PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA

EXPLORACIÓN DEL ÉXITO ACADÉMICO A TRAVÉS DE ALGORITMOS DE DATA MINING

KERIM ASSAD RUMIE GREZ

Tesis presentada a la Oficina de Investigación y Estudios de Posgrado en cumplimiento parcial de los requisitos para el Grado de Magister en Ciencias de la Ingeniería.

Tutor:

MIGUEL NUSSBAUM VOEHL

Santiago de Chile, Junio 2020

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA

EXPLORACIÓN DEL ÉXITO ACADÉMICO A TRAVÉS DE ALGORITMOS DE DATA MINING.

KERIM ASSAD RUMIE GREZ

Miembros del Comité:

MIGUEL NUSSBAUM VOEHL DocuSigned by: 8FAB758C7BD541B...

DENIS PARRA SANTANDER DocuSigned by: Denis Parra 2DE1B35BDA7B48B...

PABLO CHIUMINATTO DocuSigned by: Pablo Chiuminatto 821F48C5922C4D2...

ALONDRA CHAMORRO DocuSigned by: Alondra Chamorro 6F6EBA72F0C449B...

Tesis presentada a la Oficina de Investigación y Estudios de Posgrado en cumplimiento parcial de los requisitos para el Grado de Magister en Ciencias de la Ingeniería.

Santiago de Chile, Junio 2020

AGRADECIMIENTOS

Me gustaría agradecer a mi familia y amigos por su constante apoyo en todo el proceso. Ellos fueron mi principal motor para no sólo llevar a cabo este trabajo, si no que también para ser la persona que hoy soy. De manera especial, me gustaría agradecer a Ezio y Eduardo, ambos amigos. Gracias por siempre motivarme a seguir adelante y tomar las decisiones correctas cuando hubo incertidumbre. También quiero agradecer a todos aquellos que alguna vez les conté el tema de investigación de mi tesis por su respuesta motivadora en forma de felicitaciones por trabajar en un tema tan importante e interesante.

Quiero agradecer de sobre manera a mi profesor supervisor, Miguel Nussbaum. Gracias por darme la oportunidad de trabajar en una temática tan relevante y proveerme de un equipo de trabajo para asistir en el trabajo de investigación. Gracias por plantear un esquema de trabajo ideal para el éxito y las posibilidades para siempre crecer. También quiero agradecer a Catalina Cortázar por ayudarme y guiarme de manera permanente en el proceso desde incluso antes de que esto comenzara. Sin tu ayuda, este trabajo no sería posible.

También quiero agradecer a la Dirección de Pregrado de la Escuela de Ingeniería y a Ricardo Vílchez por su disposición cuando se requirió ayuda con las bases de datos y por proveer de estas. A mi equipo de trabajo, Francisco, Joaquín y Agustín, gracias por siempre apoyarme y acompañarme en la investigación.

Por último quiero agradecer a todos los que contribuyeron a convertirme en una persona interesada por el resto y sobre todo por temas relacionados con la educación. Agradezco al Centro de Alumnos de Ingeniería, a Kaizen, a mis profesores y funcionarios de la escuela.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABSTRACT

The increase in the number of engineering students accelerates the challenge of knowing how to care for, treat, and improve both the education and experience of the pupils. Given this increase, it is useful to analyze the success cases of the UC School of Engineering and with that, it is important to know what defines a successful student, and if these characteristics are congruent with the exit profile of the UC School of Engineering. From this it is possible to exploit the findings in order to propose improvements for the educational institution, such as, know what is the impact of the school in its students and take action on the final result of the study.

The objective of this study is to identify and understand which variables impact the academic success of the engineering students of the Pontifical Catholic University. This study used as a definition of academic success the student who completes his studies in the shortest time (Zhang, G., et al. 2004), also the student who obtains the best grades (Burton, L. et al. 2009) and finally the combination of the two definitions mentioned above was proposed.

The database of graduate students between 2007 and 2012 was analyzed, forming a sample of 2725 students. The information was first prepared and then analyzed using data mining techniques, such as decision trees and Random Forest. Demographic data, academic characteristics, participation in extra-programmatic activities, and the commitment with the school among the information of the students, disposed by the Undergraduate Directory of the School of Engineering UC, were considered. After this analysis, academic success prior to university entrance and academic performance seem to be more relevant than demographic characteristics and extra programmatic activities to achieve the

academic success defined above. This thesis had the support of FONDECYT/ CONICYT 1180024.

RESUMEN

El aumento de alumnos en ingeniería acelera el desafío de saber cómo cuidar, tratar y mejorar tanto la educación como la experiencia de los estudiantes. Dado este aumento, es útil analizar los casos de éxito de la Escuela y para eso, es importante saber qué es lo que define a un alumno exitoso, si estas características son congruentes con el perfil de egreso de la Escuela de Ingeniería UC y a partir de esto saber cómo explotar el conocimiento con el fin de proponer mejoras para la institución educacional, como por ejemplo, saber cuál es el impacto de la Escuela de Ingeniería UC en sus alumnos y tomar acción sobre el resultado final del estudio.

El objetivo de este estudio es identificar y comprender que variables son las que impactan en el éxito académico de los alumnos de ingeniería de la Pontificia Universidad Católica. En este estudio se usó como definición de éxito académico al estudiante que completa sus estudios en el menor tiempo (Zhang, G., et al. 2004), también al estudiante que obtiene las mejores calificaciones (Burton, L. et al. 2009) y por último se propuso la combinación de las dos definiciones mencionadas anteriormente.

Se analizó la base de datos conformada por los alumnos egresados entre 2007 y 2012, conformando una muestra de 2725 estudiantes. En primera instancia se preparó la información para luego analizarla con técnicas de minería de datos, tales como árboles de decisión y "*Random Forest*". Se consideraron datos demográficos, características académicas, participación en actividades extra – programáticas y el compromiso con la escuela entre la información de los alumnos, dispuesta por la Dirección de Pregrado de la escuela de Ingeniería UC. Luego de este análisis, parecen ser más relevantes el éxito académico previo al ingreso a la universidad y el desempeño académico universitario que las características demográficas y que las actividades extra programáticas realizadas por

los estudiantes para alcanzar el éxito académico definido anteriormente. Esta tesis tuvo el apoyo de FONDECYT / CONICYT 1180024.

1. INTRODUCCIÓN

Estudiar el flujo curricular, las características demográficas y sociales de los alumnos de la escuela de ingeniería es de suma importancia para entender cómo se desarrollan los alumnos, cual es su ruta crítica en cuanto a malla curricular y cómo la escuela los está afectando con su aporte desde el punto de vista educacional. En general las escuelas estudian la tasa de egreso de sus alumnos, las notas, duraciones promedio y sus distribuciones para obtener sus propias estadísticas (Naren et al., 2014).

Nuevas disciplinas como "*big data*" y "*data mining*" probablemente revolucionen la forma de tratar la información al entregar como resultado conclusiones más profundas sobre la causa de una consecuencia en un set de datos (Sagiroglu & Sinanc, 2013; Yu et al., 2016). Por ejemplo, cuando se desea estudiar un problema de clasificación, es muy probable que se intenten utilizar regresiones lineales o métodos estadísticos para lograr el objetivo, pero las técnicas de "*data mining*" proveen de metodologías automatizadas y no lineales capaces de obtener mejores resultado (Varian, 2014). Dentro de estas metodologías existen las redes neuronales, árboles de decisiones, "*random forest"*, "*support vector machines"* entre otros. Estos métodos pueden mejorar la eficiencia de los estudios cuando existen muchas variables, y además estos intentar buscar relaciones entre la información del estudio sin entregar una hipótesis a probar.

En la práctica, el "*data mining*" se ha utilizado en diversos campos, como por ejemplo, salud (Wang et al., 2017; Raghupathi, 2016), educación (Angeli et al., 2017; Romero & Ventura, 2013) y los negocios (Shmueli et al., 2017; Provost & Fawecett, 2013), entre otros. Su aplicación es variada, como por ejemplo: procesar imágenes, audios, videos (Gamal et al., 2017; Panda et al., 2017), reconocimiento de textos (Niekler et al., 2017; Cambria et al., 2013), y predicción de eventos lo que ha demostrado las funcionalidades de esta metodología.

Específicamente el rubro de la educación es interesante para la aplicación de técnicas de "*data mining*" dado que por su naturaleza, y su duración, las bases de datos cuentan con información de diversas fuentes, con grandes cantidades de variables, lo que nos podría permitir encontrar información oculta y tendencias en una gran base de datos (Naren et al., 2014).

Las preguntas de investigación que conducen este trabajo son las siguientes:

1. ¿Qué variables inciden mayormente en el éxito académico de los estudiantes de la Escuela de Ingeniería UC?
2. ¿Qué se puede concluir con respecto a estas variables?

2. PROPUESTA DE TRABAJO

En esta sección se describe brevemente la propuesta de trabajo. La que principalmente es obtener información sobre las variables más importantes en torno al éxito académico de los alumnos de Ingeniería UC a través de algoritmos de "*data mining*". Para lograr esto, se propone un proceso de 4 etapas que consisten en:

1. Preparar la información.
2. Aplicar algoritmos de "*data mining*".
3. Analizar y volver a realizar los pasos anteriores.
4. Generar conclusiones a través de los resultados del proceso antes mencionado.

2.1 Objetivo: Descubrir las variables que más inciden en el éxito académico

La propuesta de trabajo diseñada involucra un proceso iterativo de técnicas de pre procesamiento de bases de datos con algoritmos de "*data mining*" con el fin de obtener

conclusiones sobre las variables existentes de la base de datos a estudiar. Para llevar a cabo esto se requiere:

1. Una base de datos representativa de los estudiantes de Ingeniería UC, que contenga el contexto académico, socio – económico y demográfico.
2. Estar familiarizado con la institución sobre la que se obtienen los datos para facilitar la preparación de la información a estudiar, por ejemplo, para entender la relevancia o secuencia de ciertos cursos.

En este caso, se trabajará con una base de datos del rubro educacional para obtener información sobre las variables más incidentes en el éxito académico de los estudiantes.

2.2 Preparación de la información: Pre procesamiento

Tener una base de datos apta para ser trabajada es un paso muy importante al momento de hacer un estudio, ya que le permitirá a los algoritmos de "*data mining*" trabajar de la manera correcta. La información debe ser provista en la cantidad, estructura y formato preciso para que el algoritmo sea capaz de leerlo. Desafortunadamente las bases de datos del mundo cotidiano están altamente influenciadas por factores negativos como presencia de ruido e información inconsistente, lo que determina una baja calidad en la base datos y por ende un rendimiento de baja calidad en el algoritmo de "*data mining*" (García et al., 2015). Dada esta situación, el fin último del pre procesamiento de la información es mejorar una base de datos para poder ser procesada por algoritmos mediante diversas técnicas.

En cuanto al pre procesamiento de la información, este es un paso obligatorio al trabajar con bases de datos, ya que si la data no se prepara es probable que el algoritmo no funcione o entregue información sin sentido (García et al., 2015). La preparación de la información se lleva a cabo con las siguientes técnicas (resumen a continuación y Figura 2-1).

1. Limpieza de datos: esta técnica se centra en filtrar la información incorrecta y reducir los detalles innecesarios de la información (García et al., 2015).
2. Transformación de datos: esta técnica es la encargada de consolidar la información para que el proceso de "*data mining*" pueda ser aplicado o sea más eficiente (García et al., 2015).
3. Integración de datos: esta etapa es la encargada de unificar la información de distintas ubicaciones (García et al., 2015).
4. Normalización de datos: proceso encargado de dejar todos los atributos expresados en la misma unidad de medida con una escala o rango común (García et al., 2015).
5. Relleno de datos: es una subtécnica de la limpieza de datos cuyo propósito es rellenar variables que contienen información faltante que es posible estimar (García et al., 2015).
6. Identificación del ruido: este paso también es parte de la limpieza de datos y su objetivo es detectar errores aleatorios y corregirlos (García et al., 2015).

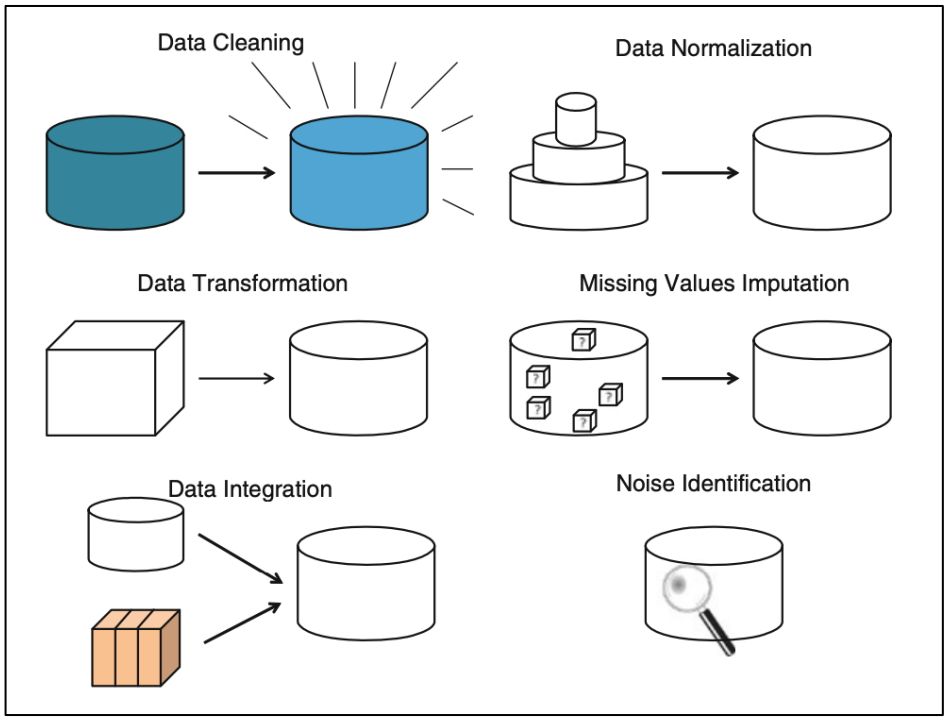


Figura 2-1:Resumen de la preparación de la Data. (García et al., 2015).

