Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

ANÁLISIS DEL RENDIMIENTO ACADÉMICO UNIVERSITARIO A TRAVÉS DE TÉCNICAS DE MINERÍA DE DATOS.

APLICACIÓN EN ALUMNOS DE LA UNIVERSIDAD NACIONAL DE SAN MARTIN.

AUTORA: CONSTANZA GRECO

DICIEMBRE 2019

Resumen

La búsqueda de la calidad educativa en el sector universitario es un bien deseado por distintos sectores en una sociedad y por el que luchan las universidades y los países desde diferentes ámbitos. Esto hace que la búsqueda de esa calidad implique una revisión integral de la universidad y que esté asociada al estudio del rendimiento académico del alumnado, pues permite favorecer el control de los recursos estatales y la mediación del impacto social. Asimismo, su análisis es de gran utilidad en procesos de toma de decisiones en aras de mejorar el sistema en general.

En el presente trabajo se analizan los factores que influyen de manera significativa sobre el rendimiento académico de los estudiantes de la Escuela de Ciencia y Tecnología de la Universidad Nacional de General San Martin. Particularmente se aplica un modelo de predicción para evaluar la tasa de abandono de los alumnos y luego poder abordar a aquellos que presentan mayor probabilidad de suspender sus estudios. Se propone un modelo de árboles predictores que emplea las variables más relevantes halladas a partir de la investigación, las cuales se considera que son las que inciden directamente en el rendimiento de los estudiantes. Se demuestra que, a partir de esta información, es posible predecir con una aceptable confiabilidad el rendimiento de un alumno dado.

PALABRAS CLAVES: Análisis de datos, rendimiento académico, modelo predictivo, universidad.

Introducción

La incorporación de alumnos que, en muchos casos, constituyen la primera generación de su familia en la educación superior es un importante factor aspiracional de movilidad social ascendente. Esa condición que conlleva un déficit en su capital social y cultural y las debilidades del proceso de educación media en su preparación para el ingreso a la vida universitaria, exigen a las Universidades desafíos mayores que no se restringen a la gratuidad o el libre ingreso para que se logre el objetivo. Investigaciones sobre el desempeño estudiantil permiten conocer un gran número de variables que entran en juego en lo que a calidad y equidad de la educación superior pública se refiere, por lo que aportan importantes elementos que repercuten en la gestión y prestigio institucional, sobre todo cuando la inversión estatal es fundamental. (Garbanzo Vargas, 2007).

En particular se elige abordar el caso particular de la Universidad de San Martín por su especial análisis que se puede realizar acerca de la asistencia y finalización de los estudios. ¨…En este quehacer nos sentimos interpelados por el llamado a dar cumplimiento efectivo al ideal de la educación superior como derecho fundamental para la transformación social. Desafío que nos incita a construir y desplegar una imaginación social situada y comprometida con el principio de justicia social, pilar de nuestra identidad institucional que reviste la misma importancia que el de innovación tecnológica, investigación y desarrollo¨. (Greco, 2019).

Por lo dicho anteriormente, resulta de gran importancia la aplicación de un modelo para analizar y predecir el rendimiento académico del alumno, es decir en qué momento alcanzará su graduación. A través de dicho análisis se busca apoyar la toma de decisiones de una universidad y contribuir a entender qué variables afectan el rendimiento académico de un alumno.

Identificar el rendimiento académico en la educación superior resulta complejo dado que es problemático y confuso relacionarlo con las notas. La valoración del rendimiento académico no conduce a otra cosa que a la relación entre lo que se aprende y lo que se logra desde el punto de vista del aprendizaje, y se valora con una nota, cuyo resultado se desprende de la sumatoria de la nota de aprovechamiento del estudiante en las diferentes actividades académicas, a las que se sometió en un ciclo académico determinado (Garbanzo Vargas,

2007). El acceso a la educación universitaria de un público estudiantil cada vez más heterogéneo en términos de su perfil socioeconómico, educativo y en aspiraciones académicas y laborales requiere que las universidades exploren nuevos caminos pedagógicos e institucionales para lograr que estos jóvenes se gradúen, adquiriendo además los conocimientos y habilidades necesarias para desenvolverse con éxito en su campo académico y profesional. Formar profesionales y científicos sin rebajar los niveles de calidad, y sobre todo buscando elevarlos, es actualmente un desafío de alta complejidad en el contexto de organizaciones de gran tamaño y modesto presupuesto. La masificación de la educación superior y las restricciones presupuestarias contribuyen entonces con el diseño de políticas tendientes a promover la elevación de la calidad y la eficiencia organizacional. (Fanelli, 2014).

En línea con lo dicho anteriormente, surge el estudio de dos indicadores esenciales a la hora de evaluar la gestión en una universidad: la tasa de graduación (disminuyendo la tasa de abandono) y la reducción del tiempo demandado para formar un graduado. En las últimas dos décadas, la mejora en los índices de rendimiento académico y graduación se ha incorporado como tema de alta relevancia en la agenda de políticas públicas e institucionales en América Latina (CINDA, 2006).

Considerando lo expuesto anteriormente, surge entonces la siguiente pregunta de investigación:

¿Cuál es el método de minería de datos que predice con mayor exactitud el abandono en alumnos universitarios?

En función a la problemática planteada entonces se define como objetivo general del presente trabajo encontrar el mejor modelo para predecir el abandono de los alumnos de la Escuela de Ciencia y Tecnología de la Universidad de San Martín mediante técnicas de minería de datos. Asimismo, como objetivos específicos se considera como primer paso definir el concepto de rendimiento académico y caracterizar y presentar a la Universidad Nacional de San Martin. Luego, analizar y preparar los datos a utilizar para el armado del modelo. En el tercer objetivo específico se define investigar y proponer modelos predictivos predeterminados, elegir una métrica para clasificarlos y comparar dichos modelos a partir de la métrica. Finalmente se plantea analizar cuáles son las variables que más influyen en el

alumno a la hora de decidir abandonar o no sus estudios y cuáles son las que determinan que el mismo alcance su graduación.

En el primer apartado se analiza el del concepto de rendimiento académico según lo definido por distintos autores. Además, se analizan los factores básicos que influyen y determinan el rendimiento de los alumnos. Se toman de referencia resultados y conceptos de investigaciones que se consideran las más relevantes en función de los objetivos de la presente investigación. Asimismo, se presenta el caso particular de la Universidad Nacional de San Martin y se analiza cuál es su situación respecto al estudio del rendimiento académico de sus alumnos.

En el segundo apartado se presenta la base de datos con la que se trabaja. En este sentido, se presentan las variables tomadas en cuenta, el tratamiento de los datos y el arribo al set de datos final.

Con respecto al tercer apartado, se describen los modelos de minería de datos puestos a prueba y los parámetros que se toman en cuenta para los mismos. Además, se define y justifica la métrica a utilizar para comparar los métodos.

Finalmente, en el último apartado se analizan los resultados. Se presentan las conclusiones derivadas de los resultados, se discuten los mismos en relación a los datos aportados por la investigación científica que ha servido de fundamentación teórica y se muestran las limitaciones, aportes, el análisis crítico y las sugerencias para trabajos futuros a partir de lo analizado en la presente investigación.

El análisis del rendimiento académico universitario. La Universidad Nacional de San Martin.

En este primer apartado, se presenta el concepto de Big Data, el rendimiento académico y la aplicación de dichos conceptos en el estudio de la Universidad Nacional de San Martin. En primer lugar, se trata de introducir el concepto de Big Data y la relación que tiene el mismo con el análisis del Rendimiento Académico. Por otro lado, se delimita el constructo ¨Rendimiento Académico¨ diferenciándolo de otros muy similares revisando la bibliografía existente en la que se investiga dicho concepto. Finalmente se introduce el estudio del caso de la Universidad Nacional de San Martin contando sus características y particularidades.

