UNIVERSIDAD ADOLFO IBÁÑEZ
FACULTAD DE INGENIERÍA Y CIENCIAS

"DETECCIÓN TEMPRANA DE ALUMNOS PROPENSOS A LA DESERCIÓN EN LA ESCUELA DE NEGOCIOS Y LA FACULTAD DE INGENIERÍA Y CIENCIAS DE LA UNIVERSIDAD ADOLFO IBÁÑEZ, MEDIANTE EDUCATIONAL DATA MINING Y MACHINE LEARNING"

Alexandra Francisca Araya Rojas
Tomás Ignacio Hasbún Hurtado

Profesor Tutor: Jorge Villalón Dinamarca

PROPUESTA DE MEMORIA PARA OPTAR AL TÍTULO DE INGENIERÍA CIVIL INDUSTRIAL CONCENRTACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN

2015

Resumen Ejecutivo

El Chile la deserción es un fenómeno que ha ido en aumento en los últimos años, lo cual es producto de un efecto secundario del fuerte aumento de jóvenes que ingresan a la educación superior. Sin embargo, este fenómeno resulta perjudicial para todos los actores del sistema de educación terciaria, es decir, los estudiantes y sus familias, el Estado y la institución que escogen para su plan académico. Estos costos fueron estimados por González y Uribe el 2005, quienes concluyeron que las pérdidas asociadas a la deserción de estudiantes que deciden abandonar el plan de estudios son de 47 mil millones de pesos anuales. (MINEDUC, 2012).

Es bajo este contexto país en que el 2013 la Universidad Adolfo Ibáñez (UAI) firmó un convenio con el ministerio de educación relacionado con el Plan de Mejoramiento Institucional que tiene por objetivo desarrollar el proyecto de "El Ingeniero Global", en el cual se fijan metas anuales y una de sus medidas de desempeño corresponde a la tasa de retención de alumnos de primer y tercer año de la Facultad de Ingeniería y Ciencias (FIC) y la Escuela de Negocios, cabe mencionar que ambas facultades comprenden el 77% de los alumnos de la Universidad. (Universidad Adolfo Ibáñez, 2015). Es importante destacar que el incumplimiento de estas metas anuales genera perjuicios para la Universidad tanto institucionales como económicos importantes y es por esto que es fundamental para la UAI poseer un modelo que permita la detección temprana de alumnos propensos a la deserción con el fin de tomar medidas de apoyo hacia estos estudiantes y lograr el término completo de su plan de estudios.

Para lo anterior, se utilizarán técnicas de Data Mining y Machine Learning, específicamente árboles de decisión que permitan encontrar con alto grado de precisión las variables de mayor relevancia en la deserción voluntaria de los alumnos de la FIC y escuela de Negocios de la UAI.

Índice de Contenidos:

Índice de Ilustraciones:

Índice de Tablas:

1 INTRODUCCIÓN

La institución educacional superior Universidad Adolfo Ibáñez, es una institución sin fines de lucro, nacida a partir de la escuela de negocios de Valparaíso, fundada en 1953 por la Fundación Adolfo Ibáñez.

En la actualidad, la Universidad cuenta con dos sedes para impartir los 14 programas de pregrado, estas sedes están ubicadas en: Diagonal Las Torres 2640, Peñalolén, Santiago y Avenida Padre Hurtado 79, Viña del Mar.

Para los programas de pregrado existen cinco facultades más la Escuela de Negocios, las carreras que se imparten son: Diseño, Psicología, Periodismo, Derecho, Ingeniería Comercial, con mención en administración de empresas y mención en Economía, e Ingeniería Civil con seis especialidades distintas que son: Industrial, Energía y Medio Ambiente, Bioingeniería, Informática y Telecomunicaciones, Minería y Obras Civiles.

Cabe destacar que estas facultades también imparten programas de postgrado en conjunto con la Facultad de Artes Liberales y la Facultad de Gobierno, las cuales no cuentan con programas de pregrado. Para estos postgrados, además de las sedes ya mencionadas, existe una tercera sede ubicada en Presidente Errázuriz 3485, Las Condes, Santiago.

La misión del modelo educativo de la Universidad Adolfo Ibáñez es: “Entregar una educación que, basada en la libertad y en la responsabilidad personal, permita a sus estudiantes desarrollar la totalidad de su potencial intelectual y humano. Para lograr esto, la UAI asume el compromiso de impartir una formación profesional con altos estándares académicos, contribuir a expandir las fronteras del conocimiento a través de investigación de alto nivel y transferir estos conocimientos para beneficio de la sociedad.” (Universidad Adolfo Ibáñez, 2015). Los valores esenciales son la libertad y la responsabilidad (Universidad Adolfo Ibáñez, 2015).

