Modelo de pronóstico de riesgo académico de los alumnos de pregrado de la Universidad Nacional de Ingeniería

Academic risk forecast model for undergraduate students of the National University of Engineering

Hernán Garrafa Aragón[1], Iván Soto Rodríguez[2]
[1]*Facultad de Ingeniería Económica, Estadística y Ciencias Sociales, Universidad Nacional de Ingeniería, Lima, Perú*
[2]*Facultad de Economía y Planificación, Universidad Nacional Agraria La Molina, Lima, Perú*
**E-mail: hgarrafa@uni.edu.pe*
Recibido (Received): 23/06/2020 **Aceptado** (Accepted): 07/10/2020 **Publicado** (Published): 15/12/2020

RESUMEN

El presente trabajo de investigación usó información no estructurada generada en las unidades académicas de la Universidad Nacional de Ingeniería, mediante el uso de técnicas de Machine Learning, a fin de predecir el nivel de riesgo académico de un estudiante. Las fases consideradas fueron:

- Fase 1: Construcción del datamart: En esta fase se realizó la integración de datos de las diferentes fuentes para construir el repositorio de datos objetivo, el cual se dividió en datos de entrenamiento y datos de prueba.
- Fase 2: Entrenamiento del modelo: Elaboración del modelo de entrenamiento basado en los datos del datamart, aplicando Máquina de Soporte Vectorial.
- Fase 3: Validación y prueba del modelo: Evaluación del modelo obtenido anteriormente, usando los datos de prueba del datamart.

***Palabras clave**: Machine learning, integración de datos, predicción.*

ABSTRACT

The present work, uses unstructured information in order to predict the academic risk of a student, making use of Machine Learning techniques. Phases:

- Construction of the datamart: The data from the different sources will be integrated to build the objective data repository, which will be divided into two: Training data and test data,
- Training of the model: This consists in elaborating the training model based on data from the datamart, applying vectorial support machine.
- Validation of the model: It consists of evaluating the model obtained previously, using the test data from the datamart.

***Keywords**: Machine learning, data integration, prediction*

1. INTRODUCCIÓN

En la UNI, un estudiante puede ser susceptible de estar en tres estados: Situación Normal, Situación de Observación académica, Separado; dependiendo ello, básicamente de su rendimiento académico (promedio ponderado semestral y acumulado). El caso más severo es que el estudiante puede ser separado de la universidad en una etapa en la que le falta pocos créditos para culminar sus estudios, con lo cual pierde la posibilidad de graduarse. Este hecho causa en el estudiante una huella de su fracaso.

Se tiene conocimiento que muchas veces la situación socio-económica, familiar y psicológica juega un rol importante en el riesgo académico. El propósito del estudio es utilizar las variables que dispone el estudiante en la UNI, a fin de poder construir un modelo predictivo de la posibilidad de riesgo académico y, según ello, realizar un acompañamiento tipo tutoría con el fin de que dicho estudiante no enfrente una situación de alto riesgo académico.

Usando la información histórica de la UNI, esta investigación estableció 3 grupos de riesgo académico: grupo 1=alumno en situación normal de estudios; grupo 2=alumno que está en la situación de "observación"; grupo 3=alumnos que está en la situación de "suspensión".

La UNI tiene información histórica de manera desagregada, en repositorios separados y, en algunos casos, de manera no estructurada. El trabajo de investigación usa la información académica histórica de la UNI de los ingresantes del año 2015, tales como: la información referida al puntaje de ingreso a la universidad mediante la modalidad del centro preuniversitario y mediante el examen ordinario, información que el postulante indica en una ficha socio-económica; información historial académica de notas en su etapa universitaria. Esta investigación permitirá establecer un modelo predictivo de un estudiante para cada uno de los grupos de riesgo académico: grupo 1=alumno en situación normal de estudios (sin riesgo académico), grupo 2=alumno que está en la situación de "observación" (riesgo académico moderado); grupo 3=alumnos que está en la situación de "suspensión" (alto riesgo académico). El modelo permitirá pronosticar la situación de riesgo futuro de un alumno nuevo o en uno que ha cursado los primeros 3 ciclos académicos. Mediante este modelo, se podrá efectuar un permanente seguimiento y acompañar al estudiante, a fin de aminorar la posibilidad de que sea integrado al grupo 3 (situación de riesgo académico).