1.1. Big Data y su relación con el estudio del Rendimiento Académico:

Big data se presenta como un proceso de negocios que procesa grandes volúmenes de datos en forma automatizada, facilitando la toma de decisiones en tiempo real para aumentar el valor económico de la organización (Schmarzo, 2013). En el caso del Estado, las organizaciones de gobierno utilizan Big data para obtener un conocimiento más específico de la población y del territorio y, a partir de esta nueva información, ejercer la soberanía a través de acciones que aumenten el valor aportado a la sociedad (Bryson, Crosby, y Bloomberg, 2015; Mergel, Rethemeyer, y Isett, 2016). En base a esto es que resulta primordial la aplicación del análisis de grandes volúmenes de datos en las organizaciones dado que la toma de decisiones basada en ellos mejora el desarrollo de la organización.
Particularmente en el caso de Argentina, las estadísticas y datos suelen ser difíciles de obtener y se torna complejo crearlas con datos oficiales. Los modelos de predicción que incluyen todos los datos personales, sociales, económicos y académicos en este caso, son necesarios para lograr una buena predicción de la continuidad o no del estudiante en la universidad. Asimismo, el análisis del fenómeno es complejo, dado que son muchos los factores intervinientes, vinculados con las inversiones en educación, las nuevas demandas del capital, los nuevos y heterogéneos perfiles de alumnado que accede a la universidad. Su relación con el Big Data puede resultar verdaderamente revolucionarias cuando se las aplica a educación. Lograr modelos con una alta precisión es beneficioso para identificar inicialmente a los estudiantes con riesgo de deserción para luego poder abordarlos y asistirlos

en sus problemáticas. Las construcciones de dichos modelos a partir del análisis de los datos son valiosas no solo para definir el rendimiento promedio del alumnado sino también para detectar los potenciales alumnos que abandonarán sus estudios, e incluso para medir cual es la variable que más impacta en su decisión de abandonar o continuar. Además, los resultados pueden ayudar a tomar medidas por parte de las instituciones para mejorar la situación definida.

A lo largo de este documento ese intenta indagar la multidimensionalidad del rendimiento académico de los estudiantes y las diversas modalidades que asumen la heterogeneidad, la segmentación y polarización académica. No obstante, cabe aclarar que estos rasgos se refieren a procesos diferentes que se yuxtaponen e influencian recíprocamente. Mientras que la heterogeneidad es un rasgo que constituye la diversidad dentro del rendimiento académico, los procesos de segmentación y polarización requieren de una mayor atención. A partir de la caracterización del desempeño de los estudiantes y la dinámica de los procesos de formación académica analizando y planteando un modelo, se espera que sea de gran aporte para pensar tanto los fenómenos de "abandono", suspensión o posposición de los estudios, así como para problematizar las dificultades inherentes al proceso de graduación en la Universidad.

Finalmente, en relación al estudio de los datos dentro del análisis del rendimiento académico, se concluye que existe una gran cantidad de desafíos vinculados con la recolección de datos y es necesario que cada institución que aborda un programa de este tipo piense como éste afectará a la privacidad de sus estudiantes, y qué mecanismos de transparencia utilizará alrededor de la información. Del mismo modo, existe el reto de no desmotivar a los estudiantes a partir de los pronósticos negativos derivados del análisis de sus hábitos y desempeño. Es por eso que las universidades están buscando el modo más indicado de encarar estos programas, y de implementar respuestas adecuadas que permitan incrementar las posibilidades de éxito de todos los estudiantes. No cabe duda que el rendimiento promedio del alumno y la prolongación de sus estudios son problemas preocupantes, por las repercusiones sociales, institucionales y personales que conlleva. El aporte que puede hacer en este sentido el análisis de datos es de gran valor tanto para la universidad, como para el alumno y finalmente para el sistema educativo en general. A través de modelos de minería de datos, que se definen como el análisis de grandes volúmenes de datos para el

descubrimiento de patrones y para resumir los datos de formas novedosas que sean comprensibles y útiles para el usuario de los datos en diferentes organizaciones (Hand, Mannilla y Smith, 2001), se busca predecir el período de tiempo en que un estudiante finalizará sus estudios.

1.2. El rendimiento académico:

En el mundo globalizado se le atribuye un lugar especial al conocimiento, aduciendo que se tenderá a valorar de manera creciente el avance teórico y la innovación tecnológica, por lo que la inversión en la formación y en la investigación se vuelve indispensable para la producción y reproducción del sistema social y económico (Beck, 1999).

En las últimas décadas, la educación superior adquirió un lugar destacado –que antes no tenía– en la agenda de las políticas públicas de América Latina. En los países en desarrollo, el Estado desempeñó un papel preponderante en la orientación de políticas y programas que favorecieron la democratización de los sistemas y la inclusión social. [...] Las universidades nacionales llevan adelante un papel social importante en promover el desarrollo económico y científico del país pues el conocimiento que ahí se transmite, genera y aplica, sin duda coloca a los sujetos ante la posibilidad de transformar su ámbito de acción y ser al mismo tiempo transformados por el saber. (Comisión Asesora del Consejo Superior de la UNSAM, 2018).

Identificar el rendimiento académico en la educación superior resulta complejo dado que es problemático y confuso relacionarlo con las notas. La valoración del rendimiento académico no conduce a otra cosa que a la relación entre lo que se aprende y lo que se logra desde el punto de vista del aprendizaje, y se valora con una nota, cuyo resultado se desprende de la sumatoria de la nota de aprovechamiento del estudiante en las diferentes actividades académicas, a las que se sometió en un ciclo académico determinado (Garbanzo Vargas, 2007). Cuando se habla de rendimiento académico se refiere al nivel de conocimientos que el alumno demuestra tener en el campo, área o ámbito que es objeto de evaluación; es decir el rendimiento académico es lo que el alumno demuestra saber en las áreas, materias, asignaturas, en relación a los objetivos de aprendizaje y en comparación con sus compañeros de aula o grupo. Así pues, el rendimiento se define operativamente tomando como criterio las calificaciones que los alumnos obtienen (Luengo, 2015). El rendimiento académico es la

resultante de un conjunto de factores personales, sociales, educativos-institucionales y económicos. La valoración de las consecuencias y repercusiones del éxito o fracaso escolar, la realidad de cómo trasciende al propio ámbito académico, la conexión directa de la función productiva de la sociedad, la adecuación de los diversos tratamientos educativos para la consecución de los objetivos propuestos junto a las inversiones realizadas en educación en base a la satisfacción de las demandas sociales; influyen en la valoración del rendimiento académico. El rendimiento está muy influenciado por el esfuerzo individual del sujeto que aprende y por la voluntad o perseverancia en el esfuerzo (Kaczynska, 1965). Se "entiende que una escuela es eficaz si consigue un desarrollo integral de todos y cada uno de sus alumnos mejor de lo que sería esperable teniendo en cuenta su rendimiento previo y la situación social, económica y cultural de las familias" (Murillo, 2003). En este sentido, si bien las calificaciones son el indicador del rendimiento, el valor añadido consiste en que ha potenciado la investigación sobre factores que facilitan el rendimiento, así como la evaluación de programas de mejora y estudios etnográficos sobre la escuela.

El acceso a la educación universitaria de un público estudiantil cada vez más heterogéneo en términos de su perfil socioeconómico, educativo y en aspiraciones académicas y laborales demanda que las universidades exploren nuevos caminos pedagógicos e institucionales para lograr que los jóvenes se gradúen, adquiriendo además los conocimientos y habilidades necesarias para desenvolverse con éxito en su campo académico y profesional. Formar profesionales y científicos sin rebajar los niveles de calidad, y sobre todo buscando elevarlos, es actualmente un desafío de alta complejidad en el contexto de organizaciones de gran tamaño y modesto presupuesto. […] En particular, el aumento en la tasa de graduación (disminuyendo la tasa de abandono) y la reducción del tiempo demandado para formar un graduado, devinieron dos indicadores que actualmente se emplean para dar cuenta del cumplimiento de tales objetivos. (Fanelli, 2014).