1.1 Contexto

El área en el cual se llevará a cabo esta memoria es la Escuela de Negocios, fundada en 1954, y la Facultad de Ingeniería y Ciencias, fundada en 1990, de la Universidad Adolfo Ibáñez. Ambas facultades comprenden el 77% de los alumnos de la Universidad, siendo 45% Ingeniería Comercial y 32% Ingeniería Civil y sus distintas especialidades (Universidad Adolfo Ibáñez, 2015). El total de matriculados en pregrado hasta en el año 2014 fueron 7.533 estudiantes.

En 2013 la Universidad Adolfo Ibáñez firmó, en conjunto con el ministerio de educación un convenio de desempeño en innovación académica, para lograr el plan denominado "El Ingeniero Global: Diseño e implementación de un modelo de formación de Ingenieros para un mundo globalizado", que tiene como objetivo la implementación de un modelo integral según estándares y prácticas internacionales, esto corresponde al Plan de Mejoramiento Institucional (PMI), propuesto por la UAI. Cabe destacar que este plan forma parte del programa MECESUP, el cual busca el mejoramiento de la calidad y equidad en la educación terciaria (MECESUP, s.f.).

1.2 Identificación de necesidades y oportunidades

Para este convenio, la Universidad se comprometió a una serie de metas anuales de indicadores de desempeño dentro de los cuales están el porcentaje de retención de alumnos de primer y tercer año de Ingeniería Civil y Comercial. Esto forma parte importante del PMI y es de vital importancia, puesto que actualmente en Chile la deserción es considerada un problema país, la cual ha ido aumentando fuertemente en las últimas décadas debido a que es una consecuencia indirecta del fuerte aumento de matrículas de enseñanza superior. Para el año 2012, más del 50% de los matriculados en enseñanza superior no concluye su programa de estudios en el que se matriculó originalmente, lo que en palabras del ministerio de educación se traduce en "… importantes pérdidas de eficiencia para el Estado y las instituciones, así como disminución de oportunidades para los estudiantes y sus familias." (MINEDUC, 2012).

Además de lo anterior, otra variable a tener en consideración es que en caso de incumplir las metas propuestas en el PMI firmado, la Universidad puede sufrir un perjuicio monetario importante para el proyecto que está llevando a cabo.

Es por todo lo anterior que es de suma importancia que la Institución cuente con un modelo que permita la detección temprana de alumnos propensos a la deserción, con el fin de tomar medidas de apoyo para dichos alumnos y así lograr conservar a estos estudiantes dentro del programa.

1.3 Objetivos del proyecto y específicos

El objetivo principal es crear, mediante Educational Data Mining y Machine learning, un modelo de predicción y detección temprana de la deserción y sus causas, en alumnos FIC y de alumnos de Negocios. Para que en un futuro cercano se pueda generar un sistema de retención para estos alumnos.

El proyecto tiene por objetivos específicos:

- Comprensión del contexto de la deserción en la UAI, es decir, los mecanismos existentes dentro de la universidad para desertar, reglamentos asociados a desvinculaciones de alumnos, entre otras variables que sean relevantes.
- Comprender los datos disponibles a analizar y observar variables relevantes con el fin de vislumbrar las primeras señales existentes en estos.
- Preparación y transformación de los datos para limpiar datos fuera de rango, llenar datos en blanco y generar nuevas variables a partir de las existentes que puedan ser de utilidad para el modelo a continuación.
- Aplicar distintas técnicas de modelado de Árboles de Decisión y Clasificadores Baysesianos Ingenuos y la calibración de dichos modelos para la obtención de mejores resultados y futura comparación.
- Evaluar dichos modelos y comparar sus resultados y precisión, con tal de determinar cuál se ajusta de mejor forma a la data original y al contexto de la universidad.
- En base a lo obtenido generar un posible plan de acción para la UAI, de forma que en los próximos años existan nuevas herramientas para combatir el problema.

1.4 Sumario

La deserción en educación terciaria es un problema nivel país relevante en estos últimos años afectando a instituciones de educación superior, así como lo es la UAI. Es por esto, que el Plan de Mejoramiento Institucional firmado el año 2013, contempla metas en relación a los índices de retención de primer y tercer año de alumnos de Ingeniería Civil y Comercial hasta el 2016. Para lograr estas metas la Universidad necesita un modelo de detección temprana de alumnos propensos a la deserción, con el fin de activar sistemas de apoyo y retención hacía estos alumnos, para que finalmente opten por terminar el programa de estudios inicialmente escogido.

