayor Épsilon de las categorías, 1.8 y 1.3, respectivamente. Dentro de los niveles de escolaridad y para el caso de las madres, los mayores Épsilon son para el nivel primaria y carrera técnica. Y en el caso de los padres, son los niveles de primaria, secundaria y carrera técnica. Cabe mencionar que las madres tienen mayores desviaciones estándar que los padres.
- Los alumnos que no tienen acceso a equipo de cómputo tienen un valor igual a 2.06 desviaciones estándar. Resulta controversial que los alumnos que no tienen acceso a equipo de cómputo sean los que tengan un mayor Épsilon, a pesar de estar en una modalidad de estudios a distancia.
- El que los alumnos entiendan, hablen y escriban el idioma inglés con dificultad o bien no hace mayor diferencia entre los valores de Épsilon.

Antecedentes académicos

- Dentro del descriptor nivel máximo de estudios, la categoría con mayor desviación estándar es para los alumnos que ya tienen un posgrado terminado.
- Los alumnos que obtuvieron en su último nivel de estudios un promedio igual o mayor a nueve tienen un mayor Épsilon, en comparación con otros promedios.
- Los alumnos que tiene más de ocho años sin estudiar tienen un Épsilon igual a 3.4.

Examen de ingreso

- De los 120 aciertos posibles que pueden obtener los alumnos en el examen de admisión, las categorías que tienen Épsilon positivos, son: 33 a 40, 49 a 56, 65 a 72 y 81 a 96. En promedio se calculó 0.34 desviaciones estándar.

Es importante señalar que así como los Épsilon positivos suman hacia el cumplimiento de la clase, los Épsilon negativos aportan hacia el no cumplimiento de la clase.

Al tener las variables y las medidas de los Scores seleccionadas, es momento de asignar cada uno de éstos a las categorías correspondientes en los datos de prueba, es decir, el valor del Score obtenido con los datos de entrenamiento vendrá a sustituir la categoría que según corresponda en los datos de prueba. Acto seguido se suman todos los valores individuales de los Score pasa obtener un Score total (St) por cada instancia[5] y se ordenan de forma descendente con su respectiva clase (0/1[6]), ver siguiente cuadro.

Núm	Clase	St
1	1	5.014
2	1	4.948
3	1	4.561
..	..	..
865	0	0.000
866	0	0.000
867	0	0.000

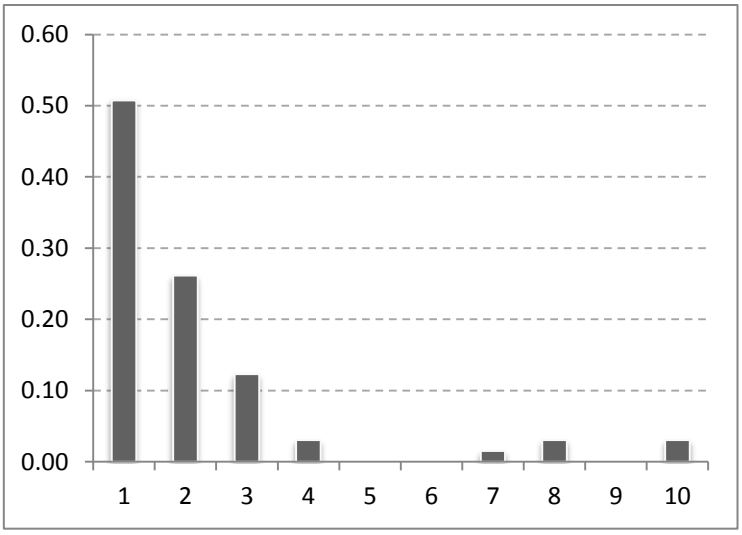
Cuadro 4. Clase y Score total

Con el cuadro cuatro completo, se construyó la siguiente tabla y gráfica del modelo clasificador por deciles, en la cual se presentan los conteos de la clase y su respectiva distribución del total de la clase para cada decil.

