

2022



MODELO PREDICTIVO DE LA MORTALIDAD ACADÉMICA DE UN PROGRAMA DE INGENIERÍA DE SISTEMAS BASADO EN TÉCNICAS DE MACHINE LEARNING

Juana Valentina Mendoza Santamaría, Universidad Santo Tomás, juana.mendoza@usantoto.edu.co

Martha Susana Contreras Ortiz, Universidad Santo Tomás, martha.contreras@usantoto.edu.co

Alexander Toro Cruz, Universidad Santo Tomás, alexander.cruz@usantoto.edu.co

Resumen.

Las instituciones de educación superior se enfrentan al desafiante reto de prevenir la deserción estudiantil y, asimismo, los estudiantes experimentan diferentes problemáticas socioeconómicas, académicas, institucionales y personales que afectan su rendimiento y permanencia. Para mitigar el riesgo de la deserción estudiantil, la gran mayoría de las universidades han implementado distintas estrategias, tales como: la creación de organismos para el apoyo estudiantil, el desarrollo de sistemas de alertas tempranas, la disposición del servicio de monitorias académicas, la caracterización de los estudiantes, entre otras. Por su parte, en los programas de Ingeniería se han identificado áreas específicas con altos niveles de mortalidad y deserción académica, en especial las asignaturas del área de ciencias básicas. Muchos estudios han diagnosticado las causas de esta manifestación, debido a factores socioeconómicos, la débil preparación del área en la educación básica y media, la apatía de algunos de los estudiantes por estas asignaturas, entre otras. Los estudiantes de Ingeniería de Sistemas no se escapan de esta situación, a pesar que ellos manifiestan motivación por su carrera, algunos han desertado de sus estudios. Además, fenómenos recientes como la pandemia del COVID-19, también han influido en la mortalidad académica y retraso en la graduación o posterior deserción, debido a diferentes factores como la afectación económica, la no disponibilidad de los recursos tecnológicos requeridos y problemas de salud propios o de algún familiar.

El objetivo de este proyecto consiste en desarrollar un modelo predictivo con técnicas de Machine Learning, que permita identificar los casos de mortalidad académica que pueden influir en la demora de la graduación o que conlleven posteriormente a la deserción de los estudiantes del programa de Ingeniería de Sistemas en la Universidad Santo Tomás Seccional Tunja. El presente trabajo inicia con un análisis bibliométrico para recopilar información de investigaciones que hayan utilizado el aprendizaje automático en el estudio de la mortalidad y deserción académica. Asimismo, se indaga sobre las técnicas de Machine Learning que mejor se ajustan para el tema y las experiencias obtenidas a nivel internacional. Dentro de la información relevante que influye en el análisis de la situación académica de un estudiante, convergen diferentes factores, tales como: académicos, socioeconómicos, médicos, familiares, psicológicos y los reportes de la unidad de atención al estudiante. Sin embargo, la mayoría de ellos corresponden a contextos sensibles y de difícil acceso. Por lo cual, en este estudio se consolidó el dataset con la información académica de los estudiantes en una institución de educación superior. Cabe aclarar que la recopilación del dataset incluye la protección y el anonimato de los datos. Después, se aplica la ingeniería de características para

2022



determinar cuáles son las más relevantes para un predictor de deserción. De esta forma, se determina un modelo de aprendizaje automático óptimo para la predicción de la deserción estudiantil, en el programa de Ingeniería de Sistemas en la universidad. Este proceso se realiza mediante la comparación de índices de desempeño, al utilizar diferentes algoritmos de Machine Learning con la información previamente recopilada. Por último, se establece la validación del modelo propuesto, realizando una confrontación con datos recientes de los estudiantes, para determinar el rango de validez de la predicción. En esta etapa se implementa la técnica de validación cruzada.

El principal aporte de este trabajo está orientado a identificar los factores académicos asociados a la mortalidad estudiantil, con el propósito de apoyar la predicción de casos que representen riesgo de deserción académica, en instituciones de educación superior. El resultado de este proyecto servirá de insumo para la siguiente fase, que consiste en la construcción de un sistema de reportes que permita la integración y la visualización de los resultados obtenidos al ejecutar el modelo predictivo.

Palabras Clave: Deserción Estudiantil, Educación Superior, Ingeniería de Sistemas, Machine Learning, Mortalidad Académica.