2 DESCRIPCIÓN DEL PROBLEMA

En diciembre del 2013 el Ministerio de Educación en conjunto con la Universidad Adolfo Ibáñez firmó un convenio de desempeño en innovación académica en donde el objetivo y plan de mejoramiento institucional fue denominado, "El Ingeniero Global: Diseño e implementación de un modelo de formación de Ingenieros para un mundo globalizado", el cual tiene como objetivo general: "Diseño e implementación de un modelo integral de formación universitaria en STEM (Acrónimo en inglés de Science, Technology, Engineering, y Mathematics) según los estándares y prácticas internacionales más avanzadas: flexible, homologable, interdisciplinario, y motivador, que permita a los egresados crear valor a través de la innovación, el emprendimiento y la práctica de su profesión en contextos de creciente complejidad, incertidumbre y globalización, con foco en ingeniería." (Universidad Adolfo Ibáñez, 2013). Este plan de mejoramiento posee ocho objetivos específicos, en donde el sexto dice relación con "Poner énfasis en la enseñanza y aprendizaje centrado en el estudiante, orientado a la formación a nivel de licenciatura." (Universidad Adolfo Ibáñez, 2013), que tiene como unos de sus indicadores de desempeño la tasa de retención de estudiantes de primer y tercer año de la Facultad de Ingeniería y Ciencias (FIC) y la Escuela de Negocios. Para estos indicadores, partiendo de una línea base, existen metas anuales desde el 2014 hasta el 2016, las que son de suma importancia por diversos factores.

El primero es de carácter social y contempla la deserción como una pérdida tanto personal como familiar para el alumno que abandona el plan de estudio, debido a los esfuerzos realizados por el estudiante y su familia para lograr completar sus estudios. Esto finalmente, conduce a que "los jóvenes que desertan ven truncados sus sueños de graduarse, lo que les genera frustración y descontento." (MINEDUC, 2012)

El siguiente factor es de transcendencia estatal, puesto que el Estado pierde eficiencia en los recursos invertidos en alumnos que no completan sus estudios, estos recursos en algunos casos son: el AFI (Aporte Fiscal Indirecto) otorgado a cada institución en relación al puntaje PSU de sus alumnos matriculados y/o becas estatales que cubren un porcentaje de arancel anual del programa de estudios.

En cuanto al tercer factor, existe una pérdida de índole institucional, puesto que, en caso de incumplimiento de las metas propuestas en el Plan de Mejoramiento Institucional, la UAI se ve afectada en su prestigio frente al ministerio, sus pares, sus docentes y sus alumnos. Lo que además puede llegar a perjudicar la percepción de la Universidad por parte de potenciales alumnos.

Tabla 1
Indicadores y metas de desempeño, objetivo específico N°6 convenio PMI 2013.

NOMBRE INDICADOR	LINEA BASE	META 2014	META 2015	META 2016
TASA DE RETENCION DE ESTUDIANTES FIC DE PRIMER AÑO	81%	82%	83%	84%
TASA DE RETENCION DE ESTUDIANTES FIC DE TERCER AÑO	54%	58%	62%	65%
TASA DE RETENCION DE ESTUDIANTES DE NEGOCIOS DE PRIMER AÑO	83%	85%	87%	90%
TASA DE RETENCION DE ESTUDIANTES DE NEGOCIOS DE TERCER AÑO	72%	75%	78%	83%

Fuente: (Universidad Adolfo Ibáñez, 2013)

Finalmente, existe un factor económico que hace referencia a los recursos que deja de percibir la institución asociados a las matrículas y aranceles de cada alumno que deja el plan de estudios.

Tabla 2
Aranceles y Matriculas 2015 de Ingeniería Comercial y Civil UAI.

CARRERA-SEDE	ARANCEL ANUAL	MATRICULA ANUAL
INGENIERIA COMERCIAL SANTIAGO	210UF	21UF
INGENIERIA COMERCIAL VIÑA DEL MAR	190UF	19UF
INGENIERIA CIVIL SANTIAGO	200UF	21UF
INGENIERIA CIVIL VIÑA DEL MAR	180UF	19UF

Fuente: Elaboración propia.

Debido a esto, como menciona el MINEDUC las instituciones “deben adaptar su funcionamiento en cursos superiores a un menor número de alumnos” (MINEDUC, 2012).