Para llevar a cabo una interpretación adecuada del fenómeno de graduación resulta conveniente articular tres tipos de indicadores. Por lo tanto, para elucidar el problema de la graduación universitaria y las dificultades inherentes al ritmo de estudio y eventual "abandono" asociados al mismo, se requiere explicitar e integrar en el análisis los alcances de los siguientes indicadores: duración media, relación entre duración media - duración teórica y tasa de graduación. Solo a partir de dicha integración se podrá establecer un

diagnóstico y análisis preciso sobre el fenómeno de graduación en la UNSAM. La duración media (DM) de una carrera se suele definir como el promedio empleado en graduarse, medido en años, por un conjunto de estudiantes de una carrera. Ese conjunto de estudiantes puede tener en común el año de ingreso, el año de egreso o estar conformado por un agregado que incluya estudiantes de diversas cohortes o promociones (de ingreso o de egreso). La conformación de dicho conjunto, la duración teórica de la carrera (esto es, la duración en años establecida en el plan de estudios) y la antigüedad de creación de la misma son factores que pueden generar importantes variaciones en la medición de la duración media. Así, por ejemplo, transcurridos 6 años desde la creación de una carrera de 5 años de duración teórica, la duración media calculada sobre la base de una cohorte de egresados será muy similar a la duración teórica, pudiendo ocurrir que este indicador asuma valores más altos a medida que pase el tiempo, y comiencen a registrarse graduados que emplean en graduarse un tiempo considerablemente mayor que el tiempo teórico. Puesta en relación con la duración teórica (DT), la duración media permite obtener un indicador derivado que torna comparables carreras de diferente duración según el plan de estudios: el cociente entre la duración media y la duración teórica (R = DM / DT). Si este indicador de relación toma valores cercanos a 1, es posible afirmar que los estudiantes de la carrera considerada se están graduando aproximadamente en el tiempo previsto en el plan de estudios; los valores mayores que 1 indican que los estudiantes emplean en graduarse un tiempo mayor al previsto. La tasa de graduación puede definirse como el porcentaje de estudiantes egresados respecto del total de estudiantes ingresantes de una determinada cohorte. Nuevamente, la duración teórica de las carreras que se consideren para calcular dicha tasa, así como la antigüedad de su creación, son factores que afectan la medida de la tasa de graduación.

Las causas asociadas al alargamiento de los estudios pueden ser variadas e incluir tanto factores endógenos (factores asociados a la organización de la institución, de la carrera, presencia o no de tutorías, regulaciones claras sobre la condición de alumno regular o el cumplimiento de requisitos mínimos de rendimiento académico, el Plan de estudios, el régimen de correlatividades o las condiciones pedagógicas del cuerpo docente, entre otras) como exógenos (la edad, la condición de actividad laboral y económica del estudiante, la formación académica previa, las aspiraciones y motivaciones individuales por ejemplo). Y la importancia de cada uno de estos factores variará según el momento o etapa de cursada.

En todo caso, los datos constituyen una alerta y ubican a la Universidad en la necesidad de una indagación más profunda para acercarse a la comprensión más acabada del problema y su mejora.

En el marco de la explosión cuantitativa y el análisis de calidad de la educación superior, crece el interés por estudiar los factores asociados al rendimiento académico en estudiantes universitarios, a fin de ofrecer herramientas de trabajo a futuras investigaciones en este campo, desde un enfoque más integral sobre el desempeño estudiantil.

1.3. La Universidad de General San Martin (UNSAM):

La UNSAM es una universidad nacional, pública y gratuita creada en 1992. Ofrece una amplia gama de carreras de grado y posgrado, tanto en el ámbito de las Ciencias Humanas y Sociales como en el de las Ciencias Exactas y Naturales. Más del 65% de sus recursos están destinados a las áreas de la ciencia y la tecnología. Con el foco puesto en las políticas de trabajo conjunto, mantiene relaciones con agencias que promueven la investigación científica y la transferencia tecnológica. Sus alianzas con el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), la Comisión Nacional de Energía Atómica (CNEA), el Instituto Nacional de Tecnología Industrial (INTI) y el Instituto Nacional de Tecnología Agropecuaria (INTA), entre otros, perfeccionan su producción teórica, el desarrollo de sus investigaciones y la formación de recursos humanos. Su sede está en el Partido de General San Martín, al noroeste de la Ciudad Autónoma de Buenos Aires. Construido sobre una antigua playa ferroviaria de más de dos hectáreas, el Campus Miguelete constituye uno de los principales atractivos del territorio bonaerense en términos de planeamiento arquitectónico y conservación patrimonial.

La UNSAM se caracteriza por ser una institución que apuesta a la formación académica de excelencia en sus programas de pregrado, grado y posgrado. Ésta reconoce a los saberes y aprendizajes que se generan por fuera de los dispositivos académicos clásicos, lo que presupone concebir a la universidad como un espacio de convergencia de distintos tipos de conocimientos y saberes con los cuales el saber universitario puede entrar en diálogo. Entre sus objetivos se encuentra la formación de profesionales con capacidad crítica, autonomía intelectual y compromiso social. La reflexión sistemática sobre el sentido de los procesos de enseñanza y aprendizajes universitarios forma parte de la experiencia formativa en UNSAM.

Asimismo, esta actividad sustantiva se nutre y se mantiene actualizada a través de la interacción permanente con las funciones de investigación, a partir de la generación de conocimientos, y de promoción del desarrollo tecnológico y social, en la medida en que se busca formar profesionales y académicos capaces de atender a las demandas y necesidades sociales. La Universidad promueve el desarrollo de prácticas pedagógicas innovadoras para todas sus ofertas así como de dispositivos alternativos de formación tales como: actividades de formación independiente con reconocimiento institucional, trabajos de campo profesional supervisado, desarrollo de proyectos y desempeño profesional, creaciones/producciones vinculadas a su formación específica. (Comisión Asesora del Consejo Superior de la UNSAM, 2018).

Durante los últimos años el sistema universitario argentino experimentó importantes transformaciones con la creación de nuevas Universidades Públicas en el área metropolitana, específicamente en distritos del conurbano bonaerense caracterizados por elevados niveles de pobreza y demandas educativas insatisfechas: Moreno, Merlo, José C. Paz, Avellaneda, Florencio Varela, entre otros. Teniendo en cuenta dicha coyuntura, surge un nuevo desafío que afrontan todas las Universidades Públicas que consiste en garantizar una formación académica de calidad con altos niveles de inclusión social de los sectores que tradicionalmente quedaron excluidos de la educación universitaria. En este escenario, la UNSAM transita los desafíos de la democratización de la educación superior que reclaman los tiempos actuales para las instituciones universitarias.

La determinación de la cantidad de estudiantes de las Universidades Nacionales se ha constituido en un inconveniente estadístico que presenta un trasfondo de definiciones conceptuales con incidencia sobre la política académica respecto a los criterios a tener en cuenta en dicha definición. En la actualidad existen diversas maneras de precisar la cifra de estudiantes de pregrado y grado que tiene la UNSAM, que se podrían reducir fundamentalmente a dos. Una de ellas refiere al empleo del concepto de estudiantes activos. Según esta perspectiva, son *estudiantes activos* aquellos que registraron algún tipo de actividad académica durante un ciclo lectivo dado, entendiéndose por tal, por ejemplo, haber completado el trámite de inscripción o reinscripción establecido por la institución, haberse anotado para rendir una materia, seminario o examen final, presentar una tesis o trabajo final

u otras actividades académicas definidas por la institución, etc. Sin embargo, este enfoque no permitiría dar cuenta de la especificidad de las diversas modalidades de permanencia y trayectoria de los estudiantes y su grado de arraigo con la Universidad durante ese período académico. Por el contrario, si se emplea el concepto de *estudiantes regulares* para la contabilización –que incluye en la definición un estándar de rendimiento académico esperado- la cifra de estudiantes de la UNSAM se reduciría sustancialmente. Es decir, esta noción se construye a partir del desempeño del estudiante y en este caso muestra las dificultades que tienen los estudiantes de la Universidad para aprobar la cantidad de materias necesarias para conservar la regularidad (2 materias). Del mismo modo que la categoría estudiantes activos, la de estudiantes regulares presenta algunas inconsistencias al no contemplar la situación de aquellos que no cumplen con el rendimiento académico requerido por la Universidad pero que desarrollan actividades académicas y persisten en continuar con sus estudios aprobando al menos 1 materia durante el año lectivo (Roig, A., 2013).

