Predicción del desempeño académico usando técnicas de aprendizaje de máquinas

Policy Brief

Ferney J. Rodríguez Dueñas[1]
Hamadys L. Benavides Gutiérrez[2]
Álvaro J. Riascos Villegas[3]

Universidad de los Andes

Presentado a

Instituto Colombiano para la Evaluación de la Educación

2 de octubre de 2018

[1]Universidad de los Andes. Decano Facultad de Ciencias. Correo: frodrigu@uniandes.edu.co
[2]Universidad de los Andes. Facultad de Economía. Correo: hl.benavides@uniandes.edu.co
[3]Universidad de los Andes. Profesor asociado. Facultad de Economía. Correo: ariascos@uniandes.edu.co

1. Antecedentes

La capacidad de predecir el desempeño académico de los estudiantes que recién inician su vida universitaria es de vital importancia para las instituciones de educación superior. En particular, es relevante para definir los estándares de admisión y es una herramienta útil para generar alertas y un adecuado acompañamiento durante los primeros semestres. En este sentido, la mayoría de investigaciones en Colombia se han centrado en el uso de técnicas estadísticas tradicionales, sin contemplar las diversas variables sociodemográficas y económicas disponibles para los estudiantes y para los establecimientos de educación secundaria de origen, dejando información valiosa sin explotar. Un conjunto de herramientas que permite enriquecer este universo de metodologías es el uso de técnicas de minería de datos que permiten manejar grandes cantidades de información y descubrir patrones y relaciones ocultas para las técnicas tradicionales. Una de las nuevas aplicaciones de la minería de datos se encuentra en la educación.

La minería de datos en educación (EDM, por sus siglas en inglés) es una nueva disciplina fundamentada en las técnicas, métodos y algoritmos de la minería de datos para explorar datos del campo de la educación en aras de encontrar patrones y predicciones que permitan, entre otras cosas, caracterizar el comportamiento y el desempeño de los estudiantes (Peña, 2014). Según Kleinberg, Ludwig, Mullainathan, y Obermeyer (2015), las técnicas empíricas clásicas no están optimizadas para la predicción porque buscan estimaciones de parámetros insesgados, mientras que las herramientas de aprendizaje de máquinas fueron desarrolladas para maximizar los resultados de la predicción por medio del intercambio entre sesgo-varianza.

Entre los estudios más recientes de la minería de datos aplicado al ámbito del desempeño académico, se destacan: Osmanbegović y Suljić (2012) quienes comparan diferentes métodos y técnicas de predicción a partir de variables socio-demográficas de estudiantes del departamento de Economía de la Universidad de Tuzla, sus resultados demuestran que el clasificador *Naive Bayes* arroja resultados superiores en la predicción del desempeño académico. Por su parte, Ahmad, Ismail, y Aziz (2015) intentan predecir el rendimiento académico en el primer año de estudiantes de pregrado en Ciencias computacionales utilizando arboles de decisión, *Naive Bayes* y reglas de clasificación. Las variables tenidas en cuenta corresponden a variables demográficas y familiares, obteniendo una tasa de precisión de 71 %.

Khobragade y Mahadik (2015) usando datos de los estudiantes (calificaciones, características familiares, condiciones sociales, entre otras) logran identificar 11 características más importantes en el desempeño académico, alcanzando una tasa de precisión en la predicción del 87.12 % con el algoritmo *Naive Bayes*. Por otro lado, Badr, Algobail, Almutairi, y Almutery (2016) utilizan métodos de clasificación basados en reglas de asociación para construir un clasificador que permita evaluar el desempeño de estudiantes universitarios en cursos de programación, logrando tasas de precisión del 63 %-67 %.

Para estudios de pregrado y maestría, Hamsa, Indiradevi, y Kizhakkethottam (2016) desarrollan un modelo de predicción del éxito académico, usando árboles de decisión y el algoritmo genético *fuzzy*. Asimismo, Costa, Fonseca, Almeida, Ferreira, y Rego (2016) evalúan el poder de predicción de cuatro técnicas de minería de datos (*Support Vector Machine*, árboles de decisión, redes neuronales y *Naive Bayes*) con el objeto de predecir el fracaso académico en cursos de programación. Sus resultados muestran que las técnicas utilizadas fueron capaces de identificar tempranamente aquellos estudiantes con riesgo de fracasar, especialmente, la técnica de *Support Vector Machine* supera en términos de error de prueba a los otros métodos mencionados.