Además, para el caso de la UAI existe un perjuicio económico en caso del incumplimiento de las metas propuestas en el PMI, ya que la Universidad puede dejar de recibir los aportes monetarios por parte del Ministerio de Educación comprometidos al proyecto “El ingeniero global”.

Es importante señalar que el fenómeno de deserción en su totalidad en Chile fue evaluado por González y Uribe (2005) quienes estimaron su costo directo en $47 mil millones de pesos. (MINEDUC, 2012)

2.1 Conclusión

Dados los factores, las dimensiones y sus consecuencias la Universidad Adolfo Ibáñez no puede evadir el problema y necesita plan de acción de forma urgente con tal de atacar eficientemente el problema, es por esto que desarrollar un modelo de alerta temprana de alumnos propensos a la deserción es de carácter prioritario para UAI.

3 MARCO TEÓRICO

3.1 KDD (*Knowledge Discovery in Databases*)

La extracción o descubrimiento de conocimiento potencialmente útil dentro de bases de datos o repositorios de información se denominan KDD (por sus siglas en inglés), y trata sobre el descubrimiento de conocimiento no trivial en extensos sets de datos y es un proceso iterativo que explora de forma exhaustiva dichos grandes volúmenes de datos para determinar relaciones. La información y conocimiento extraído sirve para proponer conclusiones respecto a los modelos que se presentan en los datos. Este proceso consta de cinco etapas explicadas a continuación:

1. Selección de datos: Se establecen las fuentes y los tipos de información con los que trabajará en las siguientes etapas, una vez definido se realiza la extracción de dichos datos.
2. Pre procesamiento: Se realiza una limpieza y preparación de los datos para su posterior uso. Se ocupan diversas estrategias para el manejo de los datos faltantes, datos fuera de rango, obteniéndose una estructura adecuada para el trabajo que le seguirá a continuación.
3. Transformación: Tratamiento preliminar de los datos, transformación y generación de variables a partir de las ya existentes.
4. Data Mining o DM (*Minería de datos*): Es la fase de modelamiento propiamente tal, en donde se aplican modelos inteligentes con el fin de extraer patrones nuevos y de potencial utilidad que se hayan encontrado previamente "ocultos" en los datos.
5. Interpretación y Evaluación: Etapa final en donde se identifican los patrones de real interés aplicando ciertas medidas y se realiza una evaluación de los resultados obtenidos.

Además, se suelen incluir dos etapas extra al proceso, la primera, que está ubicada antes de la selección de datos, consta de un análisis de las necesidades del cliente, de forma de enfocar la investigación en dicha dirección. La segunda etapa está ubicada al final y pretende integrar los resultados obtenidos en el cliente. (WebMining Consultores, 2011)

3.2 CRISP-DM

Por sus siglas en inglés (Cross Industry Standard Process for Data Mining), es una metodología de proceso de Data Mining y está descrita en términos de un proceso jerárquico en seis fases. El proceso de Data Mining suele continuar luego del despliegue de los resultados con el fin de generar nuevas preguntas que pueden traer mayores conocimientos respecto a la temática.

Las seis fases corresponden a:

- Comprensión del Negocio del cliente, en donde se capturan las reales necesidades de este para determinar los requisitos del proyecto.
- Comprensión de los datos, que comienza con la recolección de estos, con el fin de familiarizarse con ellos, detectar su calidad y formular posibles hipótesis en base a observaciones interesantes.
- Preparación de los datos, en la cual se incluyen todas las actividades que hacen relación con el manejo de los datos previo a la aplicación del modelo, como la limpieza de datos atípicos, el relleno de datos blancos o la generación de variables nuevas que puedan ser de utilidad a partir de las previamente existentes.
- Modelado, en donde se seleccionan y aplican distintos modelos y algoritmos de Data Mining en relación a las necesidades del proyecto.
- Evaluación, en la cual se comparan dichos modelos en relación a los objetivos y su precisión, además de hacer posibles ajustes correspondientes para obtener mayor exactitud en los modelos.
- Finalmente se hace un despliegue de la información obtenida, generando posibles implementaciones en el negocio en cuestión, de tal forma de generar valor. (DATAPrix, 2007).

3.3 DM *(Data Mining o Minería de Datos)*

Como se mencionó anteriormente, la minería de datos es una etapa de análisis de KDD y es un campo de ciencias de computación que refiere al proceso automático o semiautomático que intenta recuperar y acceder a información en los patrones de grandes volúmenes de datos, haciendo uso de métodos como la inteligencia artificial, las bases de datos, estadística y Machine Learning para extraer información comprensible y de interés de los datos, debido a que las herramientas tradicionales de tratamiento de datos no tienen la capacidad y los recursos necesarios.