2. DESARROLLO

La metodología se inicia con la recopilación de las notas obtenidas al ingresar; ya sea por la modalidad del examen directo del Centro Preuniversitario o por la modalidad del examen ordinario; luego, se registran los datos socio-económicos y el historial de notas de cada ingresante.

Para la elaboración del modelo predictivo se establecieron tres fases:

Fase 1:

Construcción del datamart: Consistió en la recopilación e integración de datos con el objetivo de obtener una base de datos completa, organizada y estructurada de los alumnos ingresantes en los ciclos académicos 2015 y el seguimiento de su desempeño durante 5 años, con el fin de que, a través de un posterior estudio, se puedan obtener conclusiones, mediante el uso de las diferentes variables obtenidas de esta misma base de datos. Este datamart se dividirá en dos: Datos de entrenamiento (70% del total de registros) y datos de prueba (30% del total de registros).

Fase 2:

Entrenamiento del modelo: el cual consistió en elaborar el modelo de entrenamiento basado en los datos del datamart, mediante la técnica Máquina de Soporte Vectorial y usando la librería de R e1071.

Fase 3:

Validación del modelo: Consistió en evaluar el modelo obtenido en la fase 2 usando los datos de prueba del datamart.

Fase 1: Construcción del datamart

Las variables recopiladas en las diferentes fuentes fueron:

Rendimiento Académico:

- Rendimiento académico del examen de ingreso por el Centro Preuniversitario
- Rendimiento académico del examen de ingreso ordinario
- Nota en el área de razonamiento verbal
- Nota en el área de razonamiento matemático
- Nota en aritmética
- Nota en álgebra
- Nota en trigonometría
- Nota en geometría
- Nota en física
- Nota en química
- Orden de mérito
- Carrera de ingreso

Situación socioeconómica:

- Tipo de colegio donde culminó la secundaria
- Tipo de preparación para la universidad

- Número de veces que postuló a la UNI
- Número de veces que postuló a otras universidades
- Qué concepto asocia más a la imagen de la UNI
- Edad
- Año de egreso del colegio
- Género

Resumen académico en la universidad:

- Promedio por curso
- Número de créditos por curso
- Nombre del curso
- Ciclo que llevó el curso
- Año que llevó el curso
- Código del profesor del curso
- Código de matrícula
- Especialidad
- Facultad

a) Variables dependientes:

Riesgo académico:

Situación normal, aquellos estudiantes cuyo promedio ponderado del último semestre cursado es mayor o igual a diez (10,00).

Situación de observación, aquellos estudiantes cuyo promedio ponderado del último semestre es menor de diez (10,00).

Situación de suspensión, aquellos estudiantes en situación de observación cuyo promedio ponderado de los dos últimos semestres fue menor de 10 (10,00).

La variable Riesgo académico se codificó en 1 (normal), 2 (observado) y (3) suspendido:

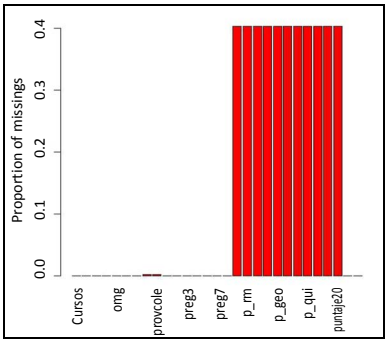
Riesgo académico	Descripción de estudiantes
Normal	No han tenido observaciones ni suspensiones en su historial.
Observado	Han tenido observaciones, pero no suspensiones en su historial.
Suspendido	Han tenido observaciones con suspensión en su historial.