Después de la revisión bibliográfica se ha encontrado que a nivel nacional no se ha explorado una evaluación comparativa entre diferentes metodologías para predecir el riesgo de reprobar una materia de los estudiantes en los primeros semestres de la universidad.

2. Problema de Investigación y Objetivos

A pesar de que, en muchas instituciones la decisión de admisión está basada en el resultado de la prueba *Saber 11*, existen muchos factores relevantes adicionales que potencialmente podrían explicar el éxito de los estudiantes. Por ende, la pregunta que plantea esta proyecto es: ¿Cuáles son los factores que más contribuyen a predecir el éxito académico de los estudiantes admitidos en la Universidad de los Andes en el período 2015-2017?. Adicionalmente, se persigue responder los siguientes interrogantes:

¿Existe realmente un efecto significativo en los resultados individuales de las pruebas Saber 11 sobre el éxito académico universitario?, ¿Cuál es el efecto de las variables familiares en la

predicción del desempeño académico de los estudiantes de la Universidad de los Andes?, ¿Qué influencia generan las variables socio-económicas en la predicción de resultados académicos de los estudiantes?, ¿Juega algún papel las condiciones del establecimiento educativo de origen en la predicción del desempeño académico?, ¿Existe alguna mejora significativa entre los resultados de predicción de los métodos tradicionales y las técnicas de aprendizaje de máquinas?, Dada la predicción del modelo, ¿qué acciones se deben tomar para mejorar los niveles de repitencia y deserción?.

Buscando responder a estas preguntas, esta investigación tiene como objetivo general identificar las variables relevantes en la predicción del desempeño académico de los estudiantes admitidos en la Universidad de los Andes entre el período 2015-2017, usando técnicas de aprendizaje de máquinas.

3. Metodología

3.1. Base de Datos

Para esta investigación se utilizan dos tipos de fuentes de datos, la primera fue proporcionada por el Icfes y corresponde a datos de la prueba Saber 11 e información adicional sobre características sociodemográficas de los estudiantes y variables sobre los establecimientos educativos. De igual forma, se usa información del desempeño en los cursos básicos de ciencias y características académicas de los estudiantes de la Universidad de los Andes entre los años 2015 y 2017. Estos datos son suministrados por el centro de Alto Cómputo de la Facultad Ciencias[4].

3.2. Plan de Análisis

Para el objetivo de predicción del desempeño de los estudiantes en los cursos de ciencias básicas de la Universidad de los Andes, se proponen los siguientes métodos: Regresión Logística con Stepwise Selection, Regresión Logística con Regularización Lasso, *Boosting* de árboles, Máquinas de Soporte Vectorial y Redes Neuronales. El objetivo es comparar los clasificadores fuera de

[4]El HPC es administrado por el departamento de física de la UNIANDES.

muestra con la medida estándar área bajo la curva (AUC) del *Receiver Operating Characteristics (ROC)*. Posteriormente, en la técnica de mejor desempeño se identificaran las variables más relevantes para la predicción.

Los diferentes clasificadores son entrenados para predecir la aprobación o reprobación del curso que depende de la nota final. Los estudiantes reprueban si su nota es inferior a 3, de lo contrario aprueban el curso. Se realiza una clasificación binaria estándar para tres tipos de modelos: Ex-Ante (sin información sobre notas parciales), EP1 (con información de nota parcial 1) y Ex-Post (con información de nota parcial 1 y 2).

Asimismo, el preprocesamiento de variables consistió en tres pasos fundamentales:

1. La marca de aprobación fue construida como sigue:

$$y_i = \begin{cases} Aprobado : Final \geq 3.0 & y = 1 \\ Reprobado : Final < 3.0 & y = 0 \end{cases} \tag{1}$$

2. Los puntajes de las pruebas Saber 11 fueron estandarizados (sustrayendo la media y dividiendo por la desviación de toda la muestra de acuerdo el tipo de partición: entrenamiento o prueba) según el año de ingreso del estudiante. Esto permite evitar posibles efectos en la distribución de los puntajes, dada la inclusión de tres años (2015, 2016, 2017).