En base a lo anterior, existen diversas técnicas a aplicar sobre los datos para poder extraer la información, todas basadas en la inteligencia artificial y el Machine Learning, pero no son solo más que distintos

algoritmos con distintos niveles de complejidad. Estos pueden ser categorizados en Algoritmos tanto supervisados como no supervisados. Los primeros se caracterizan por predecir características o datos desconocidos a priori a partir de datos conocidos. Por otro lado, los no supervisados descubren patrones y tendencias en los datos. (MINERÍA DE DATOS Y ALMACENAMIENTO WEB, n.d.)

3.3.1 Redes Neuronales

Es un procesamiento automático y paradigma de aprendizaje inspirado en el modo de funcionamiento del sistema nervioso de los animales, emulando el funcionamiento del cerebro. Imita el funcionamiento de neuronas en una red, de forma de que al recibir un estímulo de entrada se genera una respuesta de salida. (Redes de Neuronas Artificiales, 2012)

3.3.2 Regresión lineal

Es una técnica muy utilizada en el ámbito científico, que modela la relación entre la variable dependiente Y y la variable independiente X. Es un método ineficaz para grandes conjuntos de datos debido a que solo es capaz de relacionar dos variables, cuando en los problemas analizados suelen ser de múltiples dimensiones. (Bianca Cung, n.d.)

3.3.3 Árboles de decisión

Modelo de predicción del ámbito de la inteligencia artificial y análisis predictivo, que dado un set de datos construye, de forma lógica, gráfica y analítica, un diagrama que representa y caracteriza los eventos o series de condiciones de forma sucesiva para la resolución del problema presentado. (Departamento de Matemática Aplicada y Estadistica).

Un conjunto de estos árboles genera un “Random Forest” (Selva Aleatoria), y su principal objetivo es disminuir la variación aleatoria generada en distintos árboles de decisión. (Department of Statistics, University of California, Berkeley, n.d.).

3.3.4 Clustering (Algoritmo de Agrupación)

Procedimiento de agrupación de datos o vectores según criterios que suelen ser según distancia vectorial, e intenta agrupar dichos datos según características en común. Suele ser considerado el Método no supervisado de mayor importancia. (Dipartimento di Elettronica ed informazione, n.d.)

3.3.5 Naive Bayes Classifier (Clasificador Bayesiano Ingenuo)

Es un clasificador probabilístico basado en el teorema de Bayes, que asume que la ausencia o presencia de cierta característica no se relaciona con la ausencia o presencia de otra característica. (DELL, 2015)

3.4 Minería de datos educacional

Existe una disciplina dentro de la minería de datos enfocada exclusivamente a la educación y las variables pertinentes a esta. Su principal objetivo es determinar nuevos patrones de aprendizaje, metodologías educativas y cualquier conocimiento nuevo generado a partir de sets de datos educacionales.

3.5 Casos de estudio

3.5.1 Predicting Students Drop Out: A Case Study (2009)

En el departamento de Ingeniería Eléctrica de la Universidad de Tecnología Eindhoven, el ratio de deserción de los alumnos de primer año es cercano a un 40%, y en los Países Bajos existe una obligación legal a las universidades por proveer la guía e información necesaria a los estudiantes para que evalúen su elección de estudio. Por esta razón la Universidad está interesada en descubrir los motivos de la deserción temprana de sus alumnos. Para esto, la Universidad desarrolló un estudio de minería de datos educacional con datos, tanto preuniversitarios como universitarios, de sus alumnos de primer año. Los algoritmos utilizados en este estudio fueron dos árboles de decisión (algoritmo CART y C4.5), un clasificador Bayesiano, un modelo logístico de regresión lineal, un aprendizaje basado en reglas y un “Random Forest”

(Selvas Aleatorias). Estos algoritmos primero se trabajaron primero solo con información pre-universitaria, luego solo información universitaria, y finalmente con un conjunto de información tanto universitaria como pre universitaria. Como resultado se obtuvo que algoritmos más simples, como los árboles de decisión tuvieron mejores predicciones con la información completa, específicamente el algoritmo CART dio una precisión de 0.81 en la predicción de deserción y entregó como variable determinante la aprobación de Álgebra Lineal, información que hasta el momento era desconocida por la Universidad. (Gerben W. Dekker, 2009)

3.5.2 Una metodología de evaluación para una intervención de salud y su efecto en el rendimiento académico de estudiantes universitarios usando Machine Learning (2014)