1. Introducción

Teniendo en cuenta los procesos de calidad, autoevaluación y autorregulación que siguen las instituciones de educación superior, se destaca un factor de gran importancia para la caracterización del proceso de los estudiantes y es la permanencia. Este factor se ve alterado por diferentes circunstancias que enfrentan los estudiantes, los cuales pueden variar entre instituciones, programas académicos, regiones, periodos académicos, entre otros.

A partir de lo anterior, surge el interés de proponer un proyecto que se centra en cómo predecir la deserción académica con base en los datos históricos de los estudiantes de Ingeniería de Sistemas de la USTA Tunja. Además, es importante establecer la tipología de los datos recolectados, para poder determinar cuáles son los factores influyentes en la deserción académica estudiantil.

El modelo predictivo propuesto es una herramienta que puede apoyar a los directivos de UDIES (Unidad de Desarrollo Integral Estudiantil), y el Departamento de Registro y Control con la toma de decisiones y mecanismos, para la prevención de la deserción académica en los estudiantes de la Facultad de Ingeniería de Sistemas de la USTA Tunja. Cabe resaltar que, el modelo propuesto genera predicciones a partir de solicitudes, las cuales constituyen un insumo significativo para la atención oportuna a los estudiantes.

La población objetivo de esta investigación son los estudiantes de la Facultad de Ingeniería de Sistemas, no obstante, este modelo puede ser replicado en otras facultades de la universidad. Comúnmente, las facultades llevan a cabo procesos de calidad para identificar oportunidades de mejora, y generar acciones que favorezcan la permanencia y minimicen la deserción estudiantil. Actualmente, esta labor se apoya de la plataforma SIIM (Sistema Integrado de Información Multicampus) y el sistema SAC (Sistema Académico), sin embargo, esto requiere un alto porcentaje de intervención humana. Mediante la aplicación de este modelo estos procesos pueden ser más ágiles y precisos.

2022



En la comunidad estudiantil cada miembro tiene un proyecto de vida que involucra al individuo y a su entorno más cercano, y no debe limitarse a un número que hace parte de una estadística. No obstante, las estadísticas contribuyen a identificar la tendencia de los casos de estudio, y permiten apoyar la toma de decisiones y acciones de mejora según se requiera.

La deserción estudiantil es un parámetro de estudio para dar seguimiento al comportamiento y posibles fluctuaciones de la vinculación y permanencia de los estudiantes en las instituciones educativas. En el caso de la Facultad de Ingeniería de Sistemas en los periodos del 2008-1 a 2015-1, el número de estudiantes matriculados era relativamente bajo, por lo cual, los casos de deserción eran más notorios. A partir del 2016, se evidenció un crecimiento progresivo de vinculación y permanencia de estudiantes, que se sigue manteniendo a la fecha. Sin embargo, aún se presentan casos de deserción estudiantil relacionados con el bajo rendimiento académico en las asignaturas del área de ciencias básicas. También, desafortunadamente algunos estudiantes atraviesan por diferentes problemáticas socio-económicas, personales, familiares y médicas que pueden afectar su desempeño académico.

Para la Universidad Santo Tomás Seccional Tunja es fundamental tomar acciones preventivas, para reducir las tasas de deserción académica en los estudiantes, y dar acompañamiento oportuno a la comunidad estudiantil. Por lo tanto, este proyecto plantea la siguiente pregunta de investigación: ¿Cómo predecir el riesgo de deserción académica de los estudiantes de la Facultad de Ingeniería de sistemas, mediante técnicas de Machine Learning?

2. Resultados

La metodología empleada para este proyecto fue CRISP-DM (Cross Industry Standard Process for Data Mining), la cual está compuesta por 6 fases iterativas (Schröer et al., 2021): a) Comprensión del negocio, b) Comprensión de los datos, c) Preparación de los datos, d) Modelado, e) Evaluación y f) Despliegue.

Este trabajo hace parte de un proyecto final de grado de una estudiante de Ingeniería de Sistemas. A la fecha, se han desarrollado las primeras 3 fases de la metodología CRISP-DM.

La primera fase de este proyecto consistió en la búsqueda sistemática de información sobre estudios relacionados con la implementación de técnicas de Machine Learning, orientados a predecir la deserción académica en instituciones de educación superior. Para ello, la cadena de búsqueda utilizada fue la siguiente: “Student dropout” + “Machine Learning”. En la Figura 1, se describe el proceso de selección, mediante la representación de un diagrama de flujo. Cabe resaltar que las cadenas de búsqueda utilizadas para excluir los estudios fueron: MOOC, LMS (Sistema de gestión de aprendizaje), e-learning, online, y escuela media.