[5] *Cada instancia refiere a un alumno con sus datos de perfil de ingreso, antecedentes académicos y matrícula.*
[6] *0 = No clase y 1 = Sí clase; Clase: Alumnos con el 100% de créditos en tiempo curricular.*

Rango	Deciles	Conteos	Relativos
1 - 87	1	33	0.524
88 - 174	2	17	0.270
175- 261	3	8	0.127
262 - 348	4	2	0.032
349 - 435	5	0	0.000
436 - 522	6	0	0.000
523 - 609	7	1	0.016
610 - 696	8	2	0.032
697 - 783	9	0	0.000
784 - 870	10	0	0.000
Total		**63**	**1**

Cuadro 5. Conteos de la clase por deciles



Gráfica 1. Modelo clasificador

En los dos primeros deciles suman casi el 80% de los conteos de la clase, lo que demuestra que el modelo clasificador desarrollado con los datos de entrenamiento y utilizando el algoritmo Naive Bayes, sí cumple su función de predecir y discriminar a la sí clase de la no clase en los datos de prueba. Con el propósito de establecer los límites entre las cuatro categorías de la matriz de confusión: Verdadero positivo (Vp), Falso negativo (Fn), Verdadero negativo (Vn) y Falso positivo (Fp). La identificación de dicho umbral se ubica en el cambio de signo de los valores de St. Al ser cero el valor más pequeño de St, se le aplicó la siguiente constante de suavización. Con el propósito de medir el rendimiento del modelo clasificador se calculó su matriz de confusión. Para calcularla se requirió de la definición de un umbral en los datos de prueba con el propósito de establecer los límites entre las cuatro categorías de la matriz de confusión: Verdadero positivo (Vp), Falso negativo (Fn), Verdadero negativo (Vn) y Falso positivo (Fp). La identificación de dicho umbral se ubica en el cambio de signo de los valores de St. Al ser cero el valor más pequeño de St, se le aplicó la siguiente constante de suavización.

$$St\ ajustado = St + \ln\frac{Pc}{1 - Pc}$$

Con dicha constante de suavización para cada St, se logró identificar el cambio de signo en los nuevos valores de St ajustado, ver siguiente cuadro.

Núm	Clase	St	St ajustado
175	0	2.610	0.024
176	0	2.587	0.001
177	0	2.525	-0.062
178	0	2.511	-0.076

Cuadro 6. Umbral

El umbral se ubicó entre 0.001 y -0.062 de la columna St ajustado. Con la identificación del límite del umbral se obtiene la siguiente matriz de confusión y sus medidas de rendimiento: sensibilidad y especificidad.

<table>
<tr><th>Matriz de confusión</th><th colspan="2">Medidas de rendimiento</th></tr>
<tr><td rowspan="2">
<table>
<tr><th>Vp</th><th>Fp</th></tr>
<tr><td>50</td><td>126</td></tr>
<tr><th>Vn</th><th>Fn</th></tr>
<tr><td>678</td><td>13</td></tr>
</table>
Dónde:

Vp: Todos los elementos sí clase por encima del umbral
Fp: Todos los elementos no clase por encima del umbral
Vn: Todos los elementos no clase por debajo del umbral
Fn: Todos los elementos sí clase por debajo del umbral

(Vp)+(Fn) = Sí clase; (Fp)+(Vn) = No clase
</td><td>Sensibilidad</td><td>$\frac{Vp}{Vp+Fn} = 0.793$</td></tr>
<tr><td>Especificidad</td><td>$\frac{Vn}{Vn+Fp} = 0.843$</td></tr>
</table>

Cuadro 7. Matriz de confusión y medidas de rendimiento

Analizando las medidas de rendimiento, podemos observar la probabilidad que tiene el modelo para clasificar correctamente a la sí clase es del 79.3% y la probabilidad de clasificar correctamente a la no clase es igual al 84.3%.