Desde el año 2002 la UAI ha implementado un programa de intervención deportiva, se desea medir la efectividad de dicho programa en base a la información recopilada de 273 alumnos y su evolución en 9 meses se determinó, mediante Clustering Jerárquico que la intervención tuvo un efecto positivo en la condición física de los alumnos, ya que demostró que un 43% de los alumnos mejoro su desempeño físico y un 51% lo mantuvo. (Gallardo, 2014)

3.5.3 Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques (2010)

En India el 2010 Mr. M. N. Quadri y Dr. N.V. Kalyankar realizaron un estudio mediante árboles de decisión para determinar las principales razones de la deserción escolar en el país, basándose en nueve variables distintas, como lo es, el género, el ingreso familiar o la asistencia a clases y se determinó que el ingreso familiar es la variable más relevante al momento de la predicción. (Kalyankar, 2010)

3.5.4 Mining Educational Data to Reduce Dropout Rates of Engineering Students (2012)

El fuerte crecimiento de estudiantes de ingeniería en India en los últimos años ha causado un incremento en la deserción universitaria de estos mismos alumnos, es por esto, que se decidió generar u modelo de predicción de deserción basado en arboles de decisión llegando a la conclusión de que el factor más relevante en la predicción de la deserción fueron las notas de enseñanza media, lo que generó un árbol con una precisión de 85,7%. (Pal, 2012)

3.5.5 Predicting School Failure and Dropout by Using Data Mining Techniques (2013)

Utilizando técnicas de Data Mining como árboles de decisión y métodos de clasificación en conjunto a información real de 670 estudiantes de enseñanza media de Zacatecas, México, se generó un modelo de predicción de deserción de enseñanza media llegando a un árbol de decisión del 99,7% de precisión, siendo las variables de mayor importancia las notas obtenidas en física, humanidades, matemática e inglés. (Marquez-Vera, 2013)

4 METODOLOGÍA

En este trabajo se llevará a cabo un estudio con el fin de encontrar patrones de la deserción universitaria en pregrado en la Facultad de Ingeniería y Ciencias y la Escuela de Negocios de la UAI. Es importante destacar que el alcance de este trabajo será predecir los alumnos propensos a la deserción de su plan de estudios original, es decir, alumnos que abandonan sus carreras o programas en la UAI, puesto que los alumnos que se cambian de programa internamente no son parte del problema.

Para esto es necesario, en primera instancia, la recopilación de datos desde la base de datos de la plataforma en línea Omega, que posee los datos académicos y personales de los alumnos.

Luego, es importante detectar cuales variables no son de real interés en la investigación, para quitarlas del set de datos.

Después, utilizando ciertas metodologías, es necesario rellenar datos ausentes, quitar datos atípicos y así aumentar la calidad de los datos para el estudio.

Una vez, obtenida la base de datos apropiada se generan nuevas variables a partir de variables originales, que pueden ser útiles para el estudio.

Posteriormente, se procede a realizar la Data Mining, utilizando principalmente algoritmos predictivos de árboles de decisión, clasificador de Bayes ingenuo y “Random Forest”, ya que estos son los que entregan una mayor precisión en estudios similares, manteniendo su nivel de exigencia computacional bajo.

Finalmente, se analizan los resultados obtenidos para verificar su pertinencia con el estudio e identificar su nivel de precisión con la realidad, además se identifican el porcentaje de falsos positivos y falsos negativos, con el fin de concluir que el estudio sea representativo.

5 DEFINICIÓN DE LÍNEA BASE – SITUACIÓN ACTUAL

En la actualidad, la Universidad tiene pérdidas monetarias e institucionales debido a los niveles de deserción presentes tanto en la Escuela de Negocios como en la Facultad de Ingeniería y Ciencias. El primer factor tiene relación directa con los recursos que deja de percibir la Institución por conceptos de matrículas y aranceles anuales de cada alumno que decide desertar. Por otro lado, la Universidad, además pierde un potencial ingreso de estos alumnos a programas de Postgrados. Estos recursos son cuantificados en 441.674 UF anuales aproximadamente, siendo 232.860 UF perdidas por Ingeniería Civil y 208814 UF perdidas por Ingeniería Comercial. Este valor considera los 7.533 alumnos matriculados al 2014, en donde los estudiantes de Ingeniería Civil corresponden al 32% del total, y los estudiantes de Ingeniería Comercial al 45%, sumando el 77% del total de alumnos de la Universidad. (Universidad Adolfo Ibáñez, 2015).