Una vez realizado los paso de la preparación de la información, otro paso importante en el pre procesamiento es la reducción de la información. El paso de reducción de información agrupa las técnicas utilizadas para reducir la cantidad de información en la base de datos. Este paso puede ser considerado opcional, dado que si la información original está bien estructurada, no es crucial llevarlo a cabo, pero si puede aumentar la eficiencia del algoritmo (García et al., 2015).

La reducción de la información se utiliza principalmente para reducir la dimensionalidad de la base de datos, la eliminación de información redundante, simplificación de los atributos y por último para rellenar vacíos en la información. Estas técnicas serán explicadas brevemente a continuación y en la Figura 2-2.

1. Reducción de dimensionalidad: el objetivo es aumentar la velocidad del algoritmo a través de reducir la cantidad de variables sin perder la representatividad de la base de datos original (García et al., 2015).
2. Remoción de información redundante: consiste en seleccionar el subgrupo de sujetos a estudiar en la base de datos que sean representativos o que sirvan para cumplir el objetivo original del algoritmo (García et al., 2015).
3. Discretización: esta técnica transforma la información cuantitativa en cualitativa al agrupar los atributos en intervalos discretos con el fin de simplificar el proceso del algoritmo de "*data mining*" (García et al., 2015).

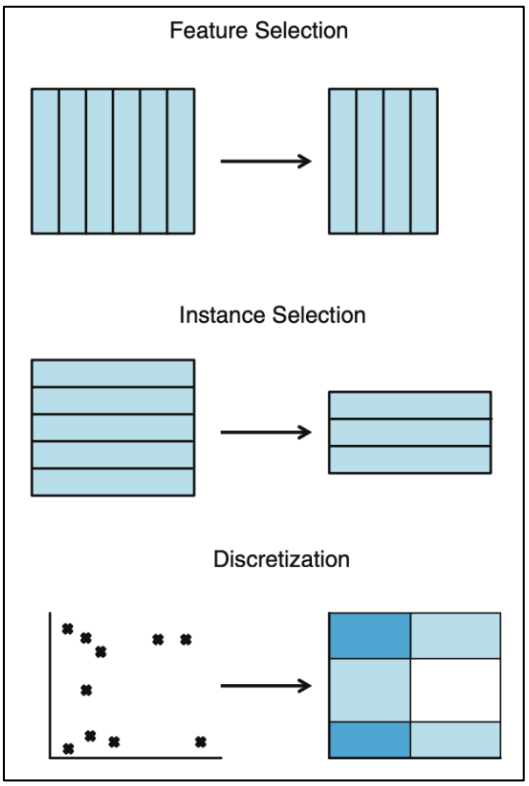


Figura 2-2:Resumen de la reducción de la Data. (García et al., 2015).

2.3 "Data mining": Elegir el algoritmo adecuado

La etapa de elegir el algoritmo adecuado para el estudio es crítica y de mucha relevancia, ya que por un lado, la base de datos se debe adecuar al algoritmo elegido y además éste debe cumplir con las necesidades del objetivo del mismo estudio.

Para comenzar, es importante introducir el concepto de "*data mining*". Hoy en día, ocurren una infinidad de procesos de manera simultánea, los que contienen mucha información, más de que las personas somos capaces de procesar, por eso surge la necesidad de automatizar el proceso de análisis de la información, y eso es lo que el "*data mining*" provee. El "*data mining*" entrega las técnicas para detectar patrones automáticamente, y predecir el futuro (Murphy, 2012).

Según Silva & Fonseca (2017), los algoritmos de "*data mining*" que se han utilizado con información educacional son las redes neurales, "*support vector machines"*, métodos basados en árboles de decisión y clasificadores bayesianos. La superioridad del un método sobre el otro va a depender del tipo de información que se tiene y el objetivo del estudio (Fernández-Delgado et al., 2014; Wolpert, 1996). Por ejemplo, para el reconocimiento de imágenes las redes neuronales son superiores a los árboles de decisión dado su forma de operar.

Algunos de los parámetros para comparar los algoritmos son:

- Calidad de los resultados. Es la capacidad del algoritmo de predecir o clasificar resultados correctos (Han et al, 2011).
- Interpretación de resultados. Evalúa la facilidad de interpretar los resultados del algoritmo. Por ejemplo, las redes neurales son difíciles de interpretar ya que funcionan como "caja negra", mientras que los árboles de decisión entregan un resultado gráfico más fácil de entender. (Han et al, 2011).

- Velocidad de procesamiento. Principalmente evalúa los recursos necesarios para funcionar con cierta velocidad y cómo estos aumentan al escalar la base de datos. (Han et al, 2011).
- Robustez. Es la habilidad del algoritmo para trabajar con ruido o información vacía. (Han et al, 2011).

Como se mencionó anteriormente, es muy importante tener claridad del objetivo del estudio. El objetivo puede ser predecir, clasificar o incluso agrupar muestras con características o buscar similitudes de la base de datos con un criterio dado.

La secuencia utilizada para elegir el algoritmo "*data mining*" se obtuvo al las siguientes preguntas:

- ¿Cuál es mi objetivo final?
- ¿De qué tamaño y tipo es mi base de datos?
- ¿Cómo voy a interpretar los resultados?
- ¿Qué algoritmo de "*data mining*" es el más adecuado?

Esta secuencia se define de manera más concreta en la sección siguiente 2.4 El proceso iterativo y se desarrolla con mayor detalle cada una de sus partes en la sección 4 Metodología

2.4 El proceso iterativo

La propuesta de trabajo consiste de 3 etapas principalmente:

1. Definir un objetivo
2. Procesar la información
3. Aplicar algoritmo de "*data mining*" y analizar los resultados.

El orden del flujo es el siguiente:

1. Primero se define un objetivo de estudio
2. Se procesa la base de datos acorde al objetivo definido
3. Se aplica el algoritmo de "*data mining*" y se analizan los resultados
4. se puede volver a la etapa de procesamiento en base a los resultados de la etapa 3 y así permitir que el proceso sea iterativo para lograr mejores resultados o comprobar que el resultado sea el convergente o adecuado.

La investigación termina cuando los resultados de la etapa 3 sean adecuados para obtener las conclusiones deseadas. A continuación la figura 2-1 resume el proceso.

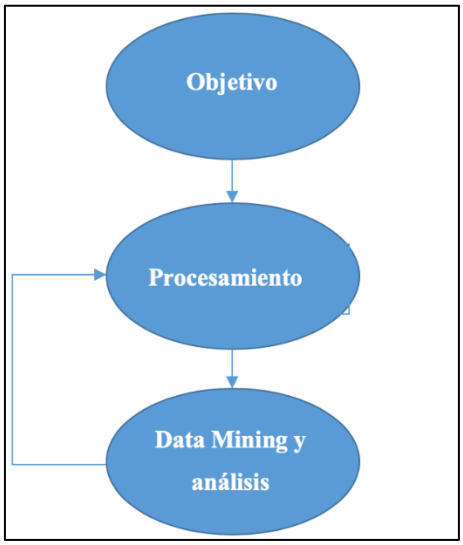


Figura 2-3:Diagrama del modelo propuesto.

3. REVISIÓN BIBLIOGRÁFICA DEL CASO DE ESTUDIO

En esta sección se presenta la revisión bibliográfica sobre el éxito académico, el proceso acumulativo del éxito académico y algunos trabajos de "*data mining*" en el campo del éxito académico.

3.1 Éxito académico

El éxito académico es una temática que se ha estudiado desde muchas perspectivas y con distintos fines. En algunos casos el objetivo es comprender cuales son los factores claves que influencian en el éxito académico para mejorar el sistema educacional, este es el caso de estudio de Burton et. Al (2009). En otro casos se busca levantar información especifica, como por ejemplo es el caso de Kamphorst et. al (2015) quien buscaba entender si el género es un factor que afecta en el éxito académico. Por otro lado también se ha estudiado el efecto de ciertas habilidades especificas en torno al éxito académico a través de pruebas estandarizadas como es el caso de Chowhan (2013) y por último se ha investigado la predicción del éxito académico en torno a variables académicas y pruebas de diagnostico como es el caso de Van den Broecka et. Al (2017).

En cuanto a las técnicas utilizadas para llevar a cabo los estudios, principalmente se han utilizado múltiples modelos de regresiones logísticas para descubrir las relaciones estadísticas entre las variables como es el caso de Thorndyke et. Al (2004) y también una combinación de descripciones estadísticas junto con modelos multi variables como es el caso de McCool et. Al (2015).

3.2 El éxito académico como un proceso acumulativo

Diversos estudios han demostrado que las experiencias educacionales previas son relevantes para el éxito académico tanto del primer como último año de estudios (Burton et al., 2009; Burton et al., 2010; Adelson et al., 2016; Thorndyke et al., 2004).

El aprendizaje es un proceso acumulativo donde los estudiantes aprenden nuevas habilidades mientras que siguen desarrollando las ya existentes (Duncan et al., 2007). Esto indica McCool et al (2015) dado que encuentra que los alumnos con mayor experiencia estadísticamente rinden mejor que sus contrapartes más jóvenes. También Thorndyke et

al., (2004) concluye que la graduación de los alumnos de ingeniería se correlaciona de manera positiva con el promedio de calificaciones del colegio y los resultados de las pruebas de matemática para ingresar a la universidad. Por lo tanto es importante tomar en cuenta estas variables al momento de hacer el estudio.

3.3 El trabajo del "data mining" en el éxito académico

El uso de "*data mining*" en el campo educacional del éxito académico se enfoca en clasificar, agrupar, predecir y encontrar reglas de asociación en torno a la información (Baradwaj, et at. 2011). En cuanto a los estudios realizados en estudiantes, Pandey & Pal et al. (2011) estudió el comportamiento de 600 alumnos a través de clasificadores bayesianos sus contexto previo a la universidad para saber si los nuevos estudiantes serán exitosos o no. Khan (2005) condujo un estudio sobre 400 estudiantes para establecer los valores de las variables demográficas, de personalidad y cognición que definan el éxito en la escuela secundaria a través de técnicas de agrupación. También Ayesha et al. (2010) describe el uso del algoritmo de *"k-means"* para predecir las actividades de aprendizaje de los alumnos. Por último, Al-Radaideh et al (2006) aplicó árboles de decisión para predecir la nota final de los alumnos de la universidad de Yarmouk. Utilizó 3 métodos de clasificación diferentes y concluyó que los árboles de decisión fue el método que mejores resultados obtuvo.

En general, este tipo de estudios es un campo emergente que se llama "Data Mining Educacional" cuyo objetivo es poder predecir el desempeño de los estudiantes el día de mañana para mejorar el sistema educacional, detectar anomalías y en general descubrir las relaciones entre las variables y su efecto en el estudiante.

3.4 El resultado de la revisión bibliográfica

A continuación se presentan las consecuencias que generó el estudio bibliográfico realizado en las secciones previas.

En primer lugar, se debe destacar que el éxito académico es una temática estudiada desde distintos puntos de vista, lo que ha indicado que el término no es fijo, es decir, que las variables que lo afectan aún no están completamente definidas. En segundo lugar, se ha ha concluido que el éxito académico es un proceso cumulativo, por lo tanto es relevante incorporar variables previas al ingreso de la universidad. Por último, previamente han utilidazo técnicas de minería de datos, concluyendo que lo árboles de decisión son una buena herramienta. Esto genera que me incline a utilizar ese tipo de técnicas dada sus características adecuadas para este caso de estudio, las que se explican con mayor detalle en la sección 4.3 de este documento.

4. METODOLOGÍA

En esta sección, se explica la aplicación de la propuesta para nuestro caso de estudio, definir el éxito académico de los alumnos de Ingeniería UC.

4.1 Objetivo

Cada estudiante es único dado que tiene notas específicas en cada una de sus asignaturas, puede o no participar en actividades extra programáticas y tiene un contexto socio económico específico. La Dirección de Pregrado de la Escuela de Ingeniería UC es el organismo que recopila toda la información de los alumnos de la escuela, creando una base de datos que incluye la información académica, socio económica, actividades extra programáticas de los alumnos.

Se estudió la relación de los alumnos con el éxito académico con el fin de explicar que es lo más relevante para el caso de los alumnos estudiados. En este caso, se utilizaron los alumnos pertenecientes a la escuela entre los años 2007 y 2012. Como variable dependiente del éxito académico se utilizó el promedio ponderado global, el tiempo hasta el egreso, y por último una combinación de estas 2 últimas variables. En cuanto a las variables independientes, se tiene el contexto socio económico y demográfico, los resultados académicos previos, las notas finales de cada asignatura y variables extra programáticas demostrando su participación a lo largo de su vida universitaria. En total cada estudiante fue asociado a 1436 variables independientes, las que incluyen todos sus cursos.