3. Transformación de variables por medio de la metodología Pesos de Evidencias (WoE, por sus siglas en inglés).

Para la construcción de los modelos se toma como muestra de entrenamiento, todos aquellos estudiantes que tomaron las asignaturas de cálculo I y II, física I y II, química I y álgebra lineal en los años 2015 y 2016, mientras que la muestra de prueba quedo conformada con observaciones pertenecientes al año 2017. Finalmente, un aspecto a considerar es la calibración de los parámetros de cada una de las técnicas utilizadas. Para esto se utilizó validación cruzada con 10 muestras.

4. Hallazgos

Los resultados obtenidos sugieren que los factores académicos, el contexto semestral y el factor geográfico son relevantes para la predicción del desempeño académico. Estos se encuentran presentes en todos los modelos.

Para el modelo Ex-Ante (sin calificaciones parciales) la métrica de desempeño AUC alcanza el 75.3 % usando la regresión lasso, que en términos de buenas prácticas en modelos de predicción se ubica dentro de un modelo predictivo bueno. Por su parte, gracias a la ganancia en información que proveen la inclusión de las calificaciones parciales, los modelos EP1 (con nota parcial 1) y Ex-Post (con nota parcial 1 y 2) se considera un modelo muy preciso (85.9 %) y un modelo excelente (93.3 %), respectivamente. A la luz del desempeño mostrado en otras investigaciones, estos resultados son consistentes con lo señalado por Shahiria, Husaina, y Rashida (2015) quienes indican que en términos de precisión y usando diferentes especificaciones, los resultados en modelos de predicción del desempeño académico usando algoritmos de aprendizaje de máquinas suelen tener una precisión por encima del 66 %. Por otra parte, es posible observar que el desempeño de la transformación WoE es superior, en términos de precisión y de parsimonia (modelos más simples), a otras formas de transformación evaluadas (transformación *dummies*-niveles y transformación categórica).

A continuación, se presentan las preguntas de investigación de la propuesta inicial y se responden puntualmente a la luz de los resultados obtenidos:

1. ¿Cuáles son los factores que más contribuyen a predecir el éxito académico de los estudiantes admitidos en la Universidad de los Andes en el período 2015-2017?

    - Es posible determinar que los factores que más inciden en el desempeño académico de los estudiantes corresponden en su mayoría al contexto semestral (repitencia, créditos totales registrados, promedio semestres anteriores,estado académicos, programa, asignatura, calificaciones parciales), a factores geográficos (municipio de residencia) y características personales como la edad. Es de resaltar que los resultados de la prueba saber 11, especialmente en el componente de matemáticas, son importantes cuando no se encuentra con ninguna información *a priori* de las calificación parcial 2 del

estudiante en la asignatura.

2. ¿Existe realmente un efecto significativo en los resultados individuales de las pruebas Saber 11 sobre el éxito académico universitario?

    - Al observar los resultados de las variables más importantes para cada modelo es posible determinar que los resultados de las pruebas saber 11 son importantes para predecir el éxito académico universitario en la medida en que no se cuente con mucha información de calificaciones parciales de las asignatura.

3. ¿Cuál es el efecto de las variables familiares en la predicción del desempeño académico de los estudiantes de la Universidad de los Andes?

    - Dentro de los resultados no se observó incidencia alguna de las variables familiares dentro de la predicción. Una posible explicación se podría encontrar en la cuasi-homogeneidad de los estudiantes de la muestra si consideramos que más del 60 % de los padres cuentan con una educación profesional y superior.

4. ¿Qué influencia generan las variables socioeconómicas en la predicción de resultados académicos de los estudiantes?

    - Teniendo en cuenta los resultados del mejor modelo de predicción, no es posible observar incidencia del factor socioeconomico. Nuevamente, es posible que la explicación se encuentre dentro de la cuasi-homogeneidad de la muestra, al considerar que más del 66 % de los estudiantes son de estrato 3 o superior.