Se puede analizar cómo las dificultades evidenciadas en el rendimiento académico y el consecuente "desgranamiento", repercuten en la graduación de los estudiantes de la UNSAM. En términos globales se observa una tendencia creciente en la producción de egresados, año a año. Es interesante resaltar que las mujeres representan alrededor del 60% de los Egresados, superando la presencia que tienen anualmente entre los nuevos inscriptos que ronda el 50% (Fuente: sistema SIU guaraní UNSAM).

En este sentido, el análisis del rendimiento académico se presenta como la puerta de ingreso para reflexionar sobre un trasfondo más amplio vinculado a los alcances reales del derecho a la universidad y su cumplimiento.

Metodología y datos para el desarrollo del modelo de predicción del abandono.

En el presente apartado se presentan los datos y los modelos que se probarán en la aplicación de la predicción. Con respecto a los datos se cuentan de dónde se obtuvieron los mismos y cómo se prepararon y dispusieron para llevar a cabo el modelo. Asimismo, se desarrollan los atributos que se tuvieron en cuenta y cómo se estructuraron. Finalmente se hace una introducción a los modelos y se describe brevemente cómo opera cada uno.

2.1. Fuente y preparación de los Datos:

La principal fuente de datos para llevar a cabo esta investigación la constituyen los registros históricos de los 8960 alumnos de la Escuela de Ciencia y Tecnología, su información académica y datos personales, recabada por la Universidad Nacional de San Martín, durante el período académico correspondiente a los años que van desde 1994 hasta 2018.
Cabe señalar que la información utilizada fue provista con consentimiento de la universidad y que los datos sobre la identidad de los alumnos fueron anonimizados para preservarlos.

Cuadro 1: Variables de Base de Datos

Año académico de inscripción
Fecha de inscripción (fecha)
Carrera
Duración teórica carrera
Edad a la fecha inscripción
Edad actual
Género
Máximo nivel de estudio del padre
Máximo nivel de estudio de la madre
Máximo nivel de ambos padres
Actividad durante la semana
Horas de trabajo (Semanales)
Última actividad (fecha)
Total materias aprobadas

La fase de análisis de datos ha comprendido la recolección de los datos de los estudiantes del sistema de información de la Universidad SIU (Sistema de Información Universitario) Guaraní.

Inicialmente se rescataron más de 20 atributos, sin embargo, manualmente se descartaron algunos de ellos ya que son considerados como irrelevantes para el estudio.

Para reordenar los datos, darles una visión más limpia y poder ejecutarlos en el programa, se procedió a eliminar algunas columnas que no agregaban valor al análisis y a crear otras. En este sentido, se toma la columna "calidad", que clasifica al alumno entre "Activo / Abandono o Pasivo / Egresado / No Regular", y a partir de ella se crea la columna objetivo definida como "Abandono" la cual es una variable binaria (se completa con 0 si el alumno sigue activo en la carrera y con 1 si el alumno abandonó sus estudios). Con respecto a ello, es importante destacar el supuesto que se ha aplicado dado que será determinante para las conclusiones. En este caso, la Universidad no considera que los alumnos puedan perder la regularidad, por lo que por ejemplo podemos ver un alumno inscripto en el año 1994, cuya última materia fue rendida en ese año y ser calificado como "Activo". A partir de este análisis se procedió a definir que, si el alumno no habría aprobado alguna materia en alguna de las formas (promocionado cursada, aprobado final o aprobado cursada) durante los últimos dos años, sería clasificado como si hubiera abandonado sus estudios (Abandono = 1). Consecuentemente con ese criterio, se crea la columna final que se define como objetivo en el modelo.

Finalmente, se obtiene una base de datos más balanceada con un 60% de alumnos que Egresaron o están Activos y un 40% de alumnos que Abandonaron.

En el mismo proceso de preparación de datos, se le dio un tratamiento especial tanto a los valores outliers como a los valores perdidos, el cual se define en el siguiente apartado.

Posteriormente, se realiza un proceso de selección de las variables de entrada para el modelo. Así se determina cuáles, de todas las obtenidas en el análisis de datos realizado, presentan una mayor relevancia para este estudio, o cuáles aportan una información redundante o secundaria.

Con respecto a las columnas "Egresado" y "Activo", fueron eliminadas de la base, previo a que se hayan extraído los datos de ellas para definir correctamente la columna "Abandono",

dado que guardan una correlación directa con la variable a predecir, y esto afecta a la modelización.

Con respecto a los modelos de minería de datos, se ponen a prueba diferentes modelos predictivos. Por otro lado, se definen las variables y el tipo de dato que contienen las mismas. Asimismo, se define la métrica para valuar los modelos. A partir de ello, se evalúan los distintos métodos y algoritmos, y se determina cuál es que predice mejor a partir del criterio definido.

2.2. Estructuración de los Datos:

Las técnicas de minería de datos aplicadas a la tarea de la predicción, tienen como objetivo desarrollar un modelo que permita predecir el valor de la variable de entrada (variable dependiente) en función de un conjunto de variables predictoras (variables independientes). En el dominio educativo, particularmente en la presente investigación, un modelo predictivo del rendimiento académico tiene como finalidad estimar el valor de la variable ¨abandono¨ que describe si el alumno continuará o no con sus estudios.
Las unidades de creación de la estructura de minería de datos son las columnas, que describen los datos que contiene el origen de datos. Estas columnas contienen información respecto al tipo de datos, el tipo de contenido y el modo en que se distribuyen los datos.
En el cuadro que sigue se muestran la estructura de los datos para poner a prueba los modelos propuestos en los párrafos anteriores, con el objeto de evaluar el comportamiento de cada modelo.

Cuadro 2: Atributos para la estructura de minería de datos

Atributo	Uso	Tipo de Dato	Tipo de Contenido
Año académico de inscripción	Entrada	polynominal	Discreto
Fecha de inscripción (fecha)	Entrada	Fecha	Continuo
Carrera	Entrada	Text	Continuo
Duración teórica carrera	Entrada	Real	Continuo
Edad a la fecha inscripción	Entrada	Entero	Continuo

Edad actual	Entrada	Entero	Continuo
Género	Entrada	Text	Discreto
Máx. nivel de estudio del padre	Entrada	Text	Discreto
Máx. nivel de estudio de la madre	Entrada	Text	Discreto
Máx. nivel de ambos padres	Entrada	Text	Discreto
Actividad durante la semana	Entrada	Text	Discreto
Horas de trabajo (Semanales)	Entrada	Text	Discreto
Última actividad (fecha)	Entrada	Fecha	Discreto
Total materias aprobadas	Entrada	Entero	Continuo
Abandono (Objetivo)	Predicción	Binomial	Discreto

Una vez definidas las variables y el tipo de dato que contienen las mismas, se procede a cargar la base en el software para comenzar con el proceso de modelado.
El primer paso es la carga del data set, es decir los datos que se van a poner a prueba. Luego, previo a la aplicación del modelo, se incorpora el operador "Split Data", con el cual se divide la base de datos en un 80% para entrenar el algoritmo y con el 20% restante se probará cuán eficiente fué, es decir en cuanto acertó sobre lo que predijo. Particularmente para el modelo SVM hay que pasar de nominales a numéricas todas las variables y además normalizarlas para que el algoritmo pueda modelar correctamente.

En cuanto a la métrica a definir para valuar los modelos, se utilizará el Área bajo la Curva ROC (AUC) dado que lo que se quiere lograr a partir del resultado es tratar a los alumnos que más probabilidad tienen de abandonar. Justamente la métrica que se adopta ordena la predicción en función a cuán probable es que se cumpla, en este caso que abandone los estudios. Por ello, en este caso el área bajo la curva ROC da una mejor solución para comparar la performance de los distintos modelos.