Por otro lado, existe un perjuicio institucional, puesto que actualmente la Universidad se encuentra llevando a cabo el proyecto del Ingeniero Global, junto con el MINEDUC. Es por esto que la Universidad firmó un contrato de proyecto de mejoramiento institucional, el cual tiene metas que deben cumplirse anualmente. Este contrato firmado el 2013, en su objetivo específico n° 6 exige tasas de retención a cumplir a finales del primer y tercer año de las carreras de Ingeniería Civil e Ingeniería Comercial, siendo la tasa de retención al tercer año de ambas carreras la más lejana en cuanto a su situación base del 2013 a la exigida a fines del 2016, las cuales deben subir 11 puntos porcentuales. (MECESUP, s.f.). Cabe mencionar que, de no cumplirse las metas propuestas en el contrato, la institución puede verse afectada tanto monetariamente como en su prestigio, debido a que el MINEDUC se comprometió a entregar aproximadamente 618 millones de pesos para el proyecto, los cuales se entregan por cuotas, y pueden dejar de pagarse. Además de que la Universidad se vería claramente afectada por no cumplir un acuerdo con el ministerio de educación.

Se espera que una vez finalizado este trabajo, la Universidad tenga las facultades para detectar tempranamente alumnos propensos a una deserción en un corto o mediano plazo. Puesto que, el objetivo es detectar patrones de comportamiento en alumnos que ya han desertado, mediante Machine Learning y Educational Data Mining, y así predecir posibles deserciones futuras en alumnos que actualmente estén cursando alguna de estas carreras y llevar a cabo un plan de acción de retención efectivo hacia estos alumnos.

5.1 Conclusión

De lo anterior se puede concluir que la deserción es un problema de gran importancia para la Universidad, puesto está en juego sus futuros recursos y su prestigio institucional ante el MINEDUC. Ante lo cual se espera que terminado este trabajo la Institución tenga las herramientas para poder detectar a alumnos que posiblemente desertarán en un corto o mediano plazo, para llevar a cabo un plan de acción de retención con estos alumnos.

6 Caracterización y análisis de los datos

Los datos a utilizar provienen directamente de la Universidad y constan de seis tablas con la información de los alumnos de los años 2008 hasta 2011 respecto a su situación académica, sus asistencias a deporte, su malla, información ramo a ramo tomado por los alumnos y su procedencia escolar. Cabe destacar que esta información se anonimizó, es decir, se borró toda información como el RUT, nombres, apellidos, mails personales, números de teléfono, etc. Se conservaron los Id puestos por la universidad a cada alumno, con el propósito de cruzar las tablas.
Esta información de consolidó en una tabla con un vector por alumno, el cual contenía la mayor información posible respecto a las variables de interés, como créditos aprobados y reprobados, diferenciados por tipo de ramo (regular, complementario, ingles o deporte), puntajes PSU, asistencias de deporte, deporte favorito y eficiencia en la aprobación de ramos en sus categorías, así como otras variables.
Esta información fue primero filtrada en base a que los alumnos deben ser solo de Ingeniería y estos debiesen haber terminado cuarto año, o haber desertado de la carrera, es decir, se eliminan los alumnos que están actualmente activos en pregrado.

Analizando las variables e información otorgada, se hizo notoria la necesidad de re-calcular ciertas variables que contenían inconsistencias. Una de estas variables fue el porcentaje de avance del alumno en su carrera, ya que en ciertas generaciones el cálculo se hizo tomando en cuenta ramos complementarios y en otras no, por lo que se tomó la decisión, dado que se tenía acceso a los ramos tomados por los alumnos, a re-calcular el porcentaje solo considerando los ramos complementarios y la base de datos de las distintas carreras y sus créditos totales.

Luego de la limpieza de la base de datos, su unión en vectores, y el filtro hecho en base a los criterios mencionados antes, los alumnos que entraron en el estudio son 4840 aproximadamente.

7 Definición de los modelos

En base a la literatura antes mencionada los modelos a ocupar serán principalmente arboles de decisión de distintos tipos y Random Forest. Los tipos de algoritmo de árbol a ocupar son específicamente el CART, C4.5 y C5.0. Esto debido a que de los algoritmos de machine learning existentes los arboles de decisión permiten observar directamente las variables predictores de una forma simple y entendible, de forma de poder entender de real forma el comportamiento del modelo.