Para realizar el procesamiento de datos se utilizó el programa R versión 3.4.1 que contiene paquetes y funciones que permiten aplicar la técnica Máquina de Soporte Vectorial.

3. PREPARACIÓN DE DATOS

En la Figura 1. se observó que, de los 661 registros simulados, el 40,2% presentaron valores ausentes en las variables relacionadas con las notas del examen de admisión (Razonamiento Verbal, Razonamiento Matemático, Álgebra, Aritmética, Geometría, Trigonometría, Física, Química y la nota final), y en las variables Años en ingresar y Provincia de colegio de secundaria. La mayoría de los registros (39,94%) presentaron valores ausentes en todas las variables relacionadas con el examen de admisión y el resto (0,3%) presentó valores ausentes adicionalmente en las variables Años en ingresar y Provincia de colegio de secundaria.

Figura 1. Proporción de valores perdidos.



Fuente: Elaboración propia

Con la finalidad de no perjudicar las conclusiones, se consideró conveniente omitir los 266 registros ausentes ya que reflejaron el 40,2% de todos los registros. Finalmente, quedaron 395 registros, los cuales presentaron la distribución según riesgo académico mostrado en la Tabla 1.

Tabla 1. Distribución de registros según riesgo académico.

Riesgo Académico	Observados	Porcentaje
Normal	177	44,81%
Observado	136	34,43%
Suspendido	82	20,76%

Fuente: Elaboración propia

Fase 2: Entrenamiento del modelo

Para la construcción de los clasificadores se tuvo que dividir el total de datos en una muestra de prueba (70%) y una muestra de entrenamiento (30%). Para ello se realizó un muestreo aleatorio simple al conjunto de 395 registros, los cuales presentaron la distribución según riesgo académico mostrado en la Tabla 2.

Tabla 2. Distribución de registros según muestra, tamaño y riesgo académico.

Muestra	Tamaño	Riesgo académico		
		Normal	Observado	Suspendido
Entrenamiento	276	42,8%	36,2%	21,0%
Prueba	119	49,6%	30,3%	20,2%

Fuente: Elaboración propia

Se utilizaron los clasificadores lineal y radial de la técnica Máquina de Soporte Vectorial. Los resultados obtenidos al clasificar la muestra de entrenamiento se presentan en la Tabla 3.

Tabla 3. Resultados de clasificación correcta según tipo de Máquina de Soporte Vectorial.

Clasificador	Riesgo académico			Total
	Normal	Observado	Suspendido	
Lineal	98,3%	85,0%	70,7%	87,7%
Radial	100,0%	99,0%	100,0%	99,6%

Fuente: Elaboración propia

En la Tabla 3 se observó que el clasificador radial presentó un mayor porcentaje de clasificación correcta (99,6%) frente al clasificador lineal (87,7%). Estos resultados fueron aún preliminares ya que se basaron en la muestra de entrenamiento.

Fase 3: Validación del modelo:

Para la evaluación de los clasificadores, se utilizó la muestra de prueba en la predicción del riesgo académico. En la Tabla 4 se muestran los resultados de la predicción correcta.

Tabla 4. Resultados de predicción correcta según tipo de Máquina de Soporte Vectorial.

Clasificador	Riesgo académico			Total
	Normal	Observado	Suspendido	
lineal	81,4%	58,3%	33,3%	64,7%
radial	93,2%	5,6%	0,0%	47,9%

Fuente: Elaboración propia

Según la Tabla 4, la Máquina de Soporte Vectorial de tipo lineal alcanzó mejores resultados en la predicción (64,7%) frente al tipo radial (47,9%). En el clasificador lineal, la categoría con mejor predicción fue la de riesgo normal, seguido por observado y suspendido. En el clasificador radial, solo se predijo correctamente las categorías de riesgo normal y observado; además, la mayor predicción correcta la obtuvo la primera categoría, seguida por una muy baja predicción correcta de la segunda categoría.

