Statistical Learning to Detect Potential Dropouts in Higher Education: A Public University Case Study

César Noboa[1], Milton Ordóñez[2], Jorge Magallanes[3]
[1] Escuela Superior Politécnica del Litoral, ESPOL, (Oficina de Admisiones), Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
enoboa@espol.edu.ec
[2] Escuela Superior Politécnica del Litoral, ESPOL, (Gerencia de Tecnologías y Sistemas de Información), Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
xordonez@espol.edu.ec
[3] Escuela Superior Politécnica del Litoral, ESPOL, (Facultad de Ingeniería en Electricidad y Computación), Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
jmagalla@espol.edu.ec

Abstract. Early detection of students tending to drop out is crucial to improve not only graduation rates but also education quality. By using basic statistical learning techniques, this work presents a simple way to predict possible dropouts based on their demographic and academic characteristics. In order to reasonably predict while gaining a better understanding of the dropout phenomenon, after a preliminary analysis, 4 classification methods are applied including 2 easy-to-interpret ones. Some of the main results of this study show that almost 22% of current students are potential dropouts while being an older student and failing many subjects tend to cause dropout; on the other hand, passing more than 12 subjects and long-term access to library materials can prevent students from leaving college.

Palabras Clave: College Dropout, Early Dropout Prediction, Dropout Risk Factors, Statistical Learning Classification Techniques.

1 Introducción

En la educación superior, la deserción estudiantil es un problema relevante, no sólo en América Latina sino en países desarrollados. Aunque no existe consenso para medir la calidad de la educación, uno de los indicadores importantes es el tiempo de titulación para graduarse (TTG), el cual está relacionado directamente con la deserción estudiantil [1]. Las estimaciones a nivel mundial sitúan esta tasa de deserción en el 40% [2]. En Estados Unidos, esta tasa es de alrededor del 30% y representa una pérdida de 9 billones de dólares en la educación de estos estudiantes [3]. Sin embargo, la deserción no sólo afecta a la calidad de la educación y a la economía de un país, sino que tiene repercusiones sobre el desarrollo de la sociedad, puesto que ésta demanda las contribuciones derivadas de la población con educación superior como son: la innovación, la producción de conocimiento y el descubrimiento científico [1].

Existen varias investigaciones que determinan en alguna medida la deserción en América Latina. En la gran mayoría se trata acerca de la determinación de los factores que conllevan a la deserción, la medición del número de desertores y los mecanismos para disminuirlo [4]. Existen dos propuestas para la cuantificación de la deserción: La primera, se establece como la proporción de estudiantes que se titulan en un tiempo determinado que corresponde a la duración de la carrera; y la segunda, simplemente es el número de estudiantes que abandona sus estudios. Para disminuir la deserción, estas investigaciones proponen mejorar los mecanismos de detección temprana de potenciales desertores.

La aplicación de los métodos de aprendizaje estadístico para abordar el problema de la deserción ya ha sido propuesta por varios estudios, analizando ya sea, la deserción o culminación de un curso [5,6] o de una carrera [3,7,8,9]. Algunos de los métodos empleados en estas investigaciones son: regresión logística, k-vecinos más cercanos, árboles de decisión incluido random forests, redes bayesianas, redes neuronales, entre otros. En el presente trabajo, se ha preferido mantener un equilibrio entre facilidad de interpretación y precisión [10], poniendo especial énfasis en la detección de desertores antes que en la reducción de malas clasificaciones. Se han escogido 2 métodos que generan modelos comprensibles: árboles de decisión y regresión logística; y 2 métodos que tienen gran capacidad de precisión: naive bayes y k-vecinos más cercanos. Estos 4 métodos empleados de manera conjunta producirán una solución de compromiso entre comprensibilidad y precisión, siendo esta última evaluada principalmente por el porcentaje de desertores detectados.

En este artículo, se presenta un marco de trabajo para los sistemas de detección temprana de potenciales desertores, en el cual se propone utilizar las 4 técnicas ya mencionadas de manera conjunta. Para medir la efectividad del marco de trabajo, estas técnicas son aplicadas sobre el conjunto de datos del Sistema Académico de la Escuela Superior Politécnica del Litoral (ESPOL), institución pública de educación superior del Ecuador.

