Ing. Paula Zanetti[a], Ing. Pamela Timpani[b], Natalia Sfiligoy[c]

[a] Dir. carrera de Lic. en Sistemas UdeMM, Correo electrónico: paula.zanetti@udemm.edu.ar

[b] Profesora Lie. en Sistemas UdeMM. Correo electrónico: ptlmpanl@docentes.udemm.edu.ar

Profesora Lie. en Sistemas UdeMM. Correo electrónico: natalia.sfiligoy@docentes.udemm.edu.ar

Uso de Data Mining y aplicación de Inteligencia Artificial para la detección temprana de la deserción de estudiantes universitarios

En la última década, debidoa diferen¬ tes políticas públicas y el avance de la clase media, el número de estudian¬ tes de educación superior aumentó a veinte millones en Latinoamérica. Sin embargo, según un informe del Banco Mundial, se calcula que sólo el 50% de los estudiantes que inician sus estudios llegan a graduarse.

En cuanto a las implicancias que con¬ llevan a la deserción, los doctores Oscar Espinoza y Luis Eduardo González [1] distinguen tres categorías principales:

- Entre las implicancias sociales está la retroalimentación del círculo de la pobreza y la gestación de una "capa social" de profesionales frustrados con posible disminución del aporte intelectual y el potencial aumento del subempleo. A su vez, se incrementa el costo para el país de la educación superior debido a una suboptimiza¬ ción de los recursos producto de la deserción.
- Entre las implicancias institucionales están la limitación para cumplir la misión institucional y un descenso en los índices de eficiencia y calidad. De igual manera, tiene implicancias económicas debido a los menores ingresos por matrícula y a los costos adicionales para las universidades tanto públicas como privadas.
- Entre las implicancias personales que pueden asociarse a la deserción está el disgusto, la frustración y la sensa¬ ción de fracaso de los desertores con los consiguientes efectos en su salud física y mental. Asimismo, se produce una pérdida de oportunidades labora¬ les dadas las menores posibilidades de conseguir empleos satisfactorios, la postergación económica por sala¬ rios más bajos con los consiguientes impactos en los costos en términos individuales y familiares.

En algunos estudios realizados en Mercosur, se recomienda la implemen-tación de políticas multidimensionales orientadas a evaluar la calidad de los contenidos y planificar un programa de apoyo e incentivos para que el estudiante logre finalizar la carrera.

En consecuencia y con la intención de dar solución a este problema, resulta menester desarrollar una herramienta que permita predecir esta situación e identificar a aquellos alumnos que se encuentren en riesgo de abandonar sus estudios. Asimismo, si bien existen múltiples trabajos de investigación y estudios que plantean distintas metodo¬ logías para abordar el problema, muy pocos han llegado a implementarlas a nivel institucional.

Por todo lo expuesto, el presente tra¬ bajo sobre el uso de Data Mining e Inteligencia Artificial está orientado a establecer los cimientos de la cons¬ trucción de un algoritmo que pueda predecir el riesgo de abandono con los datos existentes en una universidad, aplicando las últimas tecnologías de procesamiento de información, como ser: aplicación de patrones de compor¬ tamiento, minería de datos, procesos estadísticos y analíticos, inteligencia artificial, etc.

Se utilizaron una serie de preguntas que servirán de guía para la investigación:

¿Cuáles son las causas y factores que intervienen en el abandono universita-rioenMercosur?

¿Qué trabajos existen en la actualidad que aborden esta problemática? ¿Qué tecnologías utilizaron, cómo fue su

1 (González, L.E. y Espinoza, O. (2008). Deserción en educación superior en América Latina y el Caribe. *Revista Paideia 45,* 33-46.)

implementación y con qué inconve¬ nientes se encontraron?

¿Qué tecnologías son las más adecua¬ das para construir el modelo? ¿Qué recursos serán necesarios?

¿Es factible la implementación de un modelo predictivo de abandono univer- sitarioenel ámbitouniversitario?

¿Qué tareas serán necesarias para actualizar, mejorar y mantener el modelo propuesto?

¿Cómo sería la interfaz de usuario y qué datos mostraría?