En la Tabla 5 se presentan las métricas calculadas a partir de la matriz de confusión con la finalidad de comparar y evaluar las Máquinas de Soporte Vectorial propuestas.

Tabla 5. Métricas para evaluar clasificadores.

Clasificador	Clasificación correcta	Área debajo de la curva ROC	Coeficiente Kappa
lineal	64,71%	0,7254	0,4239
radial	47,90%	0,5018	-0,0011

Fuente: Elaboración propia

Según la Tabla 5, la Máquina de Soporte Vectorial tipo lineal presentó una mayor precisión, indicando que el 64,71% de los registros fueron predichos correctamente. Los resultados para los valores del promedio del área Debajo de la curva ROC corroboraron lo anterior, en el que el clasificador lineal presentó un poder discriminante muy aceptable (0,7254) frente al clasificador radial que presentó un poder discriminante nulo (0,5018). Finalmente, el coeficiente de Kappa fue mayor en el clasificador lineal (0,4239) frente al radial (-0,0011), lo que señaló que hay mayor concordancia en los resultados del clasificador lineal.

CONCLUSIONES

- La Máquina de Soporte Vectorial de tipo lineal es una técnica eficaz para predecir el riesgo académico de los estudiantes, ya que obtuvo una tasa de clasificación

correcta del 64,7%. Con mayores porcentajes de clasificación correcta para el riesgo académico normal y observado, pero menor para el riesgo académico suspendido.

- Los resultados de realizar la clasificación fueron muy precisos cuando se usó los datos de entrenamiento que los de prueba. Sin embargo, los resultados primeros fueron considerados preliminares, toda vez que la capacidad predictiva de la técnica se evalúa con los datos de prueba, datos que no fueron utilizados en la construcción del modelo. Resultó conveniente el uso de una Máquina de Soporte Vectorial tipo lineal frente al tipo radial para la predicción del riesgo académico de un estudiante.

RECOMENDACIONES

- A fin de poder incrementar la tasa de clasificación correcta, se recomienda realizar un análisis factorial confirmatorio en el preprocesamiento de datos, a fin de aminorar los datos faltantes y las inconsistencias producto de la integración de datos.

REFERENCIAS

Adams Harding, A., & Gingras, R. (2018). *Google News Initiative.* Obtenido de News Consumer Insights Playbook: https://newsinitiative.withgoogle.com/training/states/consumer_insights/pdfs/gni-new-consumer-insights-playbook.pdf

DBi Data Business Intelligence - Havas. (2019). Obtenido de Google Analytics: ¿Y tú qué necesitas? ¿la versión gratuita o 360?: https://dbibyhavas.io/es/blog/google-analytics-y-tu-que-necesitas-la-version-gratuita-o-360/

Google Analytics Developers. (2019). Obtenido de Enviar datos a Google Analytics: https:// developers.google.com/analytics/devguides/collection/analyticsjs/sending-hits?hl=es-419

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* California: Springer.

Jeffares, A. (Noviembre de 2019). *Towards Data Science*. Obtenido de K-means: A Complete Introduction: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

Kladnik, M., Stopar, L., Fortuna, B., & Mladenić, D. (2017). Audience Segmentation Based on Topic Profiles. *Jožef Stefan Institute and Jožef Stefan International Postgraduate School*, 1.

Lopez, G., Seaton, D. T., Ang, A., Tingley, D., & Chuang, I. (2017). Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data. *L@S '17: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*.

Syakur, M. A., Khotimah, B. K., Rochman, E. M., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 1.

Los artículos publicados por IECOS pueden ser compartidos a través de la licencia Creative Commons: CC BY 4.0 Perú. Permisos lejos de este alcance pueden ser consultados a través del correo revistas@uni.edu.pe.