2022

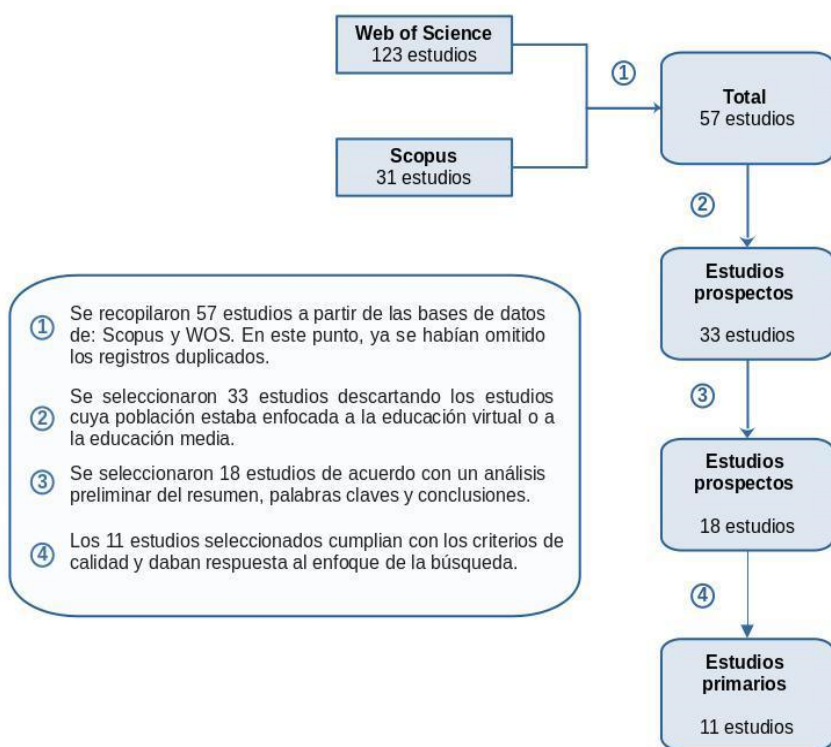


Figura 1. Diagrama de flujo búsqueda sistemática de literatura.

Además, se establecieron los siguientes criterios de inclusión:

- Los estudios deben de estar escritos en inglés.
- Los estudios deben de ser publicados a partir del año 2017 en adelante.
- Los estudios deben de estar enfocados a una población de educación superior con modalidad presencial.
- Los estudios corresponden a artículos de revista o artículos de conferencias.
- Los estudios deben de incluir técnicas de Machine Learning o de minería de datos.
- Los estudios deben aportar valor al presente proyecto y contribuir a dar respuesta a la pregunta de investigación.

Adicionalmente, para la recopilación y análisis de los artículos se emplearon las herramientas software VOSviewer y ScientoPy, las cuales permitieron realizar filtros para identificar los clusters más relevantes.

Una vez seleccionados los estudios primarios, se extrajo información relevante, tal como se presenta en la Tabla 1. Allí se pueden apreciar los factores de deserción estudiantil asociados y los algoritmos implementados de Machine Learning, con mejor porcentaje de Accuracy de cada uno de los estudios primarios.

2022



Tabla 1. Caracterización de estudios primarios.