En base a esta información se utilizaron técnicas para pre procesar la información, luego "*data mining*" tales como árboles de decisión y *"random forest"* para reducir el número de variables y lograr encontrar las variables más relevantes, para finalmente obtener una conclusión sobre qué es lo que define a un alumno exitoso en la Escuela de Ingeniería UC.

4.2 Preparación de la base de datos para el "data mining"

La tabla 4-1 a continuación presenta las etapas seguidas para preparar la base de datos. Luego se describe que se realizó en cada etapa.

Tabla 4-1: Etapas para la preparación de la base de datos.

Etapa	Objetivo	Acciones realizadas
1. Ordenar y unificar las bases de datos.	Trabajar de manera más ordenada y rápida.	Describir la base de datos. Describir las variables independientes y dependientes. Crear la base de datos unificada.
2. Definir la muestra y el pre procesamiento de la información.	Eliminar el ruido de la muestra y crear variables nuevas agrupando variables.	Describir como se definió la muestra. Explicar criterios para eliminar ruidos y variables eliminadas. Comparar la base de datos nueva con la inicial para verificar.
3. Adecuar el lenguaje de la base de datos.	Crear una base de datos que facilite al algoritmo lograr su objetivo.	Traducir el valor de las variables a un lenguaje adecuado para los algoritmos utilizados.

4.2.1 Orden y unificación de la base de datos

En esta etapa se describe la base de datos utilizada, las variables que contiene y el procedimiento de orden y unificación de la base de datos para poder visualizarla como una base de datos factible a utilizar.

En cuanto a las bases de datos brindadas por la Dirección de Pregrado, estas se componen de 1 archivo por año, con 2 hojas cada archivo, utilizando desde el año 2007 al 2012. La primera hoja contiene la información académica de cada alumno, es decir, sus cursos y respectivas notas. Estas se representan en una lista de filas. La segunda hoja contiene el resto de las variables de los alumnos, como por ejemplo, colegio de procedencia, región de procedencia, etc. En el apéndice A se entregan mayores detalles de la base de datos original.

El objetivo era tener un solo archivo que tuviese toda la información de cada alumno de manera que fuese fácil de leer. Para esto se decidió que los alumnos serían representados por las filas y las variables con las columnas. Para lograr este objetivo, principalmente se utilizaron herramientas de Excel tales como tablas dinámicas y conversión de letras a números (para aquellos cursos que cuya calificación es A o R de aprobado o reprobado). Esto se realizó para cada una de los años y así todos los años en el formato deseado.

En cuanto a las variables contenidas en las bases de datos, estas se pueden agrupar en los siguientes grupos para facilitar su comprensión. 1) Variables académicas, 2) Variables demográficas, 3) Variables que representan las actividades extra programáticas y 4) Variables representan el compromiso con la escuela y comunidad. En la tabla 4-2 se representa con mayor detalle la composición de estos grupos con sus respectivas variables.

Tabla 4-2: Clasificación de variables.

Variables Demográficas	Variables Extra - programáticas	Variables Académicas	Variables de compromiso
- Género - Forma de ingreso - Vía de ingreso - Región - Tipo de colegio - Nombre del colegio	- Deportista UC - Ayudante - Participación en trabajos sociales - Tutor - Embajador	- Notas de cada curso - Porcentaje de aprobación de cursos - Notas PSU - Causales de eliminación - Fechas de licenciatura, egreso y titulación - Especialidad - Procedencia	- Donaciones - Trabajador UC - Asistencia a eventos de la Escuela (Encuentro interno, Cena CAi, Back to school) - Padre o hijo en la escuela

Dado que el éxito académico se definió según la literatura como una función del tiempo hasta la graduación (Zhang, G., et al. 2004) y como un resultado de las notas de los alumnos (Burton, L. et al. 2009), la variable independiente sería la que represente estas definiciones en cada caso, y las dependientes serán todas las restantes.

Ya ordenada la base de datos es importante unificar todos los años en un solo archivo para evitar inconsistencias en los futuros pasos a tomar y así siempre realizar cambios sobre la base completa y no de manera separada. Para lograr esto, se realizó un proceso de 2 etapas. La primera consistió en nombrar a todas las variables de las respectivas bases de datos de la misma manera y la segunda correr una función en Python para unificar de manera correcta los archivos. En Python se corrió un código para ordenar, cuya misión es juntar los distintos archivos .csv con la particularidad de crear columnas con nuevas variables cuando no existe una coincidencia entre las variables de los distintos archivos. Esta función es muy útil, dado que no todas las variables en los distintos archivos son las mismas, principalmente porque los alumnos tomaron distintos ramos dado que son libres de elegir sus optativos y algunas siglas de cursos cambiaron a lo largo de los años.

4.2.2 Definición de la muestra y el pre procesamiento

En esta etapa se describe la metodología para seleccionar la muestra sobre la que se hará el estudio finalmente con sus respectivos pasos y justificaciones.

En cuanto a la base de datos utilizada, para poder evaluar el éxito académico de manera uniforme a través de los alumnos de la Escuela de Ingeniería, como condición era necesario que los alumnos al menos estuviesen egresados para evaluar su desempeño completo en la universidad. Por otro lado, la Dirección de Pregrado nos facilitó el registro de todos los alumnos que ingresaron a la Escuela entre el 2007 y el 2012.

En la tabla 4-3 se puede apreciar el numero de estudiantes por año versus el número de estudiantes egresados por año.

Tabla 4-3: Número de alumnos por año versus alumnos egresados por año.

	2007	2008	2009	2010	2011	2012
Alumnos inscritos	547	544	611	625	677	726
Alumnos egresados	477	457	504	505	473	305

Dado que se quiere estudiar el éxito académico, la muestra de alumnos no egresados, tales como los renunciados, expulsados o los que decidieron cambiarse de carrera no se pueden considerar en el estudio debido a que no manejamos la información sobre la razón para no completar los estudios en la escuela, por lo tanto para nuestro caso de estudio, genera ruido en la muestra al no saber qué explica estas situaciones. Esto generó una muestra de 2725 alumnos egresados entre el 2007 y el 2012.

Luego, para facilitar el trabajo de compresión en el proceso de "*data mining*", se pre-procesó la base de datos. Principalmente se agruparon ciertas variables con el fin de ver si estas en conjunto son de importancia, y en caso de serlo se desagrupan para ver su importancia individual. Otra razón para agrupar variables es que son variables equivalentes, por ejemplo cursos que han cambiado de siglas (por lo tanto son variables distintas) pero son el mismo curso. En la tabla 4-4 se puede apreciar el resumen de las variables agrupadas.

Tabla 4-4: Resumen de variables agrupadas.

Nueva variable	Variables agrupadas
Proporción de aprobación.	Total créditos inscritos + total créditos. aprobados + convalidados.
Si estuvo o no en causal de eliminación.	Número de causales de eliminación.
Cantidad de días hasta el egreso.	Fecha de egreso + Fecha de ingreso.
Forma de ingreso	Tipo de ingreso + vía de ingreso.
Curso probabilidades y estadística	Curso probabilidades + estadística.

Tabla 4-5: Resumen de cursos equivalentes.

Curso antiguo	Curso nuevo
ING1003	ING1004
MAT1202	MAT1203
MAT1503	MAT1610
MAT1512	MAT1620
MAT1523	MAT1630
MAT1532	MAT1640
FIS1532	FIS1533
IIC1102	IIC1113
ICS1102	ICS1113
ICS1502	ICS1513
FIS1512	FIS1513
FIS1522	FIS1523
MAT1102	MAT1600
QIM100	QIM100I + QIM100A

Además de agrupar variables, el pre procesamiento de la base de datos también consistió en eliminar variables que son redundantes al entregar la misma información, o que se puede obtener con un conjunto de otras variables. También fueron eliminadas variables (luego del agrupamiento) ya que generaban ruido en la muestra por los siguientes motivos: 1) existía muy poca información sobre ellas, 2) No se definió una forma clara de representarlas o 3) no aportaba información útil para el estudio. A continuación en la tabla 4-6 se resumen las variables eliminadas con su debida justificación.

Tabla 4-6: Resumen de variables eliminadas.

Variable eliminada	Motivo
Curriculum	No aporta información
Código programa	No aporta información
Estado	Irrelevante para el estudio
Código preferencia especialidad	Variable redundante
Preferencia especialidad	Variable redundante
Código preferencia major	Sin información
Preferencia major	Sin información
Código major Banner	Sin información
Major Banner	Sin información
Código minor Banner	Sin información
Minor Banner	Sin información
Código Minor Siding	Sin información
Minor Siding	Sin información
Año ingreso	Irrelevante para el estudio
Licenciado	Irrelevante para el estudio
Fecha licenciatura	Irrelevante para el estudio
Titulado	Irrelevante para el estudio
Fecha titulación	Irrelevante para el estudio
Instrumento de titulación	Irrelevante para el estudio
Examen de licenciatura aprobado	Irrelevante para el estudio
Fecha examen de licenciatura	Irrelevante para el estudio
Examen de titulo aprobado	Irrelevante para el estudio
Fecha examen de titulo	Irrelevante para el estudio
Ciclo 2 (preferencia)	Sin información

Salida al mercado	Sin información
Continuidad de estudios	Sin información
Detalle pregrado	Sin información
Detalle postgrado	Sin información
Puntajes PAA	Sin información
PSU historia	Mucha información faltante
Fecha egreso colegio	No aporta información

Para continuar con el proceso de procesamiento de la base de datos se debe decidir qué hacer con las variables que tengan información faltante. En base a este desafío se plantean las siguientes acciones a tomar:

1. Intentar rellenar información faltante de las variables en base a otras variables en caso de ser posible. Por ejemplo, si falta información en la variable que indica si el alumno está egresado, revisar la variable si el alumno está titulado.
2. Para las variables que son las calificaciones de cada curso, si la información faltante es menor al 15% se procede a rellenar con el promedio de la variable.
3. En caso de que la información faltante sea superior al 15% se debe evaluar si agrupar con otra variable o eliminar.

Por último para finalizar el proceso de orden y definición de la muestra, es necesario normalizar las variables relacionadas con los cursos para que sean comparables entre si. Esta decisión se toma debido a que en la Escuela de Ingeniería podrían existir cursos lo que tienen promedios en general menores al resto por su grado de dificultad. Para llevar a cabo el proceso de normalización, se asumió que las notas distribuyen de manera normal, por lo que se normalizo en torno al promedio y la desviación estándar de cada variable. Para esto se utilizó la siguiente formula:

$$X - \mu / \sigma$$

4.2.3 Adecuar el lenguaje de la base de datos para el algoritmo de "data mining".

Una vez ya ordenada la base de datos y realizado el primer pre procesamiento, es necesario definir un lenguaje común para las variables para que de esta forma puedan ser comparables entre si al momento de correr el algoritmo de "*data mining*". Para lograr esto, lo primero fue analizar los tipos de variables existentes en la muestra. Dentro de esta existen variables categóricas (como el sexo, la región de procedencia, etc.), variables continuas (como las notas de cada curso) y variables discretas (como la cantidad de créditos de los alumnos). Por otro lado, como vimos anteriormente, el lenguaje que requieren los árboles de decisión y *"Random Forest"* son el mismo, y básicamente pueden leer casi todos los lenguajes, pero requieren que cada variable esté en el mismo lenguaje. Para facilitar la comprensión y la comparación de las variables entre si, se decidió transformar a lenguaje binario la mayor cantidad de variables posibles. Para cumplir este objetivo, en algunos casos fue necesario agrupar la información de cada variable para facilitar el proceso de evaluación de los resultados. Por ejemplo, para la variable que indica de que región es el alumno, se cambió a si pertenece o no a la región metropolitana.

A continuación en la tabla 4-7 se presenta el resumen de las variables transformadas al lenguaje binario con su respectiva observación si se requiere.

Tabla 4-7: Resumen de variables en lenguaje binario.

Variable en lenguaje binario	Observación
Colegio de región	1 es RM, 0 es región
Compromiso	1 es comprometido, 0 no.
Extra programático	1 es participar en extra programáticas, 0 no.
Genero	1 es mujer, 0 hombre-
Causal de eliminación	1 es tener al menos 1 causal, 0 ninguna.

Colegio particular	1 es que el colegio sea particular, 0 no.
Vía de ingreso	1 es ingreso normal, 0 otro.

4.3 Aplicación de "data mining" para reducir variables y obtener resultados

Tal como se explicó previamente, es importante recordar que todo este proceso es iterativo, por lo tanto, el objetivo de los algoritmos de "*data mining*" es entregar información relevante sobre las variables con el fin de volver a la etapa de procesamiento de la base de datos para avanzar con las variables más importantes y poder concluir a partir de ellas.

En cuanto a la estructura de nuestros datos, estos siguen una estructura de matriz dado que cada estudiante es una fila y cada una de sus características (variables) son una columna. Por otro lado, como existen grandes cantidades de cursos, la base de datos es altamente dimensional, con 1436 variables y de distintos tipos: binaria, nominal, ordinales y numéricas entre otras. Es probable que varias de estas variables sean irrelevantes para predecir el éxito académico.