Para validar estos distintos modelos, además del error, se hará una validación cruzada junto con las matrices de confusión. La matriz consta de los casos que el algoritmo logró predecir con exactitud en un grupo de prueba, los falsos positivos y los falsos negativos de esta predicción, mientras que la validación cruzada de 10 veces consiste en cambiar diez veces los sets de prueba y entrenamiento, manteniendo sus proporciones, con tan de evitar que el modelo se sobre ajuste a un subset de datos en particular. Es de esperar que el porcentaje de precisión, opuesto al error, sea igual o mayor a un 60%, en donde ya se puede decir que un modelo de machine learning logra cierto grado de predicción.

REFERENCIAS Y BIBLIOGRAFÍA

Bianca Cung, J. H. (s.f.). *Regression and Machine Learning*. Obtenido de UCLA Department of Mathematics: http://www.math.ucla.edu/~wittman/10c.1.11s/Lectures/Raids/Regression.pdf

DATAPrix. (14 de Septiembre de 2007). *Metodología CRISP-DM para minería de datos*. Obtenido de DATA Prix Knowledge Is The Goal: http://www.dataprix.com/es/metodolog-crisp-dm-para-miner-datos

DELL. (2015). *Naive Bayes Classifier*. Obtenido de Technical Documentation: http://documents.software.dell.com/Statistics/Textbook/Naive-Bayes-Classifier

Departamento de Matemática Aplicada y Estadistica. (s.f.). *Introducción a los Árboles de Decisión.* Obtenido de Departamento de Matemática Aplicada y Estadistica: http://www.dmae.upct.es/~mcruiz/Telem06/Teoria/arbol_decision.pdf

Department of Statistics, University of California, Berkeley. (s.f.). *Random Forests*. Obtenido de Statistics at UC Berkeley: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Dipartimento di Elettronica ed informazione. (s.f.). *A Tutorial on Clustering Algorithms*. Obtenido de Dipartimento di Elettronica ed informazione: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

Gallardo, A. (2014). *UNA METODOLOGÍA DE EVALUACIÓN PARA UNA INTERVENCIÓN DE SALUD Y SU EFECTO EN EL RENDIMIENTO ACADÉMICO DE ESTUDIANTES UNIVERSITARIOS USANDO MACHINE LEARNING".* SANTIAGO.

Gerben W. Dekker, M. P. (2009). *Predicting Students Drop Out: A Case Study* . Eindhoven.

Kalyankar, M. M. (2010). Drop Out Feature of Student Data for Academic. *Global Journal of Computer Science and Technology*, 2-5.

Marquez-Vera, C. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *IEEE Education Society*, 7-14.

MECESUP. (s.f.). *Financiamiento Institucional*. Obtenido de MECESUP: http://www.mecesup.cl/index2.php?id_seccion=3586&id_portal=59&id_contenido=14892

MINEDUC. (30 de Septiembre de 2012). *Serie Evidencias: Deserción en la educación superior en Chile.* Obtenido de Ministerio de Educación de Chile - Mineduc: http://www.mineduc.cl/usuarios/bmineduc/doc/201209281737360.EVIDENCIASCEM9.pdf

MINERÍA DE DATOS Y ALMACENAMIENTO WEB. (s.f.). *MINERÍA DE DATOS Y ALMACENAMIENTO WEB*. Obtenido de MINERÍA DE DATOS Y ALMACENAMIENTO WEB: http://mineriadatosyalmacenamientoweb.net/

Pal, S. (2012). Mining Educational Data to Reduce Dropout. *I.J. Information Engineering and Electronic Business*, 1-7.

Redes de Neuronas Artificiales. (2012). *Redes de Neuronas Artificiales Aprendizaje*. Obtenido de Redes de Neuronas Artificiales: http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/index.html

Universidad Adolfo Ibáñez. (2013). *Plan de mejoramiento UAI 1303 convenio firmado.* Obtenido de Mecesup UAI: http://mecesup.uai.cl/download/documentos/convenios_y_bases/convenio_de_desempeno_mecesup.pdf

Universidad Adolfo Ibáñez. (2015). *Modelo Educativo UAI.* Obtenido de Acreditación Institucional UAI 2015: http://acreditacion.uai.cl/documentos/CentralesUAI/MODELO%20EDUCATIVO.pdf?VRPD=3

Universidad Adolfo Ibáñez. (2015). *Provición de carreras y programas.* Obtenido de Acreditación Institucional UAI 2015: http://acreditacion.uai.cl/documentos/Internos/PROVISI%C3%93N%20DE%20CARRERAS%20Y%20PROGRAMAS.pdf?VRPD=1

WebMining Consultores. (10 de Enero de 2011). *KDD: Proceso de Extracción de conocimiento*. Obtenido de Web Mining.cl :: Business Inteligence, Data Mining y Analitycs para Empresas: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/