2 Metodología

2.1 Análisis preliminar

En esencia este proyecto trata de comparar las características de los estudiantes desertores versus las características de aquellos que no lo son, para luego de esto definir una regla o modelo que los diferencie. En este trabajo, un estudiante es considerado desertor si ha dejado de estudiar los últimos 3 años (2015, 2016, y 2017) y no se ha graduado.

En la Fig. 1 se observa la tendencia del porcentaje de deserción de los estudiantes que ingresaron a la ESPOL en el período 2009-2013 clasificados por género. Este porcentaje de deserción ha ido decreciendo hasta situarse en el 10.56% para las mujeres y el 14.48% para los hombres; sin embargo, aún son porcentajes altos, especialmente, si se los traduce a cantidades absolutas.

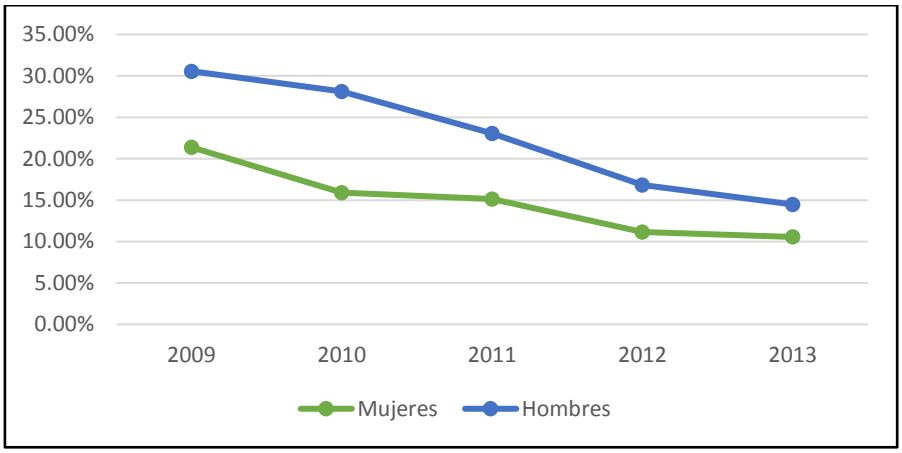

Fig. 1. Comportamiento de la deserción por año y sexo.

La Fig. 1 también refleja que existe una considerable diferencia en la deserción de ambos grupos. Ésta muestra no sólo que ha existido mayor porcentaje de desertores varones que de mujeres, sino que la brecha entre estos 2 grupos se ha acortado, pero siempre ha existido. La Tabla 1 muestra la comparación porcentual de desertores según su género. El estadístico χ^2 de Pearson para la prueba de independencia de esta tabla de contingencia es 78.96 con un valor p< 0.0001, lo cual indica que la deserción y el género no son independientes. Sin embargo, esto último no implica que el género tiene la capacidad suficiente para discriminar entre estudiantes desertores y no desertores.

Tabla 1. Deserción de estudiantes que ingresaron desde el 2009 al 2013.

	Femenino		Masculino		Todos	
	Cantidad	**%**	**Cantidad**	**%**	**Cantidad**	**%**
Desertor	498	14.75	1,105	22.62	1,603	19.40
No Desertor	2,878	85.25	3,781	77.38	6,659	80.60
Totales	3,376	100.00	4,886	100.00	8,262	100.00

2.2 Selección del conjunto de datos objetivo

Para la selección del conjunto de datos objetivo, se tomará como "instante de tiempo" el segundo semestre del año 2011. En concreto, el conjunto de datos objetivo lo conforman los estudiantes que ingresaron a la ESPOL desde el año 2009 y estudiaron en el semestre 2011-2s con las características que tuvieron en ese instante de tiempo. Este conjunto de datos consta de 4294 estudiantes de los cuales 525 son desertores.