2.3. Los Modelos:

En esta investigación, las técnicas de minería de datos para la clasificación que se proponen aplicar con la finalidad de predecir el abandono de los estudiantes son: Regresión Logística, Arboles de Decisión, Random Forest, SVM y KNN. A continuación, se presentará una descripción general sobre el proceso de aprendizaje y la inferencia para la predicción de la

clasificación de nuevas instancias en el ámbito educativo para cada una de las técnicas propuestas.

El modelo de Regresión Logística (RL) permite estudiar la dependencia funcional entre una variable dependiente categórica Y (con dos clases) y un conjunto de "p" variables independientes o predictoras X= (X1i , X2i ,...,Xpi) que pueden ser cuantitativas o categóricas. El modelo de una regresión logística binaria, permite predecir en términos de la probabilidad la ocurrencia del evento de interés (Y=1, Abandona). El proceso de inferencia con la RL, consiste aplicar la ecuación estimada al vector de datos para predecir la clasificación del rendimiento académico de un estudiante (Abandona o No Abandona).

Con respecto al Arbol de Decisión (AD), se define al mismo como un modelo jerárquico para el aprendizaje supervisado, que puede ser aplicado para un problema de regresión o clasificación. Un árbol de decisión es un modelo no paramétrico, puesto que no se asume ninguna forma paramétrica para las densidades de la variable clase y la estructura de árbol no se fija a priori, sino que se va generando durante el proceso de aprendizaje y que depende de la complejidad del problema inherente a los datos (Alpaydin, 2010). El modelo para un árbol de clasificación, se presenta como una estructura jerárquica para mostrar y establecer las relaciones entre la variable de dependiente y el conjunto de variables predictoras. El AD está compuesto por el nodo raíz que se presenta en la parte superior, un conjunto de nodos internos asociados cada uno a una variable predictora, y cuyas ramas representan validaciones o decisiones de los valores de la variable y un conjunto de nodos hojas o nodos terminales, que están etiquetadas con algún valor de la clase de la variable dependiente. La inducción de un árbol, consiste en el proceso de su construcción a partir de un conjunto de entrenamiento.

La técnica de Random Forest (RF) divide cada nodo en un árbol de decisión, el bosque aleatorio solo considera un subconjunto aleatorio de todos los atributos en el conjunto de entrenamiento. Para reducir el error de generalización, el algoritmo se aleatoriza en dos niveles, selección de registros de entrenamiento y selección de atributos, en el funcionamiento interno de cada clasificador base. En general, el modelo funciona en varios pasos. Primeramente, si hay n registros de entrenamiento con m atributos, y sea k el número de árboles en el bosque, entonces para cada árbol se selecciona una muestra aleatoria n con reemplazo. Luego se selecciona un número D, donde D << m. D determina el número de

atributos a considerar para la división de nodos. Además, se inicia un árbol de decisión. Para cada nodo, en lugar de considerar todos los atributos m para la mejor división, se considera un número aleatorio D de atributos. Este paso se repite para cada nodo. Finalmente, como en cualquier conjunto, cuanto mayor es la diversidad de los árboles base, menor es el error del conjunto. Una vez que se construyen todos los árboles en el bosque, para cada nuevo registro, todos los árboles predicen una clase y votan por la clase con igual peso. La clase más predicha por los árboles base es la predicción del bosque (Vijay & Bala. 2015).

El algoritmo SVM funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro. Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVMs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de defnir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento. Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantiza una solución única. (Carmona Suarez, 2013)

Finalmente, el modelo de KNN sirve esencialmente para clasificar valores buscando los puntos de datos más similares (por cercanía) aprendidos en la etapa de entrenamiento y hacer conjeturas de nuevos puntos basadas en esa clasificación. Cualquier registro en un conjunto de datos se puede visualizar como un punto en un espacio n-dimensional, donde n es el

número de atributos. Si bien es difícil para los humanos visualizar en más de tres dimensiones, las funciones matemáticas son escalables a cualquier dimensión y, por lo tanto, se puede realizar todas las operaciones que se pueden realizar en espacios bidimensionales en el espacio n-dimensional. Se necesita un algoritmo eficiente para resolver casos de esquina y medir la proximidad de los puntos de datos con más de dos dimensiones. Una técnica es encontrar el punto de datos de entrenamiento más cercano a partir de un punto de datos de prueba invisible en un espacio multidimensional, y usar el valor de la clase objetivo del punto de datos de entrenamiento más cercano como la clase objetivo pronosticada para el punto de datos de prueba. Esto es similar a cómo funciona el algoritmo k-NN. La k en el algoritmo k-NN indica el número de registros de entrenamiento cercanos que deben considerarse al hacer la predicción para un registro de prueba sin etiqueta. Cuando k = 1, el modelo intenta encontrar el primer registro más cercano y adopta la etiqueta de clase del primer registro de entrenamiento más cercano como el valor previsto de la clase objetivo. Como la clase del registro objetivo se evalúa mediante votación, a k generalmente se le asigna un número impar para un problema de dos clases.

Aplicación del Modelo y Resultados

Como apartado final, se plantea el análisis descriptivo de los datos y la base final a la que se arribó. Además, se incluyen gráficos de las variables más relevantes como otra forma de presentar la información. Por otro lado, se comenta de qué forma se diseña el proceso para llevar a cabo los modelos de minería de datos elegidos. Finalmente se describen los resultados y se analiza y justifica el mejor modelo elegido.

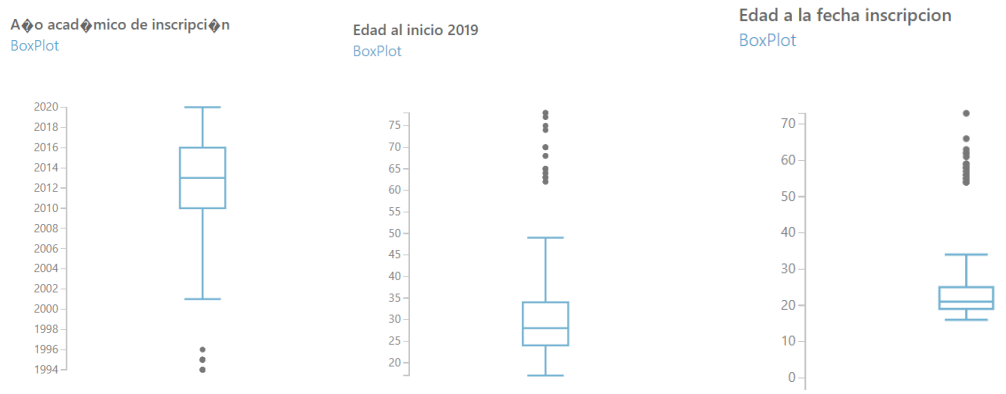
3.1. Análisis de Datos:

Como primer paso, se realiza un análisis estadístico descriptivo de cada uno de los atributos incluidos en la información. El propósito de este análisis es estructurar un modelo predictivo capaz de arrojar antecedentes significativos vinculados con el abandono estudiantil de los alumnos de la Universidad de San Martin. Para ello, se creó un modelo sustentado sobre la determinación de las variables significativas referentes a la conducta de entrada de los estudiantes que ingresan a la universidad. Previo a este análisis se estudió la relación de cada una de las variables explicativas (variables independientes) con la variable abandono estudiantil (variable dependiente) de los estudiantes de la UNSAM con el mismo modelo que se aplicó para realizar la predicción.

Inicialmente se contaba con 8605 registros para el análisis, a partir de los cuales se descartaron algunos y se concluyó con una base de 7.930 alumnos. Algunos atributos fueron eliminados dado que se consideraban redundantes (como ser las columnas ´materias promocionadas´ y ´cantidad de materias regularizadas´ se eliminaron y se integró en una sola columna ´cantidad de materias aprobadas´) y algunos otros se incorporaron a la base (por ejemplo, la columna ´cantidad de materias de la carrera´ y ´duración teórica de la carrera´) con el objetivo de ayudar al algoritmo a generar mejor la predicción. La base fue trabajada en una hoja de Excel y luego para correr los algoritmos se usó el software RapidMiner Studio.