Para el caso de estudio los algoritmos utilizados fueron *"Random Forest"* y Árboles de decisión por los siguientes motivos:

1. La base contiene grandes cantidades de variables, de las cuales algunas no son relevantes.

   Por la forma en cómo funciona *"Random Forest"* particularmente, este algoritmo se comporta bien con grandes cantidades de variables, incluso con algunas que no signifiquen un aporte. En este punto es superior a otros algoritmos como *"Support Vector Machines"*. (Han et al. 2011)
2. Fácil de implementar e interpretar

Tanto los árboles de decisión como *"Random Forest"* son fáciles de implementar dado que el programa no requiere de muchos recursos (memoria). Además los resultados son fáciles de interpretar ya que se puede configurar para que entregue un ranking de importancia de las variables. En el caso de los árboles de decisión estos al representarse de manera gráfica como un árbol también es fácil de interpretar la importancia de las variables en los resultados. (Han et al. 2011)

3. Flexibilidad de trabajo en cuanto a las variables

   Ambos algoritmos son capaces de trabajar con distintos tipos de variables, lo que facilita el trabajo de pre procesamiento de la base de datos. (Han et al. 2011)

4. Existencia de ruido en la información

   A pesar de que se realiza un trabajo importante de limpieza de la base de datos, no tenemos certeza del nivel de ruido residual en la base de datos, y como este algoritmo trabaja bien con esta característica, es ideal para nuestro caso. (Han et al. 2011)

Los árboles de decisión y *"Random Forest"* están muy relacionados, dado que *"Random Forest"* trabaja sobre los árboles de decisión, por lo tanto, se entregará una explicación de ambos algoritmos a continuación.

4.3.1 Árboles de decisión

El árbol de decisión es un modelo de predicción y clasificación. Estos se representan como gráficos que parecen árboles ya que contiene nodos, y estos a su vez, tienen ramas. En cada nodo se ubica una variable la que separa la muestra según una serie de criterios. Las variables de cada nodo se pueden comparar con otras variables o con constantes para generar las ramificaciones del árbol. Para ilustrar mejor los árboles de decisión en la figura

4-1 se puede apreciar un ejemplo elaborado de manera propia. Para el caso de la figura, podemos apreciar las distintas variables que se sitúan en cada nodo. Mientras más arriba se ubique la variable, es decir, en un nivel más elevado, más importante será para el caso de estudio. En este ejemplo se están muestran las variables más relevantes para el éxito académico definido en función de las mejores notas al egreso. A medida que disminuimos el nivel del árbol, la importancia relativa de la variable en el nuevo nodo disminuye también. El valor que acompaña a cada variable es la constante utilizada para dividir la muestra. Por ejemplo, en el primer nodo, con la variables Aprobados/Inscritos, 0.945 es el valor que divide la muestra y crea la ramificación.

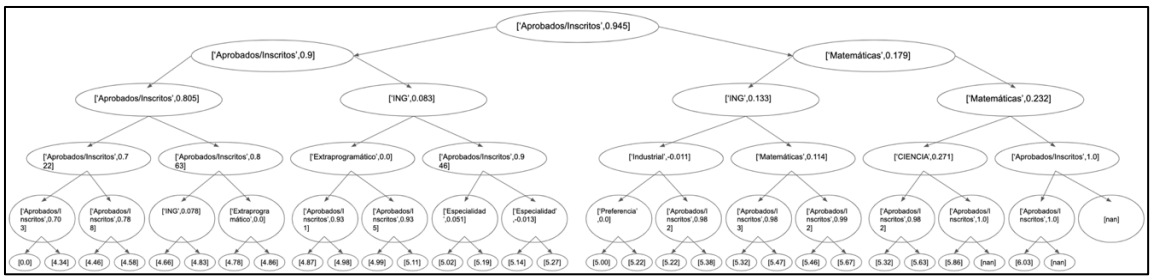


Figura 4-1:Ejemplo de árbol de decisión. Elaboración propia.

Los árboles de decisión trabajan en un proceso recursivo ya que al funcionar el algoritmo, éste selecciona la variable que mejor separa a la muestra, y divide la muestra en 2 nuevos nodos, los que repiten el proceso para continuar el proceso de separación de la muestra. Cuando un nodo tiene una muestra de un solo tipo, significa que ya no se puede separar más y esa rama ha terminado. La cantidad de ramificaciones hacia abajo denota la cantidad de niveles del árbol, esta cantidad se puede configurar a gusto para disminuir el tamaño del árbol y el sobre ajuste. Este proceso de configuración también se le llama poda del árbol. Esta técnica se utiliza para evitar la generación de resultados pobres en términos de clasificación de muestras. En cuanto a la forma en que el algoritmo divide a la muestra, este trabaja modelando la entropía de la muestra, donde esta es cero si la muestra es de la misma clase y por otro lado esta es máxima si la cantidad de clases son iguales (Witten & Frank, 2005).

4.3.2 "Random Forest"

"Random Forest" es un algoritmo que trabaja sobre los árboles de decisión. Estos serían la unidad básica del algoritmo pero con la particularidad de que *"Random Forest"* realiza 2 técnicas tanto para reducir la varianza y el sobre ajuste como para aumentar la precisión del algoritmo (Breiman, 2001).

La primera técnica consiste en realizar paquetes de muestras aleatorias utilizando la base de datos original. Luego se realiza el algoritmo del árbol de decisión a cada una de estas muestras. La cantidad de muestras que se realizan es un parámetro que se le debe informar al programa. Para nuestro caso de estudio, sería realizar paquetes de estudiantes aleatorios sobre la base de datos completa.

La segunda técnica consiste en realizar paquetes aleatorios sobre las características de la base de datos para así elegir cuál es la que mejor separa a la muestra. Esta característica (variable) es elegida sobre el subgrupo generado y no sobre la base de datos completa. Esta técnica puede aumentar la precisión del algoritmo reduciendo la correlación entre los nodos (Breiman, 2001). La cantidad de muestras que se realizan es un parámetro que se le debe informar al programa. Para nuestro caso de estudio, sería realizar muestras con todos los estudiantes con diferentes grupos de variables obtenidas de la base de datos completa.

Por lo tanto *"Random Forest"* se puede resumir en realizar 2 procesos de empaquetamiento de muestras aleatorios, lo que genera múltiples árboles de decisión y obteniendo como resultado una clasificación de variables en un ranking de importancia más precisa que al utilizar un solo árbol de decisión. A continuación se muestra la figura 4-1 que demuestra como *"Random Forest"* funciona.

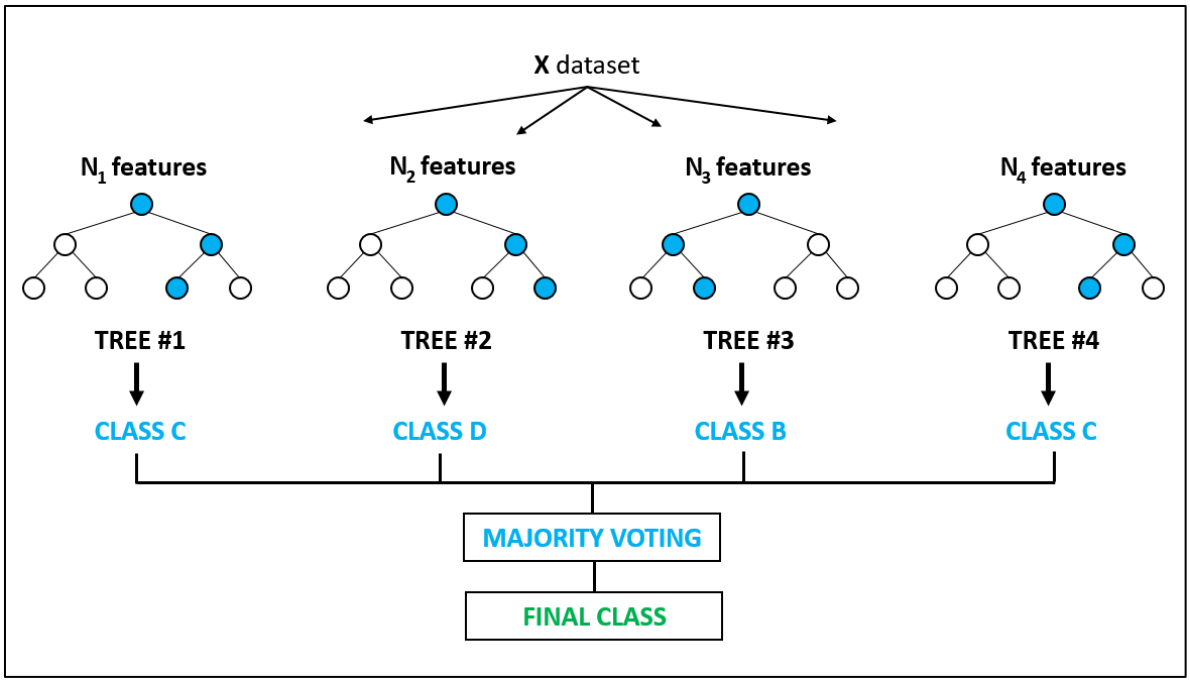


Figura 4-2:Ejemplo funcionamiento "Random Forest" (Abilash, R. 2018).

4.4 El proceso iterativo

En esta sección, se describe el proceso seguido del resultado entregado por el algoritmo de *"Random Forest"*. Principalmente se utilizaron las mismas técnicas utilizadas en la etapa de pre procesamiento como agrupación, eliminación y relleno de variables. Adicionalmente se seleccionaron sub muestras de la base de datos de manera no aleatoria sobre las cuales se iteró para comparar resultados. Por otro lado, muy importante, el proceso además de ser iterativo por los cambios en las variables es iterativo dado que se utilizaron distintas definiciones de éxito académico.

Primero se definirá el proceso iterativo dado por las distintas definiciones de éxito académico, es decir, según que criterios el algoritmo debe separar la muestra en cuestión.

Las definiciones utilizadas fueron las obtenidas en la literatura más una propuesta por nosotros:

1. El éxito académico es función de las evaluaciones.
   Para este caso se utilizó como objetivo de los criterios de separación de muestras la variable PPA, que representa la nota final con la que el alumno egresa de la escuela. Por lo tanto entregará un ranking en base a las variables más significativas para el PPA, y así poder evaluar que variables son las definen a un alumno con buenas notas.
2. El éxito académico es función de la duración de los estudios.
   En esta iteración se utilizó la variable DURACIÓN, la que se creó a partir de las fechas de ingreso a la escuela y la fecha de egreso. El algoritmo separará las variables según teniendo como objetivo la duración de los estudios. Entregará ranking de las variables y se podrá evaluar que variables son las que definen que un alumno egrese en menor o mayor tiempo que otro.
3. El éxito académico en función de las evaluaciones y la duración de los estudios.
   Por último, se creo una variable que contuviese de manera proporcional las notas y la duración de los estudios. En este caso fue importante definir la forma de crear esta nueva variable, ya que la duración de la carrera se mide en cantidad de días y las notas en un número entre 1 y 7. Por lo tanto un bajo valor en duración y un alto valor en notas significa éxito. Para este caso se decidió normalizar las variables y restarlas, resultando así una nueva variable que a medida que sea mayor, más exitoso es aquel alumno.

En cuanto a la selección de las muestras no aleatorias utilizadas, también se utilizaron los algoritmos de árboles de decisión y *"Random Forest"* con el mismo objetivo que la base

de datos completa, pero además para comparar si las variables más incidentes en el éxito académico son distintas para los siguientes casos:

1. Sub grupo de alumnos que sólo realizaron actividades extra programáticas
2. Sub grupo de alumnos que no realizaron actividades extra programáticas

A modo de resumir el proceso completo, la figura 4-3 presenta el paso a paso de nuestro proceso iterativo para facilitar la compresión. Primero se encuentra la etapa de establecer un objetivo, para el cual se debe realizar una preparación inicial de la base de datos y definir la variable de salir del algoritmo. Segundo se debe procesar la base de datos, es decir, reparar, rellenar, crear y eliminar variables para seguir adelante. Tercero se aplica el algoritmo de *"data mining"* y se analizan los resultados. Cuarto, se vuelve a la etapa dos de procesar la base de datos con el fin de reducir variables, crear nuevas o lo que sea necesario según el análisis de la etapa previa. El proceso acaba cuando se obtiene la data necesaria para concluir sobre las variables más incidentes en el éxito académico de los alumnos.

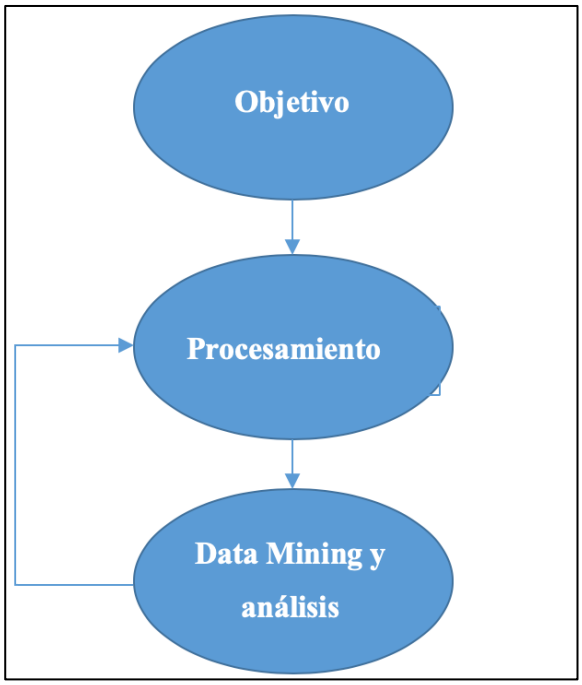


Figura 4-3:Diagrama del modelo propuesto.