5. ¿Juega algún papel las condiciones del establecimiento educativo de origen en la predicción del desempeño académico?

    - Dentro de las estimaciones de los modelos finales no se observó la participación de las característica de los establecimientos educativo de origen, esto puede ser causado por la poca variabilidad de estas variables: los estudiantes provienen en su mayoría (más del 70 %) de colegios académicos-no oficiales con jornada completa.

6. ¿Existe alguna mejora significativa entre los resultados de predicción de los métodos tradicionales y las técnicas de aprendizaje de máquinas?

- Si consideramos la regresión logística como técnica base, es posible observar que cuando se tiene información sobre calificaciones parciales de la asignatura no existe diferencia significativa alguna entre los resultados de predicción, sin embargo, si existe más eficiencia en el uso de la información en la regresión lasso para los modelos EP1 y Ex-Post. Por otro lado, cuando no se tiene información *a priori* de la evolución del estudiante en la asignatura, la diferencia en términos de AUC entre la técnica base y la mejor técnica de aprendizaje de máquinas (regresión lasso) es de 3 puntos con mejoras en eficiencia en la información.

7. Dada la predicción del modelo, ¿Qué acciones se deben tomar para mejorar los niveles de repitencia y deserción?

    - Teniendo en cuenta las variables incluidas en los modelos finales de predicción, las acciones al comienzo del semestre deben centrarse en la identificación de las categorías con mayor riesgo (en términos WoE) de las variables incluidas en el modelo Ex-Ante: municipio de origen del estudiante, el promedio obtenido en semestres anteriores, programa, estado académico actual, edad, estado de repitencia, examen de clasificación y el puntaje en matemáticas obtenido en la prueba Saber 11. En otras palabras, el estudiante presentará mayor riesgo de reprobar si: la asignatura que está tomando, dentro de las consideradas en esta investigación, es diferente a física 2, álgebra lineal y cálculo II, pertenece a los programas de ingenierías eléctrica, biomédica, ambiental, mecánica y sistemas y computación, administración, geociencias, química, arquitectura, contaduría, ciencias políticas, diseño, filosofía, gobierno y asuntos públicos, licenciatura en ciencias naturales, lenguas y cultura, medicina, microbiología, música y psicología, su municipio de residencia al presentar las pruebas Saber fue Ibagué, Neiva, Barranquilla y los contenidos en la categoría negativa de los Anexos de la investigación, se encuentra en el segundo semestre del año, ha tomado menos de 29 créditos en toda su carrera, tiene una edad diferente del rango de 16.8 y 18.6 años, nunca se ha encontrado en un estado académico diferente al normal, la calificación en el examen de clasificación de matemáticas estuvo entre 0 y 9, su promedio acumulado se encuentra entre 0 y 3.55 y su puntaje de matemáticas en las pruebas Saber es menor a 0.43 en términos

estandarizados.

Referencias

Ahmad, F., Ismail, N., y Aziz, A. (2015). The prediction of student's academic perfomance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415-6426.

Badr, G., Algobail, A., Almutairi, H., y Almutery, M. (2016). Predicting students' performance in university courses: A case study and tool in ksu mathematics department. *Procedia Computer Science*, *82*, 80-89.

Costa, E., Fonseca, B., Almeida, M., Ferreira, F., y Rego, J. (2016). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247-256.

Hamsa, H., Indiradevi, S., y Kizhakkethottam, J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*, *25*, 326-332.

Khobragade, L., y Mahadik, P. (2015). Students academic failure prediction using data mining. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(11), 290-298.

Kleinberg, J., Ludwig, J., Mullainathan, S., y Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review: Papers and Proceedings*, *105*(5), 491–495.

Osmanbegović, E., y Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review-Journal of Economics and Business*, *10*(1), 2-12.

Peña, A. (2014). *Educational data mining: Applications and trends* (Vol. 524). Switzerland: Springer International Publishing Switzerland.

Shahiria, A. M., Husaina, W., y Rashida, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *75*, 414-422.