Por otro lado, se usó la herramienta de Microsoft Azzure Machine Learning (AML) para realizar estadística descriptiva de los datos. A partir de ello, se detectaron datos atípicos (observaciones extremadamente grandes o pequeñas, que dista del resto de valores), con diagramas de Pareto.

Gráfico 1: Diagramas de Pareto de Variables Año de Inscripción, Edad al 2019 y Edad a Inscripción.



Como se puede observar en el Gráfico 1, se encontraron algunos valores atípicos en tres variables (estudiantes inscriptos en 1994, 1995 y 1996, estudiantes inscriptos a los 70 años y estudiantes con 70 años en el 2019), los cuales fueron eliminadas de la base con el fin de no distorsionar el modelo.

Por otro lado, se aplicaron las funciones estadísticas a las variables numéricas pudiendo obtener así algunas características de la base de datos como muestra el cuadro que sigue:

Cuadro 3: Estadística Descriptiva de Variables cuantitativas:

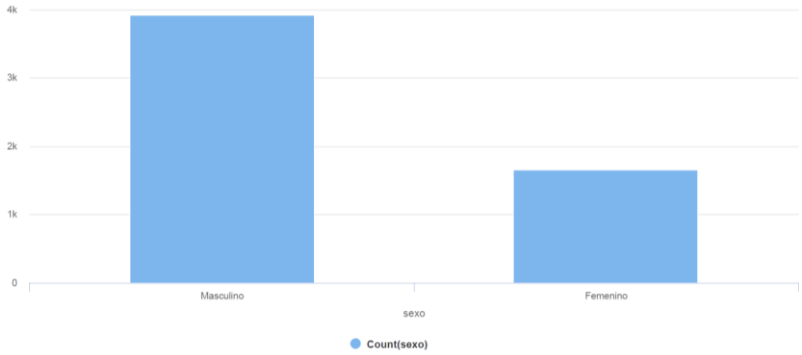
	EDAD a fecha inscripción	Total Materias Aprobadas	EDAD inicio 2019	Duración teórica Carrera
Valor Max	73	47	78.00	5.5
Valor Min	16	1	17	1
Valor Promedio	23.03391117	9.487864078	29.55270151	4.846577249
Desvío	6.274886286	9.974750133	7.872559662	1.114390404
Varianza	39.3741979	99.49564022	61.97719562	1.241865972
Moda	18	1	25	5.5
Mediana	21	5	28	5.5

Como se observa en el Cuadro 3, se presentan alumnos en el rango de edad desde los 73 años a los 16 (pudiendo ser ésta última edad correspondiente al caso en el que el alumno se anota durante el secundario para adelantar materias de la carrera de grado) promediando la edad a la fecha de inscripción a la universidad los 23 años. Siguiendo con el análisis de la

edad, se observa que al inicio del 2019 se cuenta con edades de alumnos desde los 78 a os 17 años, siendo su promedio los 29 años. Por otro lado, se observa que los alumnos cuentan con 9 materias aprobadas en promedio, 47 materias es el máximo aprobado de un alumno y 1 el mínimo. Finalmente, con respecto a la duración de la carrera se observa que en promedio las mismas duran 4,8 años.

En el Gráfico 2 que sigue se puede observar de qué forma se distribuye la variable género en la universidad:

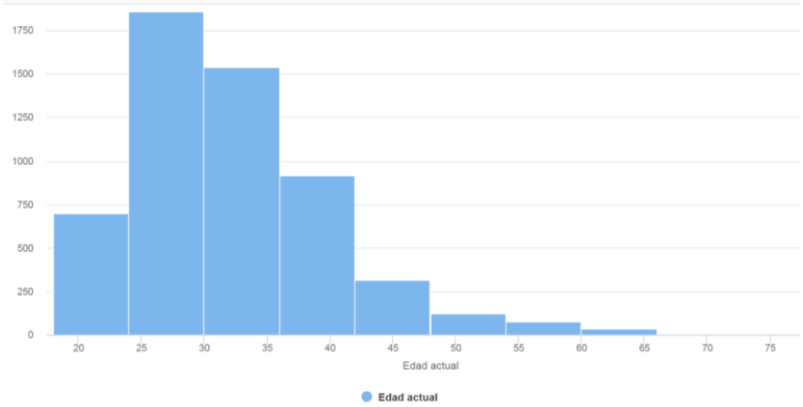
Gráfico 2: Distribución del Género



Analizando dicho Gráfico, se observa que se tiene la muestra dividida en 30% para el género femenino y 70% para el género masculino.

Por otra parte, con respecto a la variable Distribución de la Edad al Inicio del año 2019 se representa en el siguiente Gráfico:

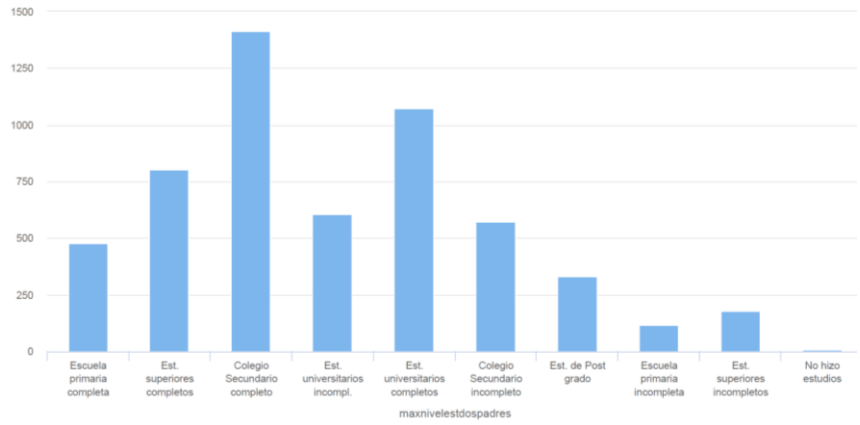
Gráfico 3: Distribución de la edad al Inicio 2019



En el Gráfico 3 se observa que la mayoría de la mayor densidad se alumnos se encuentra entre los que tienen edad entre 24 y 36 años. En las colas de la distribución se ubican los que tienen entre 19 y 21 años y los que tienen a partir de 50.

En el siguiente Gráfico se analiza el máximo nivel de estudios que alcanzaron los padres de los alumnos:

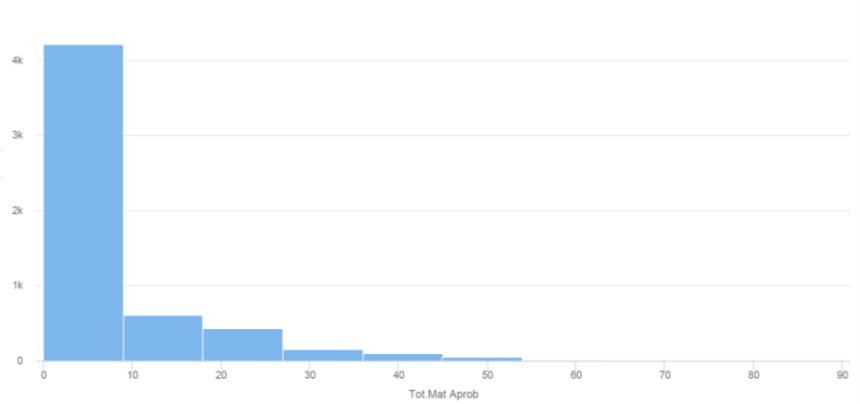
Gráfico 4: Máximo nivel estudiantil alcanzado por ambos padres



En función al Gráfico 4, se observa que por lo menos un cuarto de los padres de los estudiantes completó el colegio secundario y un 20% de los mismos ha completado la universidad.