El conjunto de variables seleccionadas se divide en 2 grupos: las variables relacionadas a las características personales del estudiante y las variables relacionadas a su comportamiento académico. En la Tabla 2 se indica la descripción de cada una de estas variables. El período de prueba mencionado en las variables "Superadas" y "Perdidas" se refiere al semestre en que un estudiante tiene la última oportunidad de

aprobar una materia luego de haberla reprobado 2 veces en semestres anteriores; reprobar dicha materia en el período a prueba restringe al estudiante de continuar en la misma carrera.

Tabla 2. Descripción de las variables seleccionadas.

#	Variable	Descripción	Tipo	Posibles valores
1	SEXO	Sexo del estudiante	Categórica	{F, M}
2	EDAD	Edad del estudiante	Numérica	16 en adelante
3	FACTOR_P	Indicador del nivel socioeconómico	Numérica	0 a 40
4	RESIDENCIA	Tipo de residencia	Categórica	{LOCAL,PROV}
5	APROBADAS	# de materias aprobadas	Numérica	0 en adelante
6	REPROBADAS	# de materias reprobadas	Numérica	0 en adelante
7	PROMEDIO	Promedio general	Numérica	0 a 10
8	ANTIGÜEDAD	# de semestres de estudio	Numérica	0 a 5
9	PERDIDAS	# de veces en que perdió un período de prueba	Numérica	Desde 0 en adelante
10	SUPERADAS	# de veces en que superó un período de prueba	Numérica	Desde 0 en adelante
11	T_AUTONOMO	# de días de consulta de material de la biblioteca principal en el semestre actual	Numérica	Desde 0 en adelante
12	DESERTOR	Etiqueta que indica si el estudiante ha o no ha desertado	Categórica, variable respuesta	{SI,NO}

2.3 Entrenamiento de los modelos de clasificación

Para la aplicación de las técnicas se toma el 70% de los datos para entrenamiento y el 30% restante para pruebas. Se toman varias muestras aleatorias con el esquema 70-30. Posterior a la generación de los modelos con las muestras de entrenamiento, se evalúa la precisión de los modelos.

Los métodos aplicados que presentan modelos fáciles de interpretar son: árbol de decisión y regresión logística. Ambos métodos permiten determinar las variables que tienen mayor influencia en la deserción universitaria. En la Fig. 2 se observa el árbol de decisión obtenido con una de las muestras.

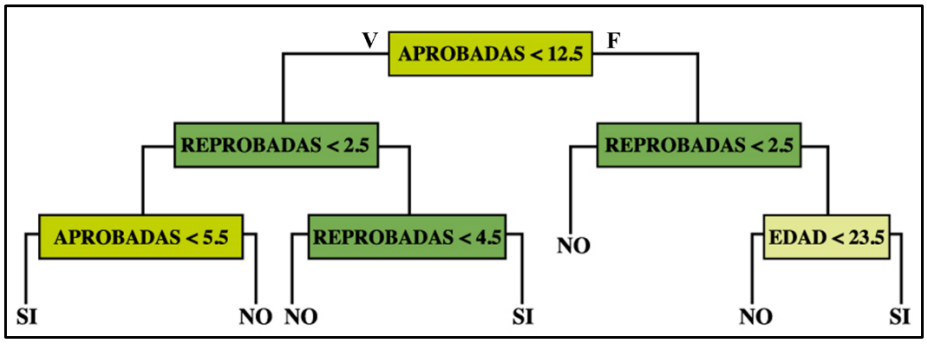

Fig. 2. Árbol de decisión obtenido de una muestra.

Los árboles de decisión pueden presentar reglas incidentales que carecen de generalidad, tomarlas en cuenta conduciría al efecto conocido como sobreajuste [11]. Luego del entrenamiento con las distintas muestras se obtiene la siguiente regla general:

Si (APROBADAS < 12.5 y REPROBADAS > 4.5)
entonces ES DESERTOR
caso contrario NO ES DESERTOR

La regresión logística es un método de clasificación que permite predecir la probabilidad de deserción del estudiante.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9148  -0.5105  -0.3360  -0.1696   3.2449