¿Qué elementos deberían modifi¬ carse para hacerlo extensivo a otras universidades?

En el análisis del Marco teórico y el Estado del Arte se analizaron experien¬ cias sobre el abandono universitario en Mercosur en el periodo 2013-2018 para generar un modelo predictivo que pronostique la deserción en el ámbito universitario de la región y establecer un plan de acción que describa las características más importantes con las que debería contar.

Con los siguientes objetivos:

- Evaluar la situación económica, social y cultural sobre estudiantes uni versitarios en Mercosur en el periodo 2013-2018, analizando experiencias sobre la deserción universitaria.
- Investigar la existencia de estudios de universidades que hayan abordado el tema de la deserción de alumnos y analizar qué tecnologías se han utilizado.
- Examinar la bibliografía relacionada con nuevas tecnologías y con la pro¬ blemática en sí, identificando varia¬ bles e indicadores que intervinieran en el abandono de sus estudios.
- Seleccionar y analizar criterios defi¬ nidos que permitan implementar un modelo de solución.
- Plantear un modelo de datos acorde a la problemática propuesta.
- Definir la forma de retroalimenta- ción del modelo a construir.
- Evaluar los eventuales riesgos, esti¬ mar la confiabilidad y precisión de los datos obtenidos.
- Generar un prototipo orientado al usuario para el acceso a las funciona¬ lidades de la herramienta.
- Componer recomendaciones para cada curso de acción según los distin¬ tos resultados del modelo a construir.

1. Sobre la investigación

En una primera revisión de la literatura existente se encontró un gran número de artículos, libros, informes y desarrollos de universidades que trataron el tema. En algunos casos, se han construido modelos para analizar la problemática, definiendo y relacionando las variables intervinientes. También se han plan¬ teado varias hipótesis sobre las causas de la deserción universitaria, se descri¬ bieron tecnologías a utilizar y se crea¬ ron indicadores. Teniendo en cuenta esta amplia bibliografía, la investiga¬ ción comenzará siendo descripúva en lo que concierne a la investigación del estado del arte y explicativa en lo que respecta al análisis de factibilidad.

La investigación comenzará siendo descriptiva ya que, además de la gran literatura existente, un estudio des¬ criptivo implica medir conceptos o variables en forma independiente, según distintos enfoques, en diferentes dimensiones. Según Dankhe (1986), este tipo de investigación *requiere con¬ siderable conocimiento del área que se investiga para formular las preguntas específicas que busca responder.*

Para el estudio de factibilidad, se plantea una invesügación explicativa ya que está dirigida a estudiar la via¬ bilidad de implementar una solución y qué condiciones son requeridas para tal fin. El estudio explicativo analiza las variables que intervienen y de qué modo. Es una investigación más estructurada y, debido a que requiere un sentido de entendimiento del fenó¬ meno, se incluye en la segunda etapa luego de haber avanzadoenel estudio descriptivo.

1.1. *Informe Socioeconómico relacio¬ nado con la Educación Superior en los países del Mercosur*

Para abordar la problemática de la deserción universitaria en el Mercosur, resulta importante realizar un estudio previo de la situación actual en lo que respecta a la educación universitaria de cada uno de los países que lo integran.

El Mercosur, Mercado Común del Sur, es un proceso de integración regional fundado en 1991 a través del Tratado de Asunción por Argentina, Brasil, Paraguay y Uruguay. En función de que el tratado está abierto a la adhesión de otros Estados miembros de Asocia¬ ción Latinoamericana de Integración, Venezuela se adhirió al tratado en 2006, y más recientemente, Bolivia encontrándose en proceso de adhesión.

Otros países como Chile, Colombia, Ecuador, Perú, Guyana y Surinam inte¬ gran los “Estados Asociados”. Éstos son aquellos miembros del ALADI con los cuales el MERCOSUR suscribe acuerdos de libre comercio y están autorizados a participar en las reunio¬ nes de órganos del Mercosur que traten temas de interés común.