Además de las matrices de confusión, los gráficos ROC (Receiver Operating Characteristic) son una excelente técnica para examinar el rendimiento de un clasificador binario (Swets, 1988). Un gráfico ROC muestra para cada iteración en el umbral, la distribución de los diferentes pares (falsos positivos y verdaderos positivos). El punto (0,1) demuestra una clasificación perfecta, es decir, clasifica correctamente todos los verdaderos positivos y ningún falso positivo. En contraste el punto (1,0) demuestra un pésimo modelo de clasificación. A continuación de muestra la curva ROC[7] que determina el rendimiento de nuestro modelo clasificador.

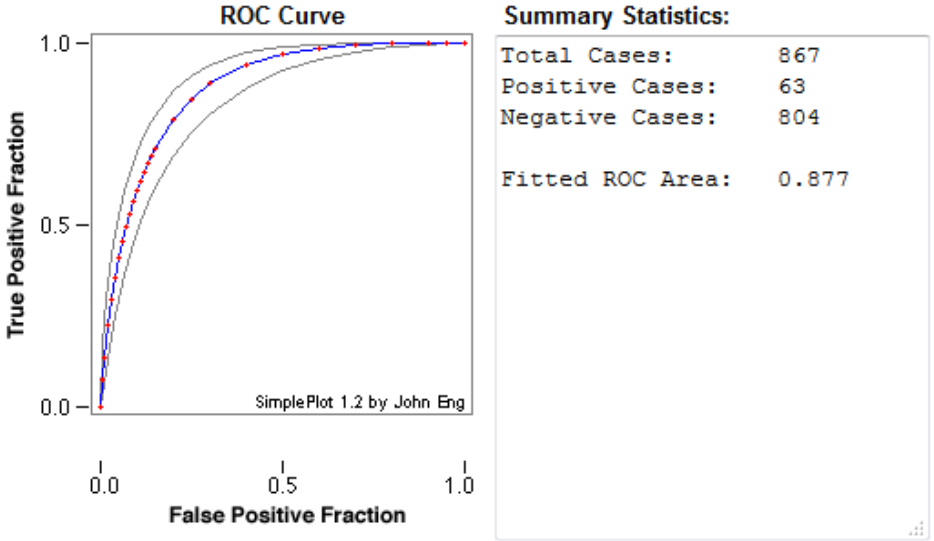


Gráfico 2. Curva ROC

Además de ver que la forma de la curva ROC tiene una mayor aproximación hacía la esquina superior izquierda (0,1), una medida importante es el área debajo la curva (AUC), la cual a través de un valor escalar entre 0 y 1, determina el rendimiento del

[7] *El gráfico se obtuvo con la calculadora web de análisis ROC de la escuela de medicina de la Universidad de Johns Hopkins. URL: http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html, consultado 15 de abril de 2016.*

modelo clasificador. El área debajo la curva es el equivalente a la probabilidad que tiene el clasificador para seleccionar aleatoriamente una instancia positiva y ubicarla por encima de una instancia negativa (Fawcett, 2006).

En nuestro caso particular, el área debajo la curva del gráfico ROC, es igual a 0.877 y hace notar la buena capacidad predictiva de nuestro modelo.

Conclusiones

Siempre será de interés para las instituciones educativas conocer los factores que motivaron en los alumnos a alcanzar la totalidad de sus créditos en tiempo curricular (10 semestres) y sobre todo, tener la posibilidad de replicar el modelo de Minería de Datos para las siguientes generaciones y así, pronosticar quién estará en posibilidades de terminar en 10 semestres, y en consecuencia sabremos que tendrán el 94% de probabilidades de tener una calificación mayor o igual a nueve (ver introducción).

El analizar los datos de los alumnos con técnicas de minería de datos permitió desarrollar un modelo clasificador para la modalidad a distancia del SUAyED, el cual predice al término del primer semestre en la licenciatura, los alumnos que cumplirán con el 100% de créditos en diez semestres. Además, identifica en menores tiempos a los alumnos que no terminarán en diez semestres la totalidad de sus estudios, lo que permitirá desarrollar estrategias de retención en corto tiempo.