ESTUDIO	REFER.	AÑO	FACTORES DE DESERCIÓN	ALGORITMOS DE MACHINE LEARNING
Predicting student dropout: A machine learning approach	(Kemper et al., 2020)	2020	Académicos Personales	Logistic Regression Decision Tree
Student dropout prediction	(del Bonifro et al., 2020)	2020	Académicos Personales	Support Vector Machine Linear Discriminant Analysis
Supervised learning in the context of educational data mining to avoid university students dropout	(K. J. O. de Santos et al., 2019)	2019	Académicos	Decision Tree Random Forest Support Vector Machine
A real-life machine learning experience for predicting university dropout at different stages using academic data	(Fernandez-Garcia et al., 2021)	2021	Académicos	Support Vector Machine Ensemble model Random Forest
Educational data mining: Analysis of drop out of engineering majors at the UnB – Brazil	(da Fonseca Silveira et al., 2019)	2019	Académicos Socioeconómicos	Generalized Linear Model Random Forest Gradient Boosting Machine
EvolveDTree: Analyzing Student Dropout in Universities	(G. A. S. Santos et al., 2020)	2020	Académicos Socioeconómicos	Decision Tree
Predictive modelling of student dropout using ensemble classifier method in higher education	(Hutagaol & Suharjito, 2019)	2019	Académicos Socioeconómicos Personales	Ensemble model Naïve Bayes K-Nearest Neighbor
Analysis of first-year university student dropout through machine learning models: A comparison between universities	(Opazo et al., 2021)	2021	Académicos Socioeconómicos Personales	Random Forest Gradient Boosting Decision Tree Decision Tree
Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study	(Maksimova et al., 2021)	2021	Académicos	Naïve Bayes Support Vector Machine Neural Network
Application of decision trees for detection of student dropout profiles	(Timaran Pereira & Caicedo Zambrano, 2017)	2017	Académicos Socioeconómicos Personales	Decision Tree
A machine learning approach to identifying students at risk of dropout: a case study	(Lottering et al., 2020)	2020	Académicos Socioeconómicos Personales	Support Vector Machine Logistic Regression Decision Tree

En el proceso de revisión de literatura se lograron identificar los factores más influyentes en la deserción estudiantil, donde se destacan los aspectos académicos, socioeconómicos y personales. Sin

2022



embargo, los dos últimos factores incluyen información sensible y de difícil acceso, por lo cual comúnmente representan una limitación para este tipo de proyectos.

Con base en la información recopilada, se puede concluir que los algoritmos de Machine Learning que obtuvieron mejores resultados fueron: Decision Tree (DT), Support Vector Machine (SVM) y Random Forest (RF).

Para la segunda y tercera fase se procedió a recopilar los datos correspondientes al rendimiento académico de los estudiantes de Ingeniería de Sistemas, durante los periodos 2018-1 al 2021-2. Esta información fue suministrada por las dependencias de Registro y Control y UDIES de la Universidad. Aunque, dentro de los factores más influyentes en la deserción estudiantil se destacan los aspectos académicos, personales y socioeconómicos (Ver Tabla 1), el presente estudio se enfocó en los factores académicos.

El dataset obtenido incluyó 5945 registros. Para la fase de preparación se procedió a normalizar los datos y se realizó un ajuste para garantizar la protección de la información. En la Tabla 2 se presentan los campos que conforman el dataset.

Tabla 2. Campos del dataset.

CAMPO DATASET	NOMBRE	TIPO DE DATO
stu_id	Código del estudiante	Numérico
stu_gender	Género del estudiante	Numérico (F = 0 o M = 1)
stu_level	Nivel del estudiante	Numérico (1 - 10)
period	Período académico	Numérico (1 - 8) 20181, 20182, 20191, 20192, 20201, 20202, 20211, 20212
sub name	Nombre de la asignatura	Numérico (1 - 5) Área Ciencias Básicas, Área Básicas de Ingeniería, Área Ingeniería Aplicada, Área Socio-Humanística, Área Institucional
sub_level	Nivel de la asignatura	Numérico (1 - 10)
teacher_id	Código del docente	Numérico consecutivo (1 - 127)
sub_repeated	Cantidad de veces que el estudiante ha cursado la asignatura	Numérico (1 - 5) Primera vez, segunda vez, tercera vez, cuarta vez, quinta vez
sub_absences	Fallas de asistencia en la asignatura	Numérico
fist_grade	Primera nota del corte del estudiante	Flotante
second_grade	Segunda nota del corte del estudiante	Flotante
third_grade	Tercera nota del corte del estudiante	Flotante
supp_aproved	Aprobación de la habilitación del estudiante	Numérico (0 = Sí, 1 = No, 2 = No presentó el examen)
final_grade	Nota final del estudiante	Flotante
stu_average	Promedio acumulado del estudiante	Flotante

2022



En el análisis de los datos se incluyeron los lineamientos estipulados por la universidad, donde se indica que un estudiante se encuentra en riesgo académico si cumple alguno de los siguientes requisitos: a) el promedio acumulado del estudiante es inferior a 3.5, o b) si está cursando por segunda o más veces una asignatura y la pierde nuevamente.