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.073171   0.565245  -7.206 5.76e-13 ***
SEXOM           0.083466   0.133207   0.627 0.530927
EDAD            0.202932   0.026429   7.678 1.61e-14 ***
FACTOR_P       -0.005980   0.009738  -0.614 0.539188
RESIDENCIAPROV -0.249330   0.228413  -1.092 0.275021
APROBADAS      -0.104725   0.011322  -9.250  < 2e-16 ***
PROMEDIO       -0.106775   0.058178  -1.835 0.066460 .
REPROBADAS      0.232576   0.026603   8.743  < 2e-16 ***
PERDIDAS        1.517568   0.643898   2.357 0.018431 *
SUPERADAS      -0.152527   0.143576  -1.062 0.288081
T_AUTONOMO     -0.006271   0.001876  -3.343 0.000828 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2277.8  on 3005  degrees of freedom
Residual deviance: 1812.6  on 2995  degrees of freedom
AIC: 1834.6
```

Fig. 3. Reporte de regresión logística de una muestra, generado con el programa estadístico R.

Los resultados de la aplicación de este método a una de las muestras de entrenamiento se exponen en la Fig. 3. Tal como se observa, las variables pro-deserción son edad y número de materias reprobadas con coeficientes de 0.2 y 0.23 respectivamente; mientras que las variables que podrían evitar la deserción son el número de materias aprobadas y el trabajo autónomo del estudiante con coeficientes de -0.1 y -0.006 respectivamente; todas estas variables de influencia con un valor p menor a 0.0001. Los otros 2 métodos que se aplican al conjunto de datos son: K-vecinos más cercanos y Naive Bayes. Estos métodos que se conocen como métodos retardados, no siempre generan un modelo explícito a la manera del árbol de decisión o la regresión logística y emplean el mayor tiempo de procesamiento cuando son consultados acerca de la clasificación de un nuevo elemento [12].

2.4 Validación de los modelos de clasificación

El principal interés es predecir con razonable precisión la tasa de deserción de un conjunto de estudiantes, es por esto, que el porcentaje de correctas clasificaciones no es muy utilizado como medida de evaluación.

Aplicando validación cruzada 10-fold para los métodos naive bayes y regresión logística se obtienen porcentajes de detección promedio de 28.57% y 28.56% respectivamente.

La técnica de validación cruzada es especialmente útil para la determinación del valor idóneo de k, en el método de los k-vecinos más cercanos. En este caso, se emplea la validación cruzada Leave-One-Out, que consiste en tomar todos los elementos excepto uno para entrenar el modelo, siendo el elemento sobrante empleado para la prueba [13]. Los resultados de esta validación se muestran en la Fig. 4, se observa que el valor de k que produce el mayor porcentaje de detección es k=2. Sin embargo, no es conveniente elegir este valor, pues en caso de existir un vecino desertor y otro no-desertor no se podría determinar la clase a la que pertenece el estudiante que se requiere clasificar. Por lo tanto, el k idóneo para este modelo es k=1, con la ventaja adicional de que el costo computacional es menor.