Es considerado una potencia econó¬ mica, con un PBI del 82,3 % del PBI total de toda Sudamérica y representa alrededor del 70% de habitantes de América del Sur. Además, se consti¬ tuye como el área económica y plata¬ forma industrial más dinámica, compe¬ titiva y desarrollada del hemisferio sur, formando el cuarto bloque económico del mundo en importancia y volumen de negocios.

En cuanto a la economía, esta es muy diversa, ya que posee las tres ciudades más ricas y pobladas de Sudamé- rica: Sáo Paulo, Buenos Aires y Rio de Janeiro, donde se permite el libre comercio y circulación de personas.

El Mercosur fue considerado en 2011 el mayor productor de alimentos en el

mundoy controla las mayores reservas energéticas, minerales, naturales, de recursos hídricos y petróleo del pla¬neta. Es una economía fuertemente industrializada con gran diversidad de productos.

1.2. *Herramientas y Metodologías*

La minería de datos se refiere al pro¬ceso que intenta descubrir patrones a partir de la extracción del conoci¬miento de los grandes volúmenes de datos. Su objetivo general consiste en obtener información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. A través de la bibliografía consultada y a partir de trabajos anteriores, se encon¬traron varias herramientas dentro de la minería de datos que pueden ser útiles para desarrollar este modelo. A modo de ejemplo, se describen a continua¬ción las más relevantes:

Arboles de decisión: es un diagrama que contiene un nodo raíz donde se encuentran todas las observaciones, nodos intemos que albergan a los nodos de división y nodos hoja que contiene la clasificación final para un conjunto de observaciones. Un árbol repre¬senta una segmentación de los datos, que se crea mediante la aplicación de reglas simples, una después de otra, originando una jerarquía de segmentos dentro de segmentos. Así, los nodos intemos representan validaciones sobre los atributos, las ramas representan las salidas de las validaciones y los "nodos hoja" representan las clases.

Para aplicarlos, primero se identifican los factores que influyen en la deser¬ción. Según Cuji (2017), estos factores se centran en variables como: ingreso económico, educación de padres, ren¬dimiento académico, etc. Se construye un árbol de decisión de varios niveles de profundidad a partir de un algoritmo especializado y se detectan aquellos factores que tienen mayor incidencia en la deserción.

Agrupamiento o Clustering: es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar casos de un conjuntodedatos dentro de clústeres que contienen características similares. Estos agrupamientos son útiles para la exploración de datos, detección de anomalías y predicciones. Los datos se agrupan según la similitud de los valores, en forma no supervisada, es decir, que no se conoce de antemano las clases del conjunto de datos. Existengran variedad de algoritmos de Clustering, entre los que se encuentran K-medianas (K-Means) o EM (Expectation/Maximization).

Independientemente de la herramienta tecnológica a utilizar, la metodología de trabajo se basa en el Proceso KDD Knowledge Discovery in Databases (Fay-yad,1996), proceso de descubrimiento del conocimiento centrado en el usuario.

El KDD se encarga de la preparación de los datos y la interpretación de los resul¬tados obtenidos; es un proceso de extracción de información potencialmente útil a partir de un gran volumen de datos en el cuál la información está implícita y apunta a encontrar relaciones o patrones. Se divide en cinco etapas, esquematiza¬das en la siguiente Figura:

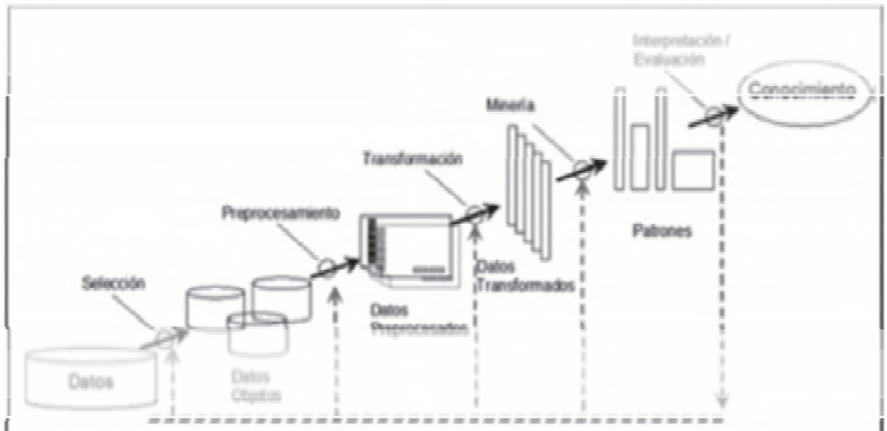

Figura 3. Etapas del KDD y sus respectivas entradas/salidas

1. **Selección de datos:** se define el objetivo y qué datos serán recolectados, a tos de entrada y salida y la justificación sobre qué se desea obtener.
2. **Procesamiento:** se diseña el esquema de un almacén de datos para unificar de manera eficiente la información. Consiste en la preparación y limpieza de los datos extraídos. Se aplican estrategias para manejar datos faltantes o en blanco, datos inconsistentes o fuera de rango.
3. **Transformación:** consiste en el tratamiento preliminar de los datos, transforma¬ción y/o generación de nuevas variables a partir de las ya existentes. Se realizan operaciones de agregación o normalización.
4. **Minería de datos:** se aplica al técnica de minería de datos más apropiada con el objetivo de extraer patrones desconocidos, válidos, nuevos y comprensibles, que están contenidos u "ocultos" en los datos.
5. **Evaluación:** se interpretan los diferentes aspectos de los datos procesados (en cuanto a utilidad, coherencia, apego a la realidad, etc.) Con los datos procesa¬dos junto con las evaluaciones, se extrapolan los casos ya contemplados y se identifican los patrones por medio de diferentes técnicas (análisis estadísticos, lenguajes de consultas, entre otros).

Otro de los puntos para tener en cuenta es la validación de los modelos aplicados. No existe un algoritmo mejor que otro, pero sí se puede determinar cuál/es resultan

de mejor aplicación a la problemática abordada. En minería de datos se utilizan diversos criterios para evaluar el desempeño de los modelos, y posteriormente seleccionar alguno de ellos.

- **Exactitud:** Estas medidas indican cuán cerca o lejos del valor medido se encuentra la predicción realizada a partir de los datos disponibles. Sin importar la métrica elegida, su resultado se verá influido por la calidad de los datos, ya que pueden existir valores perdidos o imputados que aumenten la incertidumbre en modelo y, en consecuencia, afecten el cálculo del error.
- **Confiabilidad:** evalúa la manera en que se comporta un modelo en conjuntos de datos diferentes. Un modelo es confiable si genera el mismo tipo de predicciones o encuentra los mismos tipos de patrones independientemente de los datos de prueba proporcionados.
- **Utilidad:** hace referencia a que, independientemente de la exactitud y la confiabilidad, el modelo permita generalizar correctamente y logre responder preguntas relevantes para el caso de estudio.

Estos factores, junto con otros como la disponibilidad, los costos y el tiempo de procesamiento, serán tenidos en cuenta en el estudio de factibilidad del trabajo de investigación.

Con respecto a las herramientas a utilizar, los datos se encuentran alojados en una base de datos Oracle, y se necesitará acceso de lectura a la misma para poder acceder a la información que, cumpliendo con las etapas del proceso KDD, será transformada y acondicionada para realizar un análisis exploratorio y luego aplicar los métodos predictivos pertinentes.

Por otro lado, para realizar el análisis exploratorio y el preprocesamiento de datos, han sido evaluados los lenguajes Python y R ya que, además de ser dos de los más utilizados en el ámbito de la ciencia de datos, ambos son Open Source, con comunidades de desarrolladores detrás, que los mejoran y enriquecen con múltiples librerías.

En la actualidad, ambos lenguajes poseen implementaciones tanto de las técnicas de minería de datos más tradicionales como de las más modernas. Sin embargo, es recomendado utilizar Python cuando se desea integrar el código a una WebApp o a una base de datos productiva, ya que se trata de un lenguaje de programación completo, no solo con fines estadísticos.

Para el proyecto de investigación se utilizará Python 3, ya que, de lograrse un modelo útil para el caso de estudio, se buscará desarrollar un prototipo de aplicación con interfaz de usuario a través de la cual se pueda interactuar con los datos y proveer visualizaciones con información accionable.