Así como las tasas de deserción son consideradas como un indicador de calidad en la educación a distancia (Lykourentzou, 2009; Willging, 2004), el que los alumnos cumplan con sus créditos en tiempos curriculares sin duda ayudará a mejorar los indicadores de la calidad.

El algoritmo de clasificación Naive Bayes ha demostrado ser un buen predictor para identificar los factores de éxito de los alumnos. Tal como lo mostró el modelo clasificador, discriminó correctamente el 52% y 27% de la sí clase en los dos primeros deciles, respectivamente. El rendimiento calculado con la curva ROC demostró tener una capacidad predictiva igual al 87% utilizando un análisis Feature Selection en comparación con incluir todas las categorías al modelo (0.847).

Con la ayuda de Épsilon, se logró identificar los principales predictores que contribuyen al éxito de los alumnos. Entre los más representativos se señala la importancia que los alumnos aprueben al menos cinco asignaturas en su primer semestre de estudios con un promedio mayor o igual a ocho. También se descubrió que los alumnos con edades de ingreso mayores a los 43 años, inscritos en la licenciatura en derecho y al iniciar sus estudios respondieron lo siguiente: ser casados, tener hijos, trabajar de 11 a 20 horas semanales, no contar con equipo de cómputo y tener más de ocho años sin estudiar. Son los alumnos que tienen mayores probabilidades de cumplir el 100% de créditos en diez semestres.

Sin duda, las técnicas de Minería de Datos aplicadas con método nos posibilitarán mejores análisis y entendimientos de nuestros datos, de modo que permita la extracción de conocimiento y así, tener más claro los factores de éxito que inciden en el desempeño académico de los alumnos para un tiempo curricular. Como lo menciona Zhao (2006) "*En los procesos de minería de datos usualmente se encuentra el*

elemento de serendipia". La minería de datos está diseñada para descubrir interesantes y a menudo inesperadas relaciones y estructuras que previamente eran desconocidas entre las variables.

Bibliografía

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. DOI: 10.1016/j.patrec.2005.10.010

Gayol, Y. (2016). Deserción: Explorando resultados en la educación a distancia. Disponible en: http://u2000.com.mx/PDFs/U2000_937_web.pdf, (2016, 15 de abril, 12:40).

Kotsiantis, S; Pierrakeas, C. y Pintelas, P. (2003). *Preventing student dropout in distance learning systems using machine learning techniques*. In Knowledge-Based Intelligent Information and Engineering Systems, (pp. 267–274).

Romero, C., S. Ventura, P.G. P.G. Espejo, and C. Hervas. (2008). Data mining algorithms to classify students. In Educational data mining 2008: Proceedings of the 1st international conference on educational data mining, 8–17.

Stephens, C.R; Heau, J.G; González, C; Ibarra-Cerdeña, C.N; Sánchez-Cordero, V; et al. (2009) Using Biotic Interaction Networks for Prediction in Biodiversity and Emerging Diseases. PLoS ONE 4(5): e5725. doi: 10.1371/journal.pone.0005725

Swet, J. (1988). Measuring the accuracy of diagnostic systems. Science 240, 1285-1293.

Willging, P.A. and Johnson, S.D. (2004). Factors that influence students' decision to dropout of online courses. In Journal of Asynchronous Learning Networks, 8(4), (pp. 105–118).

Yukselturk, E; Ozekes S. y Türel Y. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. European Journal of Open, Distance and E-Learning. Volume 17, Issue 1, Pages 118–133, ISSN (Online) 1027-5207.

Zang, W. and Lin, F. (2003). *Investigation of webbased teaching and learning by boosting algorithms*. In Proceedings of IEEE International Conference on Information Technology: Research and Education, 2003, (pp. 445449).

Zhao, C. y Luan, J. (2006). *Data mining: Going beyond traditional statistics*. In New Directions for Institutional Research, 131(2), (pp. 716).