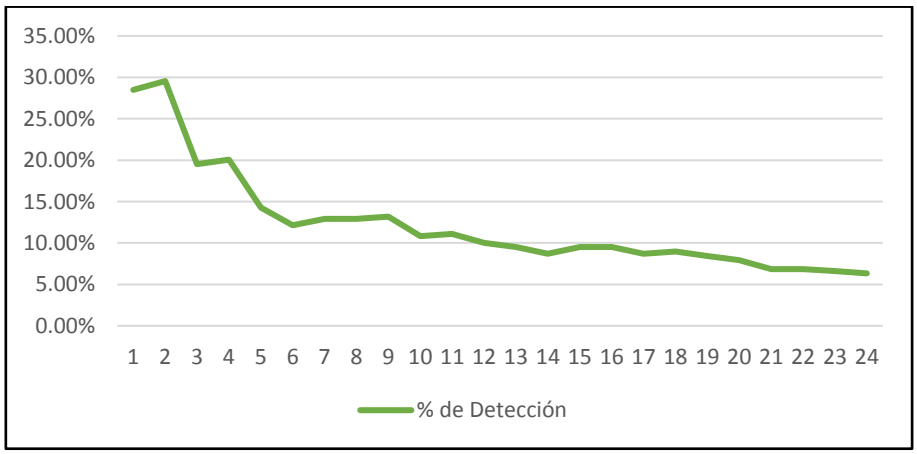

Fig. 4. Promedio del % de detección de k-nn para distintos valores de k.

3 Resultados y discusión

En esta sección, se presentan los resultados de evaluar cada método con 5 conjuntos de prueba. Al evaluar cada método se podrá obtener las tasas de: malas clasificaciones, falsos positivos, falsos negativos y detección.

En el caso de regresión logística es común etiquetar a un nuevo estudiante como desertor si la probabilidad de deserción que se obtiene es mayor que 0.5. Sin embargo, valores menores para este umbral disminuyen los falsos negativos, aunque en contraparte aumentan los falsos positivos. A este respecto, la Fig. 5 muestra el comportamiento de las distintas tasas al variar el umbral. La selección del umbral no es del todo objetiva; depende, en gran medida, de los recursos disponibles de la institución para atender a los falsos positivos. De acuerdo a la figura, un valor para el umbral pudiera ser 0.3 ó 0.4, ya que se obtiene un porcentaje de detección mayor al 30% con un porcentaje de falsos positivos menor al 10%. Es interesante notar que el porcentaje de malas clasificaciones (% de error) varía muy poco para valores del umbral entre 0.3 y 0.8.

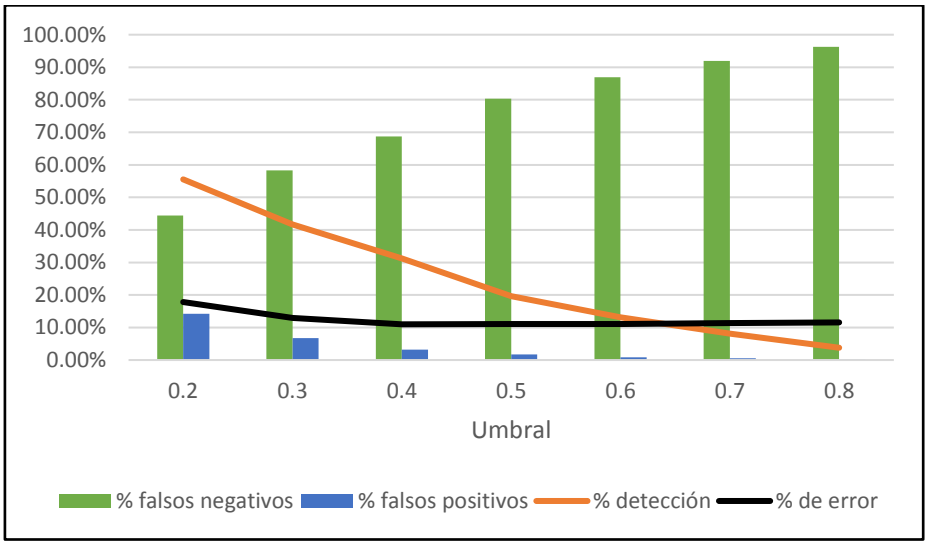

Fig. 5. Tendencias de los principales indicadores de regresión logística, promedio.

La Tabla 3 muestra un resumen de la prueba de cada uno de los métodos. En esta tabla se observan los mejores y peores resultados por cada método, siendo el método con mejor porcentaje de detección promedio, la regresión logística con umbral 0.4.

Tabla 3. Porcentajes de detección por método y por muestra.

Métodos	Árbol	Knn1	Naive Bayes	Logit (0.4)
muestra 1	26.03%	30.14%	39.04%	39.04%
muestra 2	20.00%	33.33%	34.67%	31.33%
muestra 3	14.81%	30.25%	26.54%	27.16%
muestra 4	19.38%	34.38%	27.50%	28.75%
muestra 5	18.37%	25.85%	27.21%	29.93%
Promedio	19.72%	30.79%	30.99%	31.24%

Para encontrar el porcentaje de detección global de desertores, es necesario aplicar de forma secuencial cada método e ir contabilizando los nuevos desertores que surgen.

La Tabla 4 muestra el resumen final de la aplicación sucesiva de los 4 métodos del proyecto. Tal como se observa, la estimación de la capacidad del proyecto para predecir la deserción universitaria es mayor al 52% y su precisión para clasificar correctamente es mayor al 82%.

Tabla 4. Porcentaje de detección acumulado y precisión por muestra.

Métodos	Árbol	Knn1	Naive Bayes	Logit (0.4)	Precisión
muestra 1	26.03%	41.78%	55.48%	57.53%	84.24%
muestra 2	20.00%	39.33%	52.66%	53.33%	82.14%
muestra 3	14.81%	34.57%	47.53%	48.77%	80.20%
muestra 4	19.38%	32.50%	41.88%	49.38%	81.44%
muestra 5	18.37%	36.73%	48.98%	51.70%	82.69%
Promedio	19.72%	36.98%	49.31%	52.14%	82.14%