En la fase de modelado se definieron los algoritmos a implementar, basados en la revisión de la literatura y en las pruebas realizadas con los datos ajustados.

Con base en lo anterior, se implementaron los algoritmos de K-Nearest Neighbor, Gaussian Naive Bayes y Support Vector Machine en el dataset del rendimiento académico de los estudiantes y se alcanzó un accuracy de 88.3095%, 87.2161% y 86.1227% respectivamente.

Otros algoritmos implementados fueron Decision Tree, Random Forest y Logistic Regression. Sin embargo, todos ellos presentaron sobreajuste en los resultados.

Con la información del dataset, se procedió a establecer las alertas del modelo sobre una posible deserción estudiantil (parámetro de salida). A partir de lo anterior, se identificaron los factores que más impactaron este parámetro: repeticiones de asignatura e inasistencias. Por otra parte, los resultados de las notas obtenidas en cada corte académico, presentaron una alta correlación entre sí.

3. Conclusiones

Este trabajo suministra una herramienta que apoya a las instituciones de educación superior en la implementación de escenarios de ingeniería de características, aplicación de algoritmos y técnicas de aprendizaje automático. Además, aplica una reflexión del estatus quo de los programas de ingeniería de un país latinoamericano, con respecto a la mortalidad y deserción académica.

La comunidad científica encontrará en este trabajo herramientas para apoyar la predicción de mortalidad académica en programas distintos a la Ingeniería de Sistemas.

Según los resultados de este estudio, las instituciones educativas deben enfocar sus esfuerzos para el seguimiento de estudiantes en repitencias, en el control de asistencias a clase, debido a que son los factores predominantes en la predicción de la mortalidad académica de los estudiantes.

Como trabajo a futuro se propone complementar la información del dataset, con datos de otros programas, teniendo en cuenta características distintas a las académicas. Además, se plantea probar con otros algoritmos de Machine Learning, configurando los parámetros de modo que se ajusten a los datos históricos, en especial algoritmos de Deep Learning, mediante redes convolucionales.

Referencias

- Da Fonseca Silveira, R., Holanda, M., de Carvalho Victorino, M., & Ladeira, M. (2019). Educational data mining: Analysis of drop out of engineering majors at the UnB - Brazil. Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, 259–262. <https://doi.org/10.1109/ICMLA.2019.00048>
- De Santos, K. J. O., Menezes, A. G., de Carvalho, A. B., & Montesco, C. A. E. (2019). Supervised learning in the context of educational data mining to avoid university students dropout. Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019, 207–208. <https://doi.org/10.1109/ICALT.2019.00068>

2022



Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. *International Conference on Artificial Intelligence in Education*, 12163, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11/TABLES/3

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, 9, 133076–133090. <https://doi.org/10.1109/ACCESS.2021.3115851>

Hutagaol, N., & Suharjito, S. (2019). Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education. *Advances in Science, Technology and Engineering Systems Journal*, 4(4), 206–211. <https://doi.org/10.25046/aj040425>

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, 10(1), 28–47. <https://doi.org/https://doi.org/10.1080/21568235.2020.1718520>

Lottering, R., Hans, R., & Lall, M. (2020). A machine learning approach to identifying students at risk of dropout: a case study. *International Journal of Advanced Computer Science and Applications*, 11(10), 417–422. <https://doi.org/10.14569/IJACSA.2020.0111052>

Maksimova, N., Pentel, A., & Dunajeva, O. (2021). Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study. *Advances in Intelligent Systems and Computing*, 1329, 719–726. https://doi.org/10.1007/978-3-030-68201-9_70

Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, 9(20). <https://doi.org/10.3390/MATH9202599>

Santos, G. A. S., Belloze, K. T., Tarrataca, L., Haddad, D. B., Bordignon, A. L., & Brandao, D. N. (2020). EvolveDTree: Analyzing Student Dropout in Universities. *International Conference on Systems, Signals, and Image Processing*, 2020-July, 173–178. <https://doi.org/10.1109/IWSSIP48289.2020.9145203>

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526–534.

Timaran Pereira, R., & Caicedo Zambrano, J. (2017). Application of decision trees for detection of student dropout profiles. *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, 2017-December, 528–531. <https://doi.org/10.1109/ICMLA.2017.0-107>

Agradecimientos

A la Universidad Santo Tomás, agradecemos el apoyo recibido de sus dependencias y por ser el ente financiador de este proyecto.