De las diversas distribuciones de Python se utilizará Anaconda, dado que provee todas las herramientas necesarias para la tarea y es la distribución con la que, quienes desarrollan la investigación, se encuentran familiarizados.

Versión	Descripción
ActiveState ActivePython	Distribución con una versión comercial y otra para la comunidad que incluye módulos de computación científica
Pythonxy	Distribución con orientación científica basada en Qt y Spyder)
Winpython	Distribución científica portable para Windows
Conceptive Python SDK	Distribución dirigida a aplicaciones de negocios, escritorio y bases de datos
Enthought Canopy	Distribución comercial para computación científica
PylMSL Studio	Distribución comercial para análisis numérico - gratis para uso no comercial
Anaconda Python	Distribución completa para la gestión, análisis y visualización de grandes conjuntos de datos
eGenix PyRun	Distribución portable complete con stdlib

Distribuciones de Python

Así mismo, para una mayor claridad del desarrollo, se utilizarálaherramienta Jupyter Notebook a modo de documentación, con el objetivo de realizar un paso a paso pormenorizado de las transformaciones de datos necesarias y así lograr un mayor entendimiento de la información, ya sea a través de gráficos o tablas.

Inicialmente, se requerirá el uso de las librerías cx_Oracle, Numpy y Pandas para acceder y transformar los datos, y Seabom, Skikit-leam y Matplotlib para analizar la información a través de gráficos y realizar el análisis exploratorio.

Una vez que se tenga un mayor conocimiento de los datos disponibles, se postularán posibles *métodos predictivos* y se seleccionarán librerías que provean su implementación. Para validar la exactitud, confiabilidad y utilidad de los modelos obtenidos, según lo anteriormente establecido, se llevarán a cabo las siguientes actividades:

- Definir las métricas a optimizar por los modelos.
- Particionar el set de datos en entrenamiento y prueba.
- Realizar una validación cruzada del conjunto de datos de entrenamiento.
- Construir visualizaciones para interpretar los resultados del modelo.

En ese contexto, y para el desarrollo de la investigación, se han propuesto los siguientes objetivos y actividades:

- Investigación del estado del arte.
- Análisis estructural de los datos existentes.
- Análisis exploratorio de la información.

- Análisis de factibilidad paraelestablecimiento de un algoritmo repetible que permita generar un modelo predictivo de estudiantes en riesgo de abandono de sus estudios.
- Selección de algoritmos para el análisis predictivo.

La correcta documentación y visualización de los resultados será crucial para su análisis, del cual se espera el descubrimiento de algún patrón o regla que permita estudiar el fenómeno de deserción estudiantil y su detección temprana, con el fin de proveer asistencia a la posible toma de alguna acción correctiva.

2. Bibliografía

Ahumada H, Dip H, Herrera C, Leguizamón Almendra J (2015). *Minería de datos para un Sistema de alerta temprana de deserción en Carreras de Ingeniería.* Departamento de Formación Básica / Facultad de Tecnología y Ciencias Aplicadas / Universidad Nacional de Catamarca.

CEA (Centro de estudios de la educación argentina) Abril 2015. *Revista del Centro de estudios de la educación argentina.* Año 4 Número 34. Lugar de Publicación: Universidad de Belgrano.

Charu C. Aggarwal (2015). *Data Mining The Textbook.* Springer

González Fiegehen L E (2007). *"Repitencia y Deserción Universitaria en América Latina* " en Informe sobre l; Educación Superior en América Latina y el Caribe 2000-2005: La metamorfosis de la educación superior. Capítulo 11. Caracas: IESALC & UNESCO.

Cují, B., Gavilanes, W. y Sánchez, R., *"Modelo Predictivo de Deserción Estudiantil Basado en Árboles de Decisión",* revista Espacios, julio-agosto 2017, (vol. 38, N° 55) p. 17.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P, *"From Data Mining to Knowledge Discovery in Databases",* Al magazine, 1996, (vol. 17, N° 3), p. 37.