5. RESULTADOS

En cuanto a la primera pregunta de investigación, "Qué variables inciden mayormente en el éxito académico de los estudiantes de la Escuela de Ingeniería UC?" se utilizó una metodología que principalmente utiliza algoritmos de *"data mining"* y procesamiento de bases de datos para lograr reducir de 1436 a 20 variables independientes a través de agrupamiento, eliminación y creación de estas según los resultados parciales del proceso. A continuación se presentarán los resultados de las etapas luego de una vez terminado el proceso de iteración.

5.1 Procesamiento de la información

El objetivo de este proceso es realizar una limpieza y comprobación de la información contenida en la base de datos a utilizar. Tal como se explicó en la sección de metodología, en esta etapa se eliminaron variables, principalmente porque se consideraron irrelevantes, o tenían muchos datos faltantes. También se agruparon variables, se crearon nuevas y se rellenaron otras según los criterios descritos en la sección 4.2. A continuación se presentan los resultados de las metodologías en tablas.

Tabla 5-1: Variables eliminadas luego del proceso iterativo.

Variables eliminadas	Motivo
- Puntaje PSU MAT, ciencias, lenguaje y ranking	La información no es comparable entre los distintos años, sólo en su propio año.

Tabla 5-2: Variables agrupadas luego del proceso iterativo.

Nueva variable	Variables agrupadas
- Act. Extra programáticas	Ayudante + Deportista UC + Participación en trabajos sociales + Participación como tutor + Participación como embajador.
- Act. Compromiso	Donaciones hacia la universidad + Trabajador en la UC + Participación en eventos UC post egresado.
- MAT	

- CIENCIAS	Cursos con sigla MAT.
- Industrial	Cursos con sigla FIS + QIM + BIO.
- ING	Cursos con sigla ICS.
- Especialidad	Cursos con sigla ING + IIC1103 + IIC1103.
- OTROS	Cursos de la especialidad del alumno.
	Cursos OFG + CARA + teológico + inglés + deportivos.

Tabla 5-3: Variables rellenas luego del proceso iterativo.

Variable rellenada	Motivo
Egresado	Para determinar quienes efectivamente egresaron.
Puesto	Se les agregó el puesto a aquellos alumnos que no tenían por ingresar por vía especial. El puesto de ellos es el último.
Preferencia	Se les agregó la preferencia a aquellos alumnos que no la presentaban por ingresar por vía especial.
Cursos de plan común	Aquellos alumnos sin notas en cursos por convalidarlos por cambio de carrera se les rellenó con el promedio.

Tabla 5-4: Variables normalizadas luego del proceso iterativo.

Variable normalizada	Motivo
- MAT	Necesidad de ser comparable.
- CIENCIAS	Necesidad de ser comparable.

- Industrial	Necesidad de ser comparable.
- ING	Necesidad de ser comparable.
- Especialidad	Necesidad de ser comparable.
- OTROS	Necesidad de ser comparable.

5.2 Árboles de decisión

Se utilizó el algoritmo de árboles de decisión con el fin de visualizar de manera preliminar y gráfica los resultados de las variables más importantes para las distintas definiciones de éxito académico y distintas iteraciones de bases de datos estudiadas. A continuación se resumen los resultados para cada instancia en la tabla 5-6. En el Apéndice C se pueden apreciar con mayor precisión cada árbol.

Tabla 5-5: Resultados árboles de decisión.

Configuración de la base de datos			
Definición de éxito académico	Base completa	Sólo alumnos que realizaron actividades extra programáticas	Sólo alumnos que no realizaron actividades extra programáticas
Éxito académico en función de las notas	1) % de aprobación 2) MAT 3) Ciencia 4) ING 5) Industrial 6) Extra programático 7) Especialidad 8) Preferencia	1) % de aprobación 2) MAT 3) Ciencia 4) Preferencia 5) ING 6) Especialidad 7) Industrial 8) Otros	1) % de aprobación 2) MAT 3) ING 4) Preferencia 5) Puesto 6) Ciencias 7) Especialidad 8) Otros

Éxito académico en función de la duración	1) % de aprobación 2) Preferencia 3) ING 4) Especialidad 5) Otros 6) ING 7) Ciencia 8) Puesto	1) % de aprobación 2) Preferencia 3) Otros 4) Industrial 5) ING 6) MAT 7) Especialidad 8) Ciencia	1) % de aprobación 2) Preferencia 3) ING 4) Industrial 5) MAT 6) Ciencia 7) Puesto 8) Especialidad
Éxito académico en función de las notas y duración	1) % de aprobación 2) MAT 3) Puesto 4) Preferencia 5) Especialidad 6) ING 7) Vía de ingreso	1) % de aprobación 2) MAT 3) Preferencia 4) ING 5) Puesto 6) Otros 7) Genero 8) Especialidad	1) % de aprobación 2) Preferencia 3) MAT 4) ING 5) Vía de ingreso 6) Especialidad

En cuanto a los resultados obtenidos y expuestos en la tabla 5-5, independiente de la iteración realizada, es decir, para las distintas definiciones de éxito académico y para las distintas muestras, siempre el resultado tiende a que las variables que mejor definen el éxito académico para el caso de estas muestras son académicas, restándole importancia a variables no académicas como las demográficas, relacionadas al compromiso o las socio económicas.

Si discutimos con mayor detalle, para la primera definición de éxito académico, donde este es función de las notas, el porcentaje de aprobación de notas y en general los cursos de los primeros años son los que mayor incidencia tienen en éxito. Para la definición en que el éxito es una función de la duración, las variables más relevantes son el porcentaje de aprobación, la preferencia de la carrera del alumno al momento de inscribirse a ingeniería y cursos introductorios de ingeniería. Esto podría ser relevante ya que se puede interpretar como los alumnos más determinados con estudiar ingeniería y con buenas calificaciones en ramos introductorios en ingeniería logran egresar antes. Por último en el

caso en que éxito se define como función de las notas y el tiempo, las variables más importantes son el porcentaje de aprobación, los ramos matemáticos del plan común y la preferencia del alumno al momento de inscribirse en la carrera. Al igual que en el caso anterior, es posible interpretar que los alumnos más determinados con Ingeniería, y con buenas habilidades en ramos matemáticos, son parte de un perfil definido como exitoso.

5.3 "Random Forest"

El objetivo de utilizar este algoritmo de *"data mining"* es similar al del árbol de decisión, pero con la particularidad de que este resultado, por la forma en que se calcula, es más confiable (Breiman, 2001) y además permite entregar un ranking de las variables más incidentes para nuestro caso de estudio. Otro de sus objetivos, es que dada la naturaleza del proceso iterativo, este algoritmo fue de ayuda para la etapa de procesamiento de la base para manejar las variables, pero todo esto se explicó en dicha sección. En cuanto a cómo se presentan los resultados, estos se presentan en orden de importancia decreciente. A cada variable se le asigna un factor entre 0 y 1, donde mientras mayor sea el factor, más relevante es y menor en el caso contrario. A continuación se resumen los resultados del algoritmo en la tabla 5-7. En el apéndice D se pueden encontrar más detalles de los resultados.

Tabla 5-6: Resultados "*Random Forest*".

Configuración de la base de datos			
Definición de éxito académico	Base completa	Sólo alumnos que realizan act. Extra programáticos	Sólo alumnos sin act. Extra programáticas
Éxito académico en	1) % de aprobación 2) MAT 3) ING 4) Especialidad	1) MAT 2) % de aprobación 3) ING 4) Ciencias	1) % de aprobación 2) ING 3) Ciencias 4) MAT

función de las notas	5) Ciencias 6) Industrial 7) Puesto 8) Otros	5) Industrial 6) Especialidad 7) Otros 8) Puesto	5) Puesto 6) Otros 7) Industrial 8) Especialidad
Éxito académico en función de la duración	1) Ciencia 2) Especialidad 3) ING 4) MAT 5) Puesto 6) Industrial 7) Otros 8) % de aprobación	1) Ciencia 2) ING 3) Industrial 4) Puesto 5) MAT 6) Otros 7) % de aprobación 8) Especialidad	1) % de aprobación 2) Ciencia 3) Industrial 4) Especialidad 5) Puesto 6) MAT 7) ING 8) Otros
Éxito académico en función de las notas y duración	1) MAT 2) Puesto 3) Ciencia 4) Industrial 5) ING 6) Otros 7) Especialidad 8) % de aprobación	1) Puesto 2) MAT 3) Industrial 4) ING 5) Especialidad 6) Ciencia 7) Otros 8) % de aprobación	1) MAT 2) % de aprobación 3) Otros 4) ING 5) Industrial 6) Ciencia 7) Especialidad 8) Puesto

Dado que *"Random Forest"* es una ponderación de múltiples resultados de árboles de decisión, no es extraño que los resultados sean similares a los árboles de decisión. De la misma forma que el caso anterior, las variables de mayor importancia en todos los casos son académicas. Esto puede interpretarse en que las actividades extra programáticas no generan impacto en que el alumno egrese con mejores calificaciones o en menor tiempo de la carrera universitaria. En general la importancia de estas 8 variables para cada iteración es similar, por lo que no se puede concluir que una tiene un efecto muy superior por sobre otra. De hecho la importancia, que se mide en escala relativa entre 0 y 1, indicando el porcentaje de importancia, todas estas variables académicas se encuentran en un rango entre 0,095 y 0,14.

Observando mayores detalles, se puede apreciar que en este tipo de resultados el porcentaje de aprobación no es la variable más importante en todos los casos como si lo fue en los árboles de decisión. También es importante recalcar que la variable que indica el puesto del alumno es relevante para el éxito académico en todos los casos, esto quiere

decir que la posición en la que entro el alumno a la Escuela de Ingeniería UC si es determinante al momento de evaluar su éxito.

6. DISCUSIÓN

En esta sección, se discutirán los resultados obtenidos en el estudio. Se hará referencia a la salida de los algoritmos utilizados y en torno a esto, responder las preguntas de investigación propuestas.

6.1 Principales descubrimientos

En cuanto a la primera pregunta de investigación "¿Qué variables inciden mayormente en éxito académico de los estudiantes de la Escuela de Ingeniería UC?" según los resultados proporcionados en primera instancia por los árboles de decisión donde se indica que la mayoría de las variables incidentes en el perfil exitoso de un alumno de Ingeniería UC son variables académicas, lo que es lógico dado que las diferentes definiciones de éxito académico que se usaron están fuertemente basadas en esas variables. Esta tendencia de variables relevantes para el éxito académico nuevamente se repite en los resultados de *"Random Forest"*, donde el algoritmo mismo claramente hace una diferencia, en términos de importancia relativa, entre las variables académicas y el resto, ya que las académicas tienen importancias entre 0,095 y 0,14, mientras que las no académicas fluctúan entre 0,005 y 0,02. Por otro lado, discutir si los cursos de ciencias de plan común son más relevantes que los cursos de matemática de plan común no tiene mucho sentido, dado que todos tienen importancias relativas similares, donde difieren entre si en un máximo (entre el mayor y el menor) de 5% y un 1,5% en el caso de ser consecutivas. Lo que si es importante discutir, entre las variables académicas más relevantes, todas excepto una se desarrollan a lo largo de la vida universitaria del alumno, como por ejemplo las calificaciones o el porcentaje de aprobación, y la excepción es el puesto de ingreso, el que

demuestra de manera relativa también el éxito previo del alumno o una ventaja comparativa de un estudiante sobre el otro. Tal como se mencionó previamente, según la literatura, el éxito futuro está relacionado con el éxito previo y acá se puede evidenciar también.

De todas formas, en cuanto a las variables no académicas más relevantes en el caso de estudio, estas son si el alumno es o no de región y si realizó o no actividades tales como ser ayudante, deportista destacado, trabajos sociales, tutor o embajador. Cabe destacar que la importancia relativa de estas variables es 5 veces menor que las académicas.

Al hacer referencia a la segunda pregunta de investigación "¿Qué se puede concluir en base a las variables más incidentes en el éxito académico del alumno?", dado que para todos los casos, los resultados indican que la tendencia es fuertemente académica. De esta información es posible extraer que el hecho de realizar actividades extra programáticas, las que podrían desarrollan habilidades distintas a los cursos de la Escuela, no promoverá ser más exitoso según las definiciones utilizadas. Este fenómeno es sin duda relevante de estudiar dado que en este estudio no se aprecia los beneficios que pueden traer estas actividades extra programáticas en el éxito académico. Para este caso específico, es probable que estas variables queden fuera ya que los alumnos al optar por estas actividades probablemente alargan sus carreras o comprometen sus calificaciones para realizar más actividades al mismo tiempo. No obstante, es altamente probable que estas actividades generen un impacto en el desarrollo del estudiante, lo que sería interesante a estudiar a futuro y analizar su importancia.