En el Gráfico que sigue se presenta la variable cantidad de materias aprobadas por los alumnos:

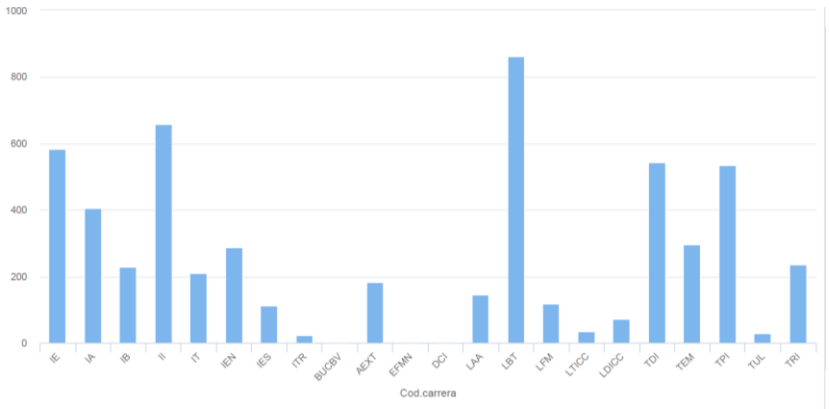
Gráfico 5: Total de materias aprobadas



En función al Gráfico 5, se nota que un 75% de los alumnos cuenta sólo con entre 0 a 9 materias aprobadas. Contrariamente los alumnos que poseen más de 40 materias aprobadas (siendo el promedio de materias por carrera 30) son sólo un 10% del total.

Finalmente se presenta el Gráfico de la distribución de los alumnos según la carrera a la que están inscriptos:

Gráfico 6: Distribución Carreras



En el Gráfico 6 se observa que la mayor cantidad de alumnos se centra en las carreras Licenciatura en Biotecnología, Ingeniería Industrial y Ingeniería Electrónica, siendo las que cuentan con menor cantidad de alumnos Especialización en Física de la Medicina Nuclear e Ingeniería en Transporte

3.2. Aplicación de modelos:

Los atributos categóricos o polinominales fueron transformados a binominales. De esta manera, para cada atributo se generaron n nuevas columnas, donde n es la cantidad de distintas categorías únicas del atributo. De estas n nuevas columnas se seleccionaron $n - 1$ para evitar problemas de multicolinealidad. Una vez definidas las variables y el tipo de dato que contienen las mismas, se procedió a cargar la base en el software para comenzar con el proceso de modelado.

Para la generación del modelo predictivo, se utiliza RapidMiner Studio. Éste cuenta con una gran librería de algoritmos de máquinas de aprendizaje. Para una misma máquina existen distintos algoritmos, cuya configuración está descrita en la sección de ayuda de la plataforma y pueden ser usados según la necesidad del usuario. Todos los algoritmos reciben como entrada la base de datos con la que aprenderán y validarán el modelo generado. Sin embargo, para que tengan un correcto funcionamiento, deben ser configurados a través del ingreso de parámetros que varían según el algoritmo de la máquina de aprendizaje. Es aquí donde se aplicaron las distintas configuraciones de los parámetros anteriormente descritas y se obtuvo el conjunto que obtenía el mayor desempeño de los modelos. Se probaron los modelos citados en el apartado anterior y se optimizaron los hiperparámetros de cada uno, tomando como variable objetivo a "Abandono". Asimismo, se tomó el 80% de los datos, para entrenar el modelo y el 20% restante para probar y validar el mismo, con el operador ¨Split Data¨.

Particularmente para el modelo SVM se debe pasar de nominales a numéricas todas las variables y además normalizarlas para que el algoritmo pueda modelar correctamente.

En función a lo descripto en el apartado anterior, se probaron los modelos Random Forest, Arboles de Decisión, KNN, SVM y Regresión Logística. Para cada modelo se generó un ranking de los mismos considerando la precisión de predicción. Para la medición de los indicadores de desempeño se definió como positiva la clase Abandona y negativa la clase No Abandona. De esta manera, el software arroja la predicción y la confianza que tiene esa predicción (es decir la probabilidad de que la predicción sea acertada).

3.3. Resultados

A partir de las pruebas realizadas con los distintos modelos se puede concluir que, usando la métrica Área bajo la Curva ROC, el mejor algoritmo para la predicción definida en el presente trabajo es Random Forest. En este caso, dicho algoritmo mejora la clasificación en casa observación en un 42% respecto al caso en que son clasificados por "azar" (predecir abandono si/ abandono no):

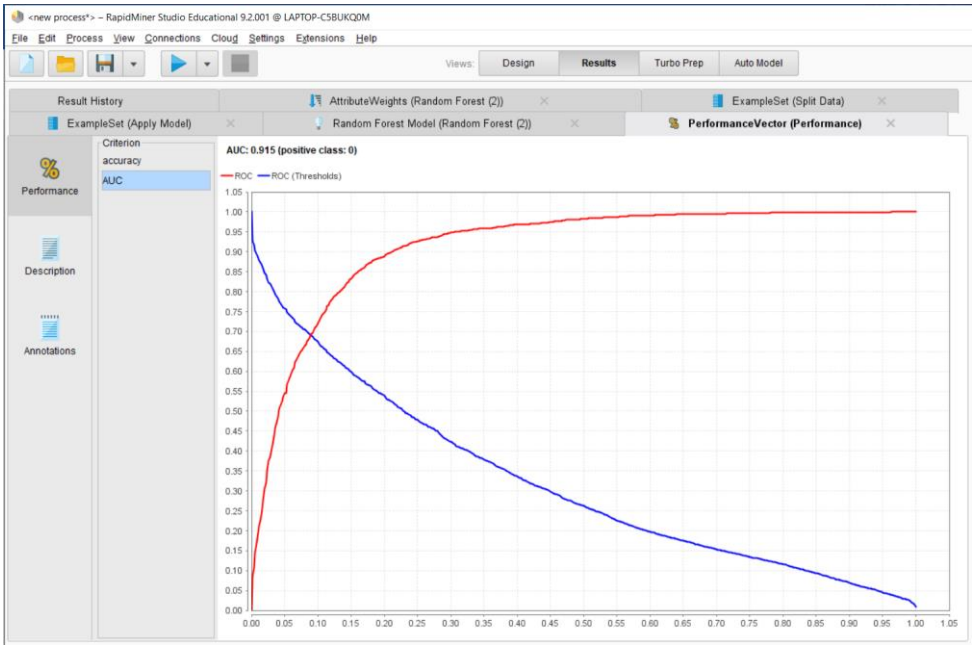
Cuadro 4: Comparativo entre modelos en función de su AUC ROC

	Precisión (AUC)
Random Forest	0.915
Decision Tree	0.823
KNN K=10	0.812
KNN K=20	0.803
SVM	0.791
Regresión Logística	0.743

En el Cuadro 4 se ordenaron los modelos probados según la eficiencia que presenta cada uno. En este sentido, los mismos fueron evaluados bajo el criterio Areá bajo la Cruva ROC. Se concluye que el que mejor predicción arroja es el modelo Random Forest y en contraposición a éste, el peor predictor es la Regresión Logística.