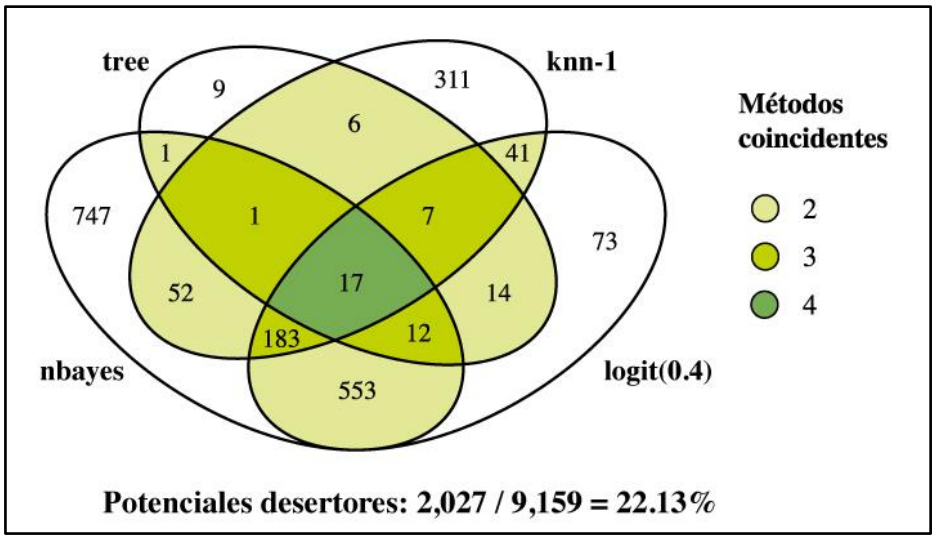

Fig. 6. Diagrama de Venn de posibles desertores 2017-2S.

Finalmente, se aplican los métodos entrenados al conjunto de 9,159 estudiantes registrados en el segundo semestre del año 2017, de los cuales por supuesto se desconoce su futura deserción. La Fig. 6 muestra los potenciales desertores empleando los 4 métodos. De acuerdo a esto, más del 22% de los estudiantes posiblemente desertarán. Es notorio que los métodos incomprensibles naive bayes y knn-1 agregan gran cantidad de desertores que los otros 2 métodos no detectan (311 + 747 vs. 9 + 73).

4 Conclusiones y trabajos futuros

De manera general, los resultados obtenidos muestran que mientras más avanza el estudiante en sus estudios menos probable es su deserción; y a excepción de la edad, las características personales de los estudiantes poco inciden en su retiro de la universidad.

De acuerdo al método de árbol de decisión, reprobar más de 4 materias en las primeras fases de la carrera contribuye significativamente a la deserción. En el caso de la regresión logística, las variables que más contribuyen a la deserción son: la edad y la cantidad de materias reprobadas. En promedio, se obtuvo que los estudiantes con mayor edad tienen 22% más posibilidades (odds) de desertar frente a los que son un año menor; y, por cada materia reprobada las posibilidades de desertar frente a no hacerlo se incrementan en un 28%; en cambio, cada materia aprobada reduce la razón entre la probabilidad de desertar versus no desertar en un 16% y la consulta de material bibliográfico reduce esta misma razón en 1% por cada día de consulta.

Puesto que las variables "perdidas" y "superadas" no influyen en la deserción, se concluye que reprobar una materia estando a prueba no es garantía de deserción universitaria; así como superar un período de prueba tampoco implica mayor resiliencia en los estudios.

Luego de los experimentos realizados, se estima que la capacidad promedio del proyecto para detectar un posible desertor es mayor al 52%; y, la capacidad promedio para clasificar a un estudiante en el grupo correcto es mayor al 82%.

En el caso de la predicción sobre los datos actuales, al aplicar los 4 métodos de discriminación a los 9,159 estudiantes registrados en el segundo semestre del 2017, se obtuvo que alrededor del 22% de los estudiantes fueron detectados como posibles desertores por al menos uno de los métodos; mientras que 220 estudiantes fueron detectados por más de 2 métodos aumentando así su riesgo de deserción.