En cuanto a la variable "puesto" esta es relevante para todos los resultados de las distintas muestras (diferentes bases de datos utilizadas) y para las distintas definiciones de éxito académico. Esto indica que un alumno que ingresó a la carrera en un mejor puesto que otro alumno, es probable que su éxito a lo largo de la universidad sea mayor, es decir, termine sus estudios o con mejor promedio final o la termine antes o incluso ambas juntas. En otras palabras, se obtiene que un alumno que es exitoso previamente (mejor resultado

al ingreso), es más probable que sea exitoso de manera posterior (mejor resultado al egreso). A simple vista puede sonar lógico, pero analizando este resultado, se puede concluir que el paso por la Escuela de Ingeniería UC no define si el alumno será o no exitoso académicamente. Esto puede implicar que la base educacional previa define en una medida importante en el desempeño de un alumno en la Escuela. Una posible causa de este fenómeno es que las herramientas entregadas por la Escuela de Ingeniería UC hoy en día no generan movilidad en sus alumnos y no promueven que los alumnos sean más homogéneos en términos académicos, incluso, la mantienen o podrían volverlos más heterogéneos, es decir, que el alumno que ingresó con un mejor puesto, puede egresar con un mejor puesto y al mismo tiempo el que ingresó con un mal puesto, puede egresar con uno más bajo.

Por otro lado, parámetros socio económicos como si el alumno pertenece a la región metropolitana o si proviene de un colegio particular pagado o incluso su género, no parecen ser relevantes para determinar el éxito académico en la Escuela de Ingeniería UC.

6.2 Consideraciones metodológicas

A continuación se discutirá sobre las metodologías utilizadas para obtener los resultados, las diferencias entre estos y las limitaciones de la ruta elegida. Para llevar a cabo esto se responderán las siguientes preguntas:

1. ¿Qué problemas insolubles se presentaron en la metodología y cómo se abordó?
2. ¿Cuál es la diferencia entre los resultados obtenidos en ambos algoritmos, cuál es que se toma más en cuenta y por qué?

Para responder la primera pregunta, lo primero es hablar de los problemas insolubles. Dentro de estos se presentaron 2 principalmente. El primero relacionado con el sesgo de

la información y de los resultados y el segundo relacionado con la definición del éxito académico.

En cuanto a la problemática del sesgo, el primer problema fue sobre el sesgo de la base de datos utilizada. Al ser todos los alumnos de la misma casa de estudios, de una cantidad de años no muy extensa y seguidos, puede no existir gran diferencia entre los distintos alumnos a pesar de ser cerca de 2700 alumnos estudiados. Puede incluso existir un sesgo de los mismos alumnos al momento de ingresar a una determinada casa de estudio, lo que puede provocar que en mayor proporción ingresen estudiantes con ciertas características a cada institución. Para abordar esta problemática, como no fue posible acceder a la información de otras casas de estudio por confidencialidad de la información y tampoco agregar una muestra más extensa por la capacidad de la Dirección de Pregrado de la Escuela de Ingeniería UC, se debe aceptar el sesgo e incluirlo en los resultados, es decir, sin importar el resultado de la investigación, este resultado será válido bajo ciertas condiciones, en este caso, para los alumnos de ingeniería de la UC para los 6 años estudiados.

Por otro lado, existe un sesgo generado por las técnicas utilizadas de minería de datos. Dado que principalmente se utilizaron árboles de decisión y "*Random Forest"*, técnicas que operan de manera similar según lo explicado en la sección 4.3.1 y 4.3.2. Este sesgo provoca que los resultados de ambas técnicas sean similares. Para abordar esta problemática fue necesario hacer una revisión bibliográfica previa sobre los resultados de otras técnicas de minería de datos para comprobar si se llegó a resultados similares.

En cuanto a la problemática sobre el sesgo de la definición de éxito académico utilizada para esta investigación, esta se genera principalmente dado que el resultado depende de la definición tomada. Para esta investigación se utilizó las definiciones de éxito académico encontradas en la literatura explicadas en la sección 3. Para abordar la problemática de obtener resultados similares a los obtenidos por otros autores, de manera adicional se creo

una nueva definición de éxito académico, la que es combinación de las otras 2 utilizadas con le fin de descubrir si se obtienen distintos resultados. El detalle de esta definición se encuentra en la sección 4.4.

Para visualizar de mejor manera los resultados del estudio, estos se representan en las tablas 5-5 y 5-6 de la sección 5.2 y 5.3 respectivamente. Los resultados de ambos no muestran mayores diferencias, ya que siempre las variables más destacadas son las mismas en ambos casos. Los resultados indican que las variables que mayormente explican el éxito académico bajo las distintas definiciones son variables del grupo académicas (la clasificación de las variables se representa en la tabla 4-2 de la sección 4.2.1). Sólo varía la importancia relativa de las variables destacadas. En cuanto a qué resultado se toma más en cuenta, los obtenidos de la técnica *"Random Forest"* son los indicados ya que según la literatura y por cómo funcionan (son múltiples árboles de decisión ponderados, esto se explica con mayor detalle en las secciones 4.3.2) son más precisos que los árboles de decisión por si solos.

7. CONCLUSIONES

En primer lugar, la contribución de este estudio es la metodología utilizada sobre una muestra de alumnos de Ingeniería UC, donde primero se trabajó con una base de datos en la que se pre proceso la información con el fin de tener un elemento sobre el cual trabajar técnicas de minería de datos de la manera correcta. Este procedimiento es relevante dado que evidencia como trabajar sobre una base de datos universitaria, dejando claro los distintos desafíos y potenciales inconvenientes que pueden surgir. En segundo lugar, elegir y testear modelos de *"data mining"* para automatizar la detección de patrones para realizar análisis más profundos con los alumnos de nuestra casa de estudios, es un paso importante para comprender mejor los hechos que suceden en la Escuela. También es

importante para mostrar que estas nuevas técnicas se pueden aplicar en un caso cercano y complementarlas con estudios tanto previos como futuros para poder obtener resultados interesantes.

En este trabajo se detectaron patrones de manera automática para poder responder la primera pregunta de investigación "¿Qué variables son las que más inciden en un alumno de la Escuela de Ingeniería, en si es o no exitoso en su carrera universitaria?". Para lograr esta automatización se detalló la metodología de trabajo y las definiciones de éxito académico a considerar en el caso de los alumnos de Ingeniería UC. La sección de metodología y resultados entregan los detalles necesarios para comprender cómo se aplicaron tanto los algoritmos de árboles de decisión y *"Random Forest"* en un proceso iterativo de procesamiento de la base de datos para determinar la respuesta a nuestra pregunta de investigación. En cuanto a la segunda pregunta de investigación, "¿Qué se puede concluir en base a las variables más incidentes en el éxito académico del alumno?, es posible concluir que dado las definiciones actuales de éxito académico y los resultados los alumnos estudiados, en este caso, las variables más incidentes son del ámbito académicas, siendo incluso hasta 5 veces más importantes que las no académicas. Por otro lado, gracias al nivel de significancia relativo de cada variable (que se puede apreciar en los apéndices), es posible concluir que no existe una sola variable significativamente superior en términos de importancia al explicar el éxito académico dado que todas están muy cercanas en cuanto a importancia relativa. Por último es relevante destacar que los resultados indican que el puesto de ingreso a la Escuela al momento de inscribirse si es muy relevante para el éxito académico, lo que indica que el éxito en la Escuela en parte está definido por el éxito previo.

En cuanto a las limitaciones del estudio, se reconocen principalmente de 2 tipos, la primera relacionada con la información y la segunda relacionada con las técnicas de minería de datos utilizadas.

Al referirse a las limitaciones de la información, primero es importante reconocer que la muestra utilizada no es representativa de lo que es la Escuela de ingeniería UC ya que sólo incluye a los alumnos egresados de 6 años de la historia de la Escuela. Segundo es la definición utilizada de éxito académico. Este concepto no es simple, y por lo mismo representarlo con un solo indicador, en nuestro caso, el promedio ponderado o la duración de los estudios es difícil ya que sabemos a cierta si el éxito es una ponderación de esas variables o si hay algo más que agregar. En tercer lugar, realizar el estudio sólo en alumnos de Ingeniería UC. A pesar de que la metodología es un aporte al ser similar para otra casa de estudios, esto provoca que los resultados y conclusiones sean válidas solo para nuestra Escuela, ya que se genera un sesgo al sólo analizar un tipo de estudiante.

Al referirse a las limitaciones de las técnicas de minería de datos utilizadas, en primer lugar es relevante destacar que ninguna técnica es perfecta al realizar su trabajo. En segundo lugar, a pesar de utilizar los árboles de decisión y *"Random Forest"*, faltó utilizar una tercera técnica basada en otro funcionamiento, como por ejemplo, técnicas de clasificación y segmentación, o métodos bayesianos, ya que tanto los árboles de decisión como *"Random Forest"*, funcionan de manera similar, generando un sesgo en los resultados dada su similitud.

En cuanto potenciales trabajos futuros, se espera que sigan trabajando este tema pero a mayor escala, es decir, con una muestra más voluminosa y con más técnicas de minería de datos para obtener mejores resultados en términos de representatividad. Los potenciales trabajo del futuro pueden descubrir con mayor detalle qué variables explican el éxito académico o incluso pueden testear nuevas definiciones de éxito académico. El hecho de probar nuevas definiciones de éxito académico es relevante dado que al menos, según el perfil de egreso de los alumnos de Ingeniería UC, se espera que estos tengan habilidades que no son tomadas en cuenta por las definiciones de éxito en la literatura, como por ejemplo ser reflexivos y proactivos o tener sólidos valores entre otras.

Por otro lado, se espera que en el futuro este tipo de trabajos se complemente con estudios sobre el éxito laboral y ver como se relaciona el éxito académico con la fase posterior a la educacional. Este tipo de trabajos será un desafío dado que será necesario trabajar con las mismas muestras en el mundo laboral y profesional para poder compararlas y obtener conclusiones a partir de sus resultados.

REFERENCIAS

York, T. T., Gibson, C., & Rankin, S. (2015). *Defining and Measuring Academic Success - Practical Assessment, Research & Evaluation.* Practical Assesment, Research & Evaluation, 20(5), 1–20.

Breiman, L. (2001). *Random forests. Machine learning*, 45(1), 5-32

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining. Intelligent Systems Reference*

Perfil del ingeniero. (2018) Admisión e información al postulante. Ingeniería. Recuperado de http://admisionyregistros.uc.cl/

Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). *Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering Education, 93(4), 313–320.*

Burton, Lorelle J. & Dowling, David G. (2009). *Key factors that influence engineering student's academic success: a longitudinal study.* USQ eprints.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

K. Murphy. (2012). "*Machine Learning, A Probabilistic Perspective*". MIT Press

C. Bishop. (2006). "*Pattern Recognition and Machine Learning*". Springer.

T. Hastie, R. Tibshirani, and J. Friedman. (2001). "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*", Springer.

Adelson, J. L., Dickinson, E. R., & Cunningham, B. C. (2016). A Multigrade, Multiyear Statewide Examination of Reading Achievement: Examining Variability Between Districts, Schools, and Students. *Educational Researcher*, *45*(4), 258-262.

U . K. Pandey, and S. Pal. (2011). *"Data Mining: A prediction of performer or underperformer using classification"*, (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690.

Z. N. Khan. (2005). *"Scholastic achievement of higher secondary students in science stream"*, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87.

Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar. (2006). *"Mining student data using decision trees"*, International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan.

Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan. (2010). *"Data mining model for higher education system"*, Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29.

Baradwaj, Brijesh & Pal, Saurabh. (2011). Mining Educational Data to Analyze Students' Performance. International Journal of Advanced Computer Science and Applications. 2. 63-69.

Zhang, Guili & Anderson, Tim & Ohland, Matthew & Carter, Rufus & Thorndyke, Brian. (2004). Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study. Journal of Engineering Education. 93.

Kamphorst, Jan & Hofman, W & Jansen, Ellen P.W.A. & Terlouw, Cees. (2015). Explaining Academic Success in Engineering Degree Programs: Do Female and Male Students Differ?. Journal of Engineering Education. 104.

Chowhan, S. (2013). 'Academic Performance of Engineering Students: The Role of Abilities & Learning Style'. World Academy of Science, Engineering and Technology, Open Science Index 73, International Journal of Educational and Pedagogical Sciences, 7(1), 308 - 314.

Van den Broeck, Lynn & De Laet, Tinne & Lacante, Marlies & Pinxten, Maarten & Soom, Carolien & Langie, Greet. (2018). Predicting the academic achievement of students bridging to engineering: the role of academic background variables and diagnostic testing. Journal of Further and Higher Education. 1-19.

Rauri McCool, Sinead Kelly, Moira Maguire, Dermot Clarke, Damian Loughran. (2015). "Factors which influence the academic performance of level 7 engineering students". AISHE-J, Volume 7, 2.

Abilash, R. (2018). Ilustración de Random Forest. [Figura]. Recuperado de: https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57

Naren, J. & Elakia, & Gayathri, & Aarthi,. (2014). Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students. International Journal of Engineering and Technology. vol 5(3). 4469-4472.

Silva, C., & Fonseca, J. (2017). Educational Data Mining: A Literature Review. In *Europe and MENA Cooperation Advances in Information and Communication Technologies* (pp. 87-94). Springer International Publishing.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1), 3133-3181.

APÉNDICES O ANEXOS

Apéndice A: Bases de datos originales

En esta sección se mostrará por un lado cómo lucían las bases de datos en un comienzo y luego cómo quedaron para evidenciar de manera visual el proceso de reducción de variables obtenido por el proceso iterativo. En la figura 9-1 se podrá apreciar cómo la información relacionada con las calificaciones de cada curso es registrada por la Dirección de Pregrado de la Escuela de Ingeniería UC. En la figura 9-2 se puede apreciar cómo es registrada el resto de las variables de cada alumno. Por último en la figura 9-3 se presenta cómo luce la base de datos luego de todo el proceso iterativo.