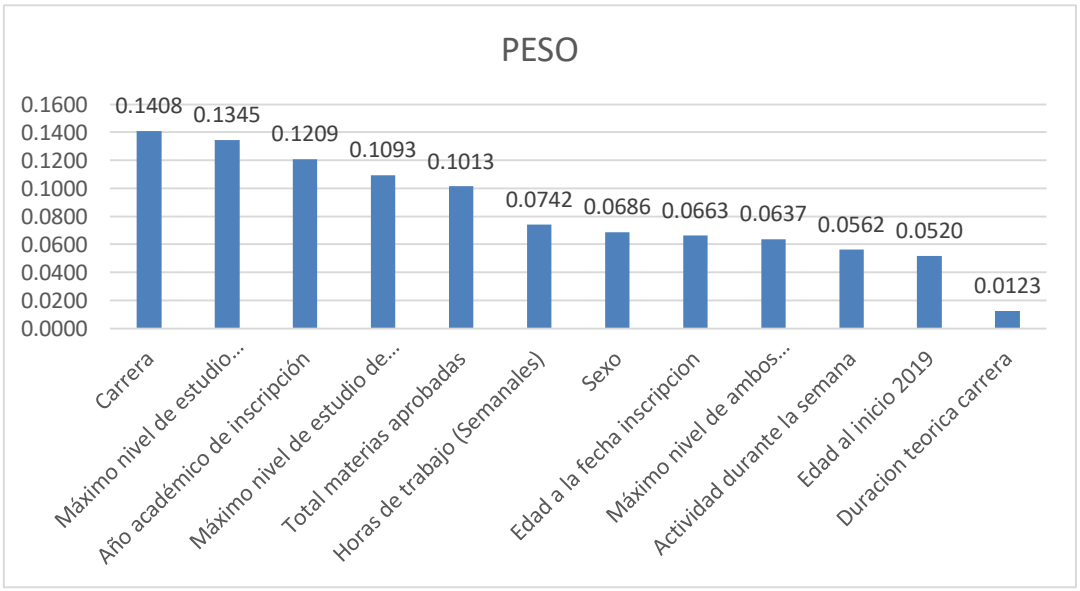
En función a la predicción del modelo Random Forest se presenta el Grafico del AUC ROC:

Gráfico 7: Área bajo la Curva ROC – Random Forest



Las variables que más explican la variable "Abandono" se muestran en el Grafico que sigue:

Gráfico 8: Peso de los atributos en el modelo



En este sentido, se observa que la carrera y la educación alcanzada por los padres es determinante en la continuación de los estudios de los hijos como así también el año

académico de inscripción y las horas de trabajo que deben cumplir. Se puede inferir que dado que el alumno no tiene incentivos o referentes por parte de su familia no sigue adelante estudiando. Por otro lado, las características que tiene cada carrera, serán determinantes a la hora de continuar con los estudios. Otra teoría podría ser que dado que el alumno tarda más en recibirse se desanima y abandona los estudios con el paso de los años.

Discusión y Conclusiones

La deserción es una de las variables más importantes en una institución estudiantil. Muchas son las teorías sobre las que se basa la causa de las mismas, desde el desincentivo familiar hasta el poco apoyo que cuentan los alumnos por parte de la institución. Este estudio tiene el objetivo de mostrar como herramientas de minería de datos pueden ser usadas para generar modelos predictivos que sirvan para apoyar a aquellos estudiantes en riesgo de deserción o de insuficientes desempeños académicos. Las instituciones universitarias podrán utilizar la metodología aquí aplicada para identificar y establecer procedimientos que permitan capturar en forma temprana la información de las variables relevantes con el objeto de mejorar los índices de retención.

En primer lugar, se destaca que la problemática de rendimiento académico afecta principalmente a aquellos estudiantes que estudian la Licenciatura en Biotecnología y la Tecnicatura Universitaria en Diagnóstico por Imágenes, dado que presentan la mayor tasa de abandono. En efecto, se considera que el rendimiento académico impactará prolongando la duración de los estudios con implicancias directas en la disminución de la tasa de graduación por carrera. Por otro lado, se observa que los alumnos con menor cantidad de materias aprobadas (menos de 5) tiene una tendencia a abandonar, resultando el mismo argumento un punto clave en el desincentivo por continuar con sus estudios.

A partir del presente trabajo, verificada que fuese la metodología aplicada, se pueden formular y reformular nuevas investigaciones complementarias al objetivo “abandono”. En particular la expectativa sobre el resultado de este trabajo, es que pueda ser utilizado para implementar diferentes políticas públicas que permitan abordar e implementar dispositivos pedagógicos, didácticos y socio económicos sobre las razones que más se correlacionan con el abandono y evaluar a partir de esa misma base de datos la posibilidad de analizar otras variables tendientes a una mejora continua de la función universitaria. En otras palabras, utilizar el conjunto de estudiantes que los modelos predicen como desertores para focalizar el grupo de individuos que deberían participar en programas de apoyo tales como, talleres de contextualización, programas de apoyo académico y programas de apoyo psicológico.

Finalmente, es necesario señalar que el carácter preliminar de este estudio obliga a buscar mejoras al modelo propuesto, integrando nuevas variables y perfeccionando la fase de recolección de datos. Al respecto, conviene mencionar que para la realización de este este documento no se tuvieron en cuenta las diversas causales tanto endógenas como exógenas que repercuten en el rendimiento académico. Estudios futuros debieran considerar la incidencia de las exigencias, condiciones sociales y económicas que determinan precisamente la heterogeneidad que se expresa en el desempeño académico de los estudiantes. En este sentido y a medida que los sistemas de información lo permitan se deberá considerar la articulación de estas variables con factores tales como los ingresos promedio de los estudiantes y/o de su familia, el tipo de gestión de la escuela secundaria de procedencia, entre otros aspectos que podrían explicar la diversidad del rendimiento académico.

Como propuesta de trabajo futuro, se puede plantear el armado de un código de programación para desarrollar alertas tempranas hacia los alumnos en vías de abandonar sus estudios con el objetivo de abordarlos y brindarles apoyo desde la institución universitaria para que continúen con los mismos. Asimismo, el modelo propuesto abre oportunidades, para la creación de nuevos, modelos de predicción, usando técnicas de clasificación, más complejas como redes neuronales y regresión logística, que permitan un análisis comparativo, de los factores que influyen en la deserción estudiantil.

Referencias bibliográficas

Alpaydin. (2010). Introduction to Machine Learning. Second Edition. Massachusetts Intitute of Tecnology.

Beck, Urlich. (1999) ¿Qué es la globalización? Falacias del globalismo, respuestas a la globalización. Barcelona: Piadós.

Carmona Suarez Enrique (2013). Tutorial sobre Máquinas de Vectores Soporte (SVM). Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED, España).

Comisión Asesora del Consejo Superior de la UNSAM. (Septiembre 2018). El rol transformador de las universidades nacionales: el caso de la UNSAM. Consejo Superior de la UNSAM.

Díaz, M., Peio, A., Arias, J., Escudero, T., Rodríguez, S., Vidal, G. J. (2002). Evaluación del Rendimiento Académico en la Enseñanza Superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU. Revista de Investigación Educativa, 2(20), 357-383.

Fanelli, A. (Junio 2014). Rendimiento académico y abandono universitario: Modelos, resultados y alcances de la producción académica en la Argentina. Revista Argentina de Educación Superior.

Garbanzo Vargas, GM. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. Revista Educación [Internet]. 31(1):43-63.

Greco, C. (2019, 22 de Agosto). Educación superior para transformar la realidad. *Pagina12*.

Hand, D.J.; Mannila, H. & Smyth, P. (2000). Principles of Data Mining. The MIT Press. USA.

Herrera, M.,, Lund, M., Ruiz, S., Mallea, L., Gema Romagnano, M., Torres, E. (2017). Determinación del Rendimiento Académico Universitario. XX Workshop de Investigadores en Ciencias de la Computación.

Kaczynska, M. (1965). El rendimiento escolar y la inteligencia. Madrid: Espasa Calpe.

Luengo, Luis. (2015). Rendimiento de los estudiantes de secundaria obligatoria y su relacion con las aptitudes mentales y las actitudes ante el estudio. UNED.

Murillo, F. J. (2003). Una panorámica de la investigación iberoamericana sobre eficiencia escolar. Revista electrónica iberoamericana sobre calidad, eficiencia y cambio en educación, (1), 1-14

Roig, A. (Agosto 2013). El rendimiento académico de los estudiantes de pregrado y grado de la UNSAM. Heterogeneidad, segmentación y polarización. Secretaría Académica, Dirección de Evaluación e Información Académica.

Vijay Kotu, Bala Deshpande. (2015). Predictive Analytics and Data Mining_ Concepts and Practice with RapidMiner