Los resultados preliminares obtenidos en este artículo indican que el proceso de enseñanza-aprendizaje pudiera verse beneficiado al enfocarse en los estudiantes detectados como posibles desertores, permitiendo que éstos tengan mayor acceso no sólo a material bibliográfico especializado sino a mejores oportunidades de incrementar su trabajo autónomo favoreciendo así su aprendizaje activo. Algunos de los siguientes pasos para potenciar estos resultados serían, estimar el tiempo que tienen los directivos antes de que el estudiante deserte, como se calcula en [3]; y, la incorporación al análisis de aspectos no cognitivos, como se sugiere en [7]. También un análisis longitudinal semestre a semestre para obtener la precisión promedio para detectar potenciales desertores, la incorporación de otros métodos como SVM para incrementar la capacidad de detección y el aumento en la recolección de datos relativos al trabajo autónomo del

estudiante que va más allá de consultas bibliográficas, son algunas de las propuestas para futuras investigaciones.

Referencias

1. Ferreyra, M.; Avitabile, C.; Botero, J.; Haimovich, F.; Urzúa, S.: *Momento decisivo La educación superior en América Latina y el Caribe Resumen*. Grupo Banco Mundial (2017)
2. El Telégrafo: La deserción universitaria bordea el 40%. https://www.eltelegrafo.com.ec/noticias/sociedad/4/la-desercion-universitaria-bordea-el-40 (2016). Accedido el 19 de Mayo de 2018
3. Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J.: Predicting student dropout in higher education. *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, pp. 16-20 (2016)
4. Munizaga, F.; Cifuentes, B.; Beltrán, A.: Retención y Abandono Estudiantil en la Educación Superior Universitaria en América Latina y el Caribe: Una Revisión Sistemática. *Education policy analysis archives*, Vol. 26, No. 61, pp. 1-36 (2018)
5. Oedaa, S.; Hashimoto, G.: Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes. Procedia Computer Science, Vol. 112, pp. 614-621 (2017)
6. Badr, G.; Algobail, A.; Almutairi, H.; Almutery, M.: Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. Procedia Computer Science, Vol. 82, pp. 80-89 (2016)
7. Hutt, S.; Gardener, M.; Kamentz, D.; Duckworth, A.; D'Mello, S.: Prospectively Predicting 4-year College Graduation from Student Applications. Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 280-289 (2018)
8. Ahuja, R.; Kankane, Y.: Predicting the probability of student's degree completion by using different data mining techniques. Fourth International Conference on Image Information Processing (ICIIP), pp. 1-4 (2017)
9. Martins, L.; Carvalho, R.; Victorino, C.; Holanda, M.: Early Prediction of College Attrition Using Data Mining. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078 (2017)
10. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.: *An Introduction to Statistical Learning*. Springer 7th Ed, pp. 25 (2014)
11. Russell, S.; Norvig, P.: *Artificial Intelligence A Modern Approach*. Pearson Education 3rd Ed, pp. 705 (2010)
12. Makhabel, B.: *Learning Data Mining with R*. Packt Publishing 1st Ed, pp. 143 (2015)
13. Witten, I.; Frank, E.; Hall, M.; Pal, C.: *Data Mining Practical Machine Learning Tools and Techniques*. Elsevier 4th Ed, pp. 167-169 (2016)