ID FALSA	AÑO	SEMESTRE	SIGLA	SECCIÓN	NOMBRE CURSO	CREDITOS C	NOTA FINAL	NOTA FINAL
2009_011	2009	1	VRA100C	1	Examen de Comunicac	0		A
2009_011	2009	2	QIM100	3	Quimica General	10	3,8	
2009_011	2009	1	ICS1502	0	Introduccion a la Econ	10	5,3	
2009_011	2009	1	ICS2522	5	Contabilidad	10	5,8	
2009_011	2009	1	IIC1102	3	Introduccion a la Progr	10	6,1	
2009_011	2009	2	IIC1222	1	Programacion Avanzac	10	4,6	
2009_011	2009	1	MAT1600	2	Introducción al Cálculo	10	4,6	
2009_011	2009	2	MAT1610	7	Cálculo I	10	4,3	
2009_011	2009	2	MAT1203	2	Álgebra Lineal	10	4,2	
2009_011	2009	1	ING1004	1	Desafíos de la Ingenie	10	5,3	
2009_011	2009	1	ING1001	1	Practica I	0		A
2009_011	2009	1	VRA2010	1	English Placement Tes	0		A
2009_011	2009	1	COE0060	0	Cr Optativos Otras Disc	60		C
2009_011	2009	1	FIS1503	8	Física General	10	5,9	
2009_011	2009	1	CTE0010	0	Cr Formacion Teologica	10		C
2009_011	2009	1	CAE010	0	Curso Antropologico-Et	10		C
2009_011	2009	1	OPT050	0	Creditos Optativos Ger	50		C
2009_011	2010	1	MAT1630	2	Cálculo III	10	4,1	
2009_011	2010	1	MAT1640	2	Ecuaciones Diferencial	10	5,2	
2009_011	2009	3	MAT1620	2	Cálculo II	10	4,5	
2009_011	2010	1	FIS1513	2	Estática y Dinámica	10	5	
2009_011	2010	1	IEE2712	1	Sistemas Digitales	10	5,3	
2009_011	2010	1	ICS3502	4	Marketing	10	5,3	
2009_011	2010	2	FIS1533	5	Electricidad y Magnetis	10	4,5	
2009_011	2010	2	IEE2103	1	Señales y Sistemas	10	4	
2009_011	2010	2	IIQ1003	1	Termodinámica	10	5,8	
2009_011	2010	2	ICS3532	3	Finanzas	10	5,1	
2009_011	2010	2	FIS0152	8	Laboratorio de Termod	0		A
2009_011	2010	2	ICS1113	4	Optimización	10	5	
2009_011	2010	2	FIS0153	5	Laboratorio de Electrici	0		A
2009_011	2011	1	ICS2123	1	Modelos Estocásticos	10	4,7	
2009_011	2011	1	IEE2123	2	Circuitos Eléctricos	10	4	
2009_011	2011	1	IEE2513	1	Comunicaciones	10	4,3	
2009_011	2011	1	ICS2523	2	Microeconomía	10	4,5	
2009_011	2010	3	EYP1113	1	Probabilidades y Estad	10	4,5	
2009_011	2011	2	IEE2183	1	Laboratorio de Medicio	5	6,2	
2009_011	2011	2	IEE2213	1	Máquinas Eléctricas	10	4	
2009_011	2011	2	IEE2413	1	Electrónica	10	4,6	
2009_011	2011	2	IEE2613	1	Control Automático	10	4,5	
2009_011	2012	1	ICS2813	1	Organización y Compo	10	4,1	
2009_011	2012	1	IEE2313	1	Sistemas de Potencia	10	4,9	
2009_011	2012	1	IEE2783	1	Laboratorio de Sistema	5	6,6	
2009_011	2012	2	ICS3332	1	Gestion de Operacione	10	4,9	
2009_011	2012	2	IEE2683	1	Laboratorio de Control	5	5,6	
2009_011	2013	1	IEE3373	1	Mercados Eléctricos	10	4,6	
2009_011	2013	1	IEE3912	1	Diseno Electrico	10	5,3	
2009_011	2013	2	IEE2113	1	Teoría Electromagnétic	10	5,7	
2009_011	2013	2	IEE3313	1	Planning, Efficiency an	10	5,1	
2009_011	2013	1	IEE2273	3	Laboratorio de Máquin	5	5,6	
2009_011	2013	2	ING2001	1	Practica II	0		A
2009_011	2013	2	ICS3013	5	Evaluación de Proyect	10	6,4	
2009_011	2014	1	QIM100	4	Quimica General	10	4,5	
2009_011	2014	1	FIS0151	0	Laboratorio de Estatica	0		C
2009_011	2014	1	ICH1104	2	Mecánica de Fluidos	10	1	

Figura 0-1:Aspecto de la base de datos original.

En este caso podemos ver que cada curso es registrado en una fila nueva para cada alumno.

ID-falsa	CURRICULU	CODIGO PRO	PROGRAMA	ESTADO	AÑO INGRES	VÍA DE INGR	TIPO INGRES	PROM. PPA	TOTAL CRÉDITOS APROBADOS	TOTAL CRÉDITOS CONVALIDADO	TOTAL CRÉDITOS APROBADOS+	TOTAL CRÉDITOS REPROBADOS	TOTAL CRÉDITOS INSCRITOS	N° CAUSALES ELIMINACIÓN (POR	LICENCIADO	FECHA LIC.	EGRESADO
2009_001	C2009	60056	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	DEPORTISTA	5,55	45	0	45	0	45	0	SI	31/07/2014	SI
2009_002	C2009	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	BECA EXCEL	6,49	55	0	55	150	705	1			
2009_003	C2009	60006	ING CIVIL ME	REGULAR	2009	ESPECIAL	HIJO DE PRI	6,61	20	0	20	150	590	0			
2009_004	C2009	60005	ING CIVIL EL	NO VIGENTE	2009	ESPECIAL	HIJO DE PRI	5,33	60	150	210	10	270	0	SI	31/12/2014	SI
2009_005	C2009	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	HIJO DE PRI	6,98	30	0	30	0	30	0	SI	31/12/2013	SI
2009_006	C2009	60052	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	HIJO DE PRI	5,34	30	50	80	0	130	0	SI	31/07/2014	SI
2009_007	C2009	60002	ING CIVIL PL	NO VIGENTE	2009	ESPECIAL	BECA EXCEL	6,31	200	0	200	50	290	2			
2009_008	C2009	60062	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,85	550	50	640	120	870	0			SI
2009_009	C98M	60094	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,45	525	50	575	0	625	0	SI	31/12/2013	SI
2009_010	C98M	60052	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,32	462	50	512	0	662	0	SI	31/12/2012	SI
2009_011	C98M	60062	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	EGRESADO	6,91	500	146	646	50	820	0			SI
2009_012	C98M	60007	ING CIVIL CC	NO VIGENTE	2009	ESPECIAL	EGRESADO	6,38	320	319	639	0	920	0	SI	01/03/2012	SI
2009_013	C98M	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,17	395	155	550	0	995	0	SI	31/12/2012	SI
2009_014	C98M	60093	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,24	385	165	550	0	985	0	SI	01/03/2012	SI
2009_015	C98M	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,33	335	250	585	0	335	0	SI	31/12/2012	SI
2009_016	C98M	60052	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,13	360	210	570	0	760	0	SI	31/07/2014	SI
2009_017	C98M	60052	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,46	420	155	575	0	620	0	SI	31/12/2012	SI
2009_018	C2009	60095	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,95	510	50	560	50	660	0			SI
2009_019	C98M	60093	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,01	580	50	630	10	690	0	SI	31/12/2013	SI
2009_020	C98M	60094	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,04	512	50	562	25	537	0			SI
2009_021	C98M	60095	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,54	665	50	555	0	665	0	SI	31/07/2013	SI
2009_022	C2009	60054	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	6,83	505	40	545	50	535	0			SI
2009_023	C2009	60072	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	6,00	530	0	530	70	600	0	SI	31/12/2014	SI
2009_024	C2009	60004	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CONVENIO I	6,54	555	130	685	0	755	0	SI	31/07/2011	SI
2009_025	C2009	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CONVENIO I	5,21	170	280	680	0	170	0	SI	31/12/2011	SI
2009_026	C98M	60095	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	Cambio de C	6,57	430	115	545	0	630	0	SI	31/07/2013	SI
2009_027	C2009	60007	ING CIVIL CC	NO VIGENTE	2009	ESPECIAL	Cambio de C	6,80	555	15	570	50	670	0	SI		SI
2009_028	C98M	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	Cambio de C	6,06	510	80	590	50	670	0	SI	31/07/2014	SI
2009_029	C2009	60054	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	Cambio de C	6,84	515	60	575	0	715	0	SI	31/12/2013	SI
2009_030	C2009	60002	ING CIVIL PL	NO VIGENTE	2009	ESPECIAL	Cambio de C	6,43	515	65	580	70	785	0			SI
2009_031	C2009	60022	ING CIVIL-ES	NO VIGENTE	2009	ESPECIAL	ENSENANZA	6,14	570	0	570	20	590	0	SI	31/12/2014	SI
2009_032	C2009	60095	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	6,17	530	0	530	0	530	0	SI	31/12/2013	SI
2009_033	C2009	60072	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	5,22	540	0	540	50	580	0	SI	31/07/2014	SI
2009_034	C2009	60002	ING CIVIL PL	NO VIGENTE	2009	ESPECIAL	CARRERAS	5,79	0	170	170	0	0	0			
2009_035	C98M	60005	ING CIVIL EL	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,06	585	110	695	50	695	0	SI	31/12/2013	SI
2009_036	C98M	60022	ING CIVIL-ES	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,87	555	115	670	50	585	0	SI	31/07/2013	SI
2009_037	C98M	60092	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,58	530	320	850	10	540	0	SI	03/03/2014	SI
2009_038	C98M	60052	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,51	545	345	890	70	615	0	SI	31/07/2013	SI
2009_039	C98M	60054	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,19	580	175	755	0	580	0	SI	31/12/2013	SI
2009_040	C98M	60093	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,31	445	110	555	120	565	0			SI
2009_041	C98M	60022	ING CIVIL-ES	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,53	425	115	540	50	515	0			SI
2009_042	C2009	60096	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	5,43	535	0	535	50	555	0	SI	31/07/2014	SI
2009_043	C2009	60002	ING CIVIL PL	NO VIGENTE	2009	ESPECIAL	ENSENANZA	5,45	520	0	520	10	530	0			
2009_044	C2009	60093	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	ENSENANZA	5,47	535	0	535	210	745	1			
2009_045	C2009	60002	ING CIVIL PL	NO VIGENTE	2009	ESPECIAL	ENSENANZA	6,02	55	0	55	70	155	0			
2009_046	C98M	60018	ING CIVIL-GE	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,81	540	210	750	50	990	0	SI	31/07/2013	SI
2009_047	C98M	60093	ING CIVIL IN	NO VIGENTE	2009	ESPECIAL	CAMBIO CAR	6,93	505	150	655	50	645	0	SI	31/12/2013	SI
2009_048	C98M	60006	ING CIVIL ME	NO VIGENTE	2009	ESPECIAL	CAMBIO DE	6,05	280	265	545	50	500	0	SI	31/07/2012	SI
2009_049	C2009	60004	ING CIVIL IN	NO VIGENTE	2009	ORDINARIA	I	5,09	545	50	595	10	555	0	SI	31/12/2013	SI
2009_050	C2009	60095	ING CIVIL IN	NO VIGENTE	2009	ORDINARIA	I	5,19	530	50	580	50	580	0	SI	31/12/2013	SI
2009_051	C2009	60004	ING CIVIL IN	VIGENTE EG	2009	ORDINARIA	I	5,37	580	20	600	50	600	0	SI	31/12/2013	SI
2009_052	C2009	60096	ING CIVIL IN	NO VIGENTE	2009	ORDINARIA	E	5,25	590	20	610	20	610	0	SI	31/12/2014	SI

Figura 0-2:Continuación base de datos original.

APROBADOS	COLEGIO RE	COMPROMI	EXTRAPROG	FECHA EGR.	GENERO	ID-FALSA	N-=> CAUSAL	PARTICULAR	PREFERENCI	PROGRAMA	PUESTO	VIA DE INGR	CIENCIA	MATEMATICAS	ING	OTROS	ESPECIALIDA	INDUSTRIAL	DURACION Y	PPA
0,93442623	1	1.0	0	31-12-12	0	2007_001	0	1	1	ING CIVIL INI	519	0	-0.40111534	0.048981326	90367027	-0.34043844	0.017482847	0.052380970	0.035761441	-0.021746025
0,96747968	1	1.0	0	31-07-13	0	2007_002	0	1	1	ING CIVIL INI	642	0	-0.14452802	0.066397931	55031809	0.015784222	0.018663977	-0.008526251	0.296226468	0.260549205
0,96460177	0	0.0	1	31-07-13	1	2007_003	0	0	0	ING CIVIL INI	778	0	-0.34602266	0.061851651	75454366	0.175634000	-0.002137739	0.007477160	0.017606753	0.682685088
0,88059702	1	0.0	1	31-07-13	0	2007_004	0	0	0	ING CIVIL INI	779	0	-0.41310080	-0.007281091	69608300	0.009607989	0.012095416	0.037468301	0.068576632	##########
0,96581197	1	0.0	1	31-12-12	1	2007_005	0	0	0	ING CIVIL INI	780	0	-0.38455836	0.037747180	48047602	-0.27333276	0.018957332	0.026484097	0.264269034	##########
0,88235294	1	0.0	1	31-07-14	0	2007_006	1	1	1	ING CIVIL INI	777	0	-0.05669802	0.002784367	71965145	-0.04276151	0.034093711	-0.037398170	0.006404313	##########
0,7	1	0.0	1	31-07-11	0	2007_007	3	0	0	ING CIVIL INI	782	0	-0.06608784	##########	##########	-0.18329728	0.146916896	0.271565174	0.005994569	##########
1	1	0.0	1	31-07-12	0	2007_008	0	1	0	ING CIVIL-GE	784	0	-0.37686417	0.176950595	99467666	-0.21529080	0.016996435	0.138273621	-0.226842761	##########
0,86407767	1	0.0	0	31-07-12	0	2007_010	0	1	0	ING CIVIL INI	785	0	-0.17577227	0.002349949	99067676	-0.34128897	-0.012767020	0.000980071	0.462060890	-0.573587796
0,95955882	1	0.0	1	31-07-12	0	2007_012	0	1	0	ING CIVIL INI	786	0	-0.43676140	0.041098565	22838573	-0.23596118	-0.00128647	0.013549781	0.086716024	##########
0,85714286	1	0.0	0	31-07-13	0	2007_013	0	1	0	ING CIVIL INI	787	0	-0.42208237	0.095738157	0787836	-0.26172610	0.171290043	-0.113222250	0.191271465	##########
0,91304348	1	0.0	0	31-12-12	0	2007_014	0	1	0	ING CIVIL INI	789	0	-0.45600940	0.011144408	51853739	-0.29371963	0.016550727	-0.002837070	0.104779551	-0.667365610
1	1	0.0	0	31-12-09	0	2007_015	0	1	0	ING CIVIL INI	790	0	-0.40532518	0.129555202	66668464	-0.30816140	0.087031952	0.135824565	0.082844678	##########
0,96226415	0	0.0	1	31-12-09	0	2007_016	0	0	0	ING CIVIL-ES	791	0	-0.42195487	0.129555202	66668464	-0.30816140	0.012604848	0.263692652	-0.234216551	##########
0,86885246	1	0.0	0	31-07-10	0	2007_017	0	1	0	ING CIVIL INI	792	0	-0.42195487	0.129555202	66668464	-0.30816140	0.006417959	-0.020195011	0.024459991	##########
0,95061728	0	0.0	0	31-12-10	0	2007_018	0	0	0	ING CIVIL-ES	793	0	-0.42195487	0.089906574	27806796	-0.30816140	0.001654481	0.208704235	-0.221704491	##########
0,94871795	0	0.0	0	31-07-11	1	2007_019	0	0	0	ING CIVIL-HI	794	0	-0.43031559	0.089906574	27806796	-0.30816140	-0.027010270	0.197122042	-0.212495331	##########
0,95238095	0	0.0	1	31-07-11	0	2007_020	0	0	0	ING CIVIL-GE	795	0	-0.42195487	0.080989248	13876579	-0.30816140	-0.032785580	0.218388303	-0.222319170	##########
0,89655172	0	0.0	0	31-12-09	0	2007_021	0	0	0	ING CIVIL INI	796	0	-0.42195487	0.129555202	66668464	-0.30816140	-0.018224120	0.060218501	0.031567460	##########
0,89956332	1	0.0	1	31-12-10	0	2007_022	0	1	0	ING CIVIL INI	798	0	-0.40959257	0.117660614	15010083	-0.30816140	0.015083246	0.066522934	0.316809875	##########
1	0	0.0	0	31-12-10	0	2007_023	0	1	0	ING CIVIL-ES	799	0	-0.42195487	0.089906574	27806796	-0.30816140	-0.01825824	0.242604036	-0.239065041	##########
0,828125	0	0.0	0	31-07-14	0	2007_024	1	1	0	ING CIVIL INI	800	0	-0.25410258	0.005793342	04306913	0.018577942	0.053347034	-0.033909080	0.014968775	##########
0,90909091	1	0.0	0	31-07-12	0	2007_025	0	1	0	ING CIVIL INI	801	0	-0.42993348	0.053899640	71080389	-0.25606453	-0.003650670	0.027044421	0.165800754	0.667988328
0,8041958	0	0.0	0	31-12-13	0	2007_027	0	0	0	ING CIVIL INI	802	0	-0.48646530	0.003895967	12634799	-0.00461352	-0.012549930	-0.028232690	0.031619644	##########
0,96774194	0	0.0	1	31-12-12	0	2007_028	0	0	0	ING CIVIL INI	804	0	-0.44125497	0.040710870	74773458	0.022233082	0.002756029	-0.010446814	0.196314061	0.251400721
0,86046512	0	1.0	0	31-07-13	0	2007_029	0	0	0	ING CIVIL INI	806	0	-0.48787345	0.004991586	73007477	0.107006036	-0.024681310	0.058020923	0.086126701	##########
0,83458647	1	0.0	0	31-12-13	0	2007_030	0	1	0	ING CIVIL CC	807	0	-0.35672901	-0.012301779	30952651	-0.00986508	-0.001832580	0.239090518	-0.140160577	##########
0,95876289	1	0.0	0	31-12-12	0	2007_031	0	1	0	ING CIVIL INI	808	0	-0.40186074	0.083179450	50470715	0.331693396	-0.038500300	0.131671051	0.027873805	0.847357261
1	0	0.0	0	31-07-08	0	2007_033	0	0	0	ING CIVIL ELI	809	0	-0.92831560	-0.982569669	55546255	-0.97354793	-0.013193660	0.220540606	-0.301444291	##########
1	1	0.0	0	31-12-12	1	2007_036	0	1	2	ING CIVIL INI	77	1	-0.62884609	-0.607572904	77662211	-0.97354793	-0.049967071	-0.003605110	0.016725392	##########
1	1	0.0	1	31-07-13	1	2007_037	0	1	2	ING CIVIL INI	36	1	-0.79727519	-0.659488781	99215114	-0.97354793	-0.015253074	-0.054876841	0.055402307	0.434369863
1	1	0.0	1	31-07-13	1	2007_038	0	1	2	ING CIVIL INI	217	1	-0.68827137	-0.645797422	19374435	-0.97354793	-0.053281360	-0.002966690	-0.006913510	0.310212250
1	1	0.0	1	31-07-13	0	2007_039	0	1	2	ING CIVIL INI	112	1	-0.65406519	-0.787624630	72730133	-0.97354793	-0.064573340	-0.095880930	0.105041622	##########

Figura 0-3:Base de datos final.

Apéndice B: Variables

A continuación se entrega una descripción de cada una de las variables presentes en la base de datos original entregada por la Dirección de Pregrado de la Escuela de Ingeniería UC. La información se resume en la tabla B-1.

Tabla B-1: Descripción de las variables.

Variable	Descripción
Curriculum	Indica el código del curriculum al que el alumno pertenece.
Código programa	Código del programa que el alumno elige.
Programa	Programa elegido. Por ejemplo: Ing. Civil Eléctrica.
Código preferencia especialidad	Código preferencia declarada.
Preferencia especialidad	Preferencia declarada.
Código preferencia major	Código preferencia declarada.
Preferencia major	Preferencia declarada.
Código major Banner	Código preferencia declarada.
Major Banner	Preferencia declarada.
Código minor Banner	Código preferencia declarada.
Minor Banner	Preferencia declarada.
Código Minor Siding	Código preferencia declarada.
Minor Siding	Preferencia declarada.
Estado	Estado actual del alumno: Regular, no vigente, abandono.
Año ingreso	Año de ingreso a la carrera.
Vía de ingreso	Declara si es ordinaria vía PSU o especial.
Tipo de ingreso	Detalle de ingreso especial: Deportista, beca o cambio carrera.
Prom. PPA	Promedio final del alumno.
Total créditos aprobados	Créditos aprobados del alumno.
Total créditos convalidados	Créditos convalidados del alumno.
Total créditos aprobados + convalidados	Créditos totales debidamente realizados por alumno.
Total créditos reprobados	Créditos reprobados por alumno.
Total créditos inscritos	Créditos inscritos por alumno. No cuenta créditos eliminados.
N causales de eliminación	Cantidad de causales de eliminación de un alumno.
Licenciado	Indica si tiene el grado o no.
Egresado	Indica si tiene el grado o no.
Fecha egreso	Fecha de egreso.
Fecha licenciatura	Fecha de licenciatura.
Titulado	Indica si tiene el grado o no.
Fecha titulación	Fecha de titulación.
Instrumento de titulación	Tipo de titulación: examen o tesis.
Examen de licenciatura aprobado	Indica si cumple o no.
Fecha examen de licenciatura	Fecha del examen de licenciatura.
Examen de titulo aprobado	Indica si cumple o no.
Colegio de procedencia	Indica colegio de egreso.
Colegio región	Indica la región del colegio de egreso.
Colegio tipo	Indica si colegio es particular, municipal o subvencionado.
Fecha egreso colegio	Indica año de egreso del colegio.
Fecha examen de titulo	

Ciclo 2 (preferencia)	Fecha del examen de titulo.
Salida al mercado	Indica preferencia de especialidad.
Continuidad de estudios	Indica si trabaja o no.
Detalle pregrado	Indica la continuidad de estudios.
Detalle postgrado	Información extra.
Tipo ingreso	Información extra.
Puntajes PAA	Indica si ingresó por college o bachillerato.
Puntajes PSU	Indica puntajes obtenidos en diversas pruebas.
Puesto	Indica puntajes obtenidos en diversas pruebas.
Preferencia	Puesto de ingreso a la universidad ordenado según puntaje.
Cursos	Preferencia declarada en DEMRE al postular a la carrera.
	Calificaciones de cada uno de los cursos del alumno.

Apéndice C: Resultados árboles de decisión

En este apéndice se presenta el detalle de los resultados obtenidos por los árboles de decisión para cada una de las iteraciones y muestras. La figura 9-4, 9-5 y 9-6 representan el resultado para el caso que se utilizó la base de datos completa, donde 4, 5 y 6 representan las iteraciones con las diferentes definiciones de éxito académico, cuyo orden es, el éxito académico en función de las notas (promedio ponderado de egreso), en función de la duración de la carrera y por último la combinación de estas dos respectivamente.

Por otro lado, 9-7, 9-8 y 9-9 representan el resultado para el caso que se utilizó como base de datos sólo a los alumnos que realizaron alguna actividad extra programática. De manera similar al caso anterior, 9-7 representa la iteración con definición de éxito académico en función de las notas (promedio ponderado de egreso), 9-8 en función de la duración de la carrera y el las notas de egreso y 9-9 en función de la duración de la carrera.

Por último, 9-10, 9-11 y 9-12 representan el resultado para el caso que se utilizó como base de datos a los alumnos que no realizaron alguna actividad extra programática en su vida académica. 9-10 representa la iteración con definición de éxito académico en función

de las notas (promedio ponderado de egreso), 9-11 en función de la duración de la carrera y por último 9-12 es el resultado de la definición de éxito académico como combinación de ambas variables (duración y notas).

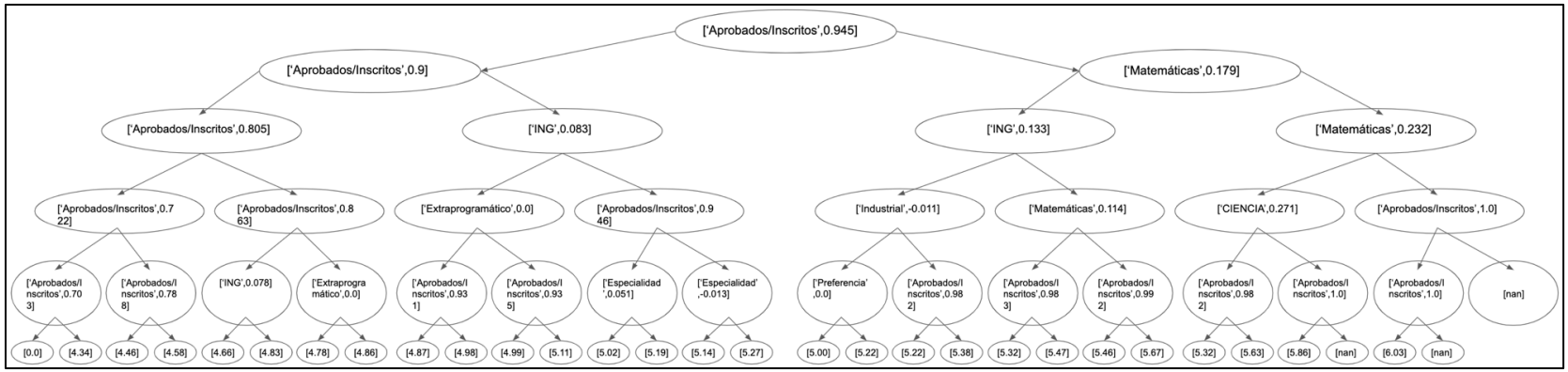


Figura 0-4:Resultado de la base de datos completa con el éxito académico definido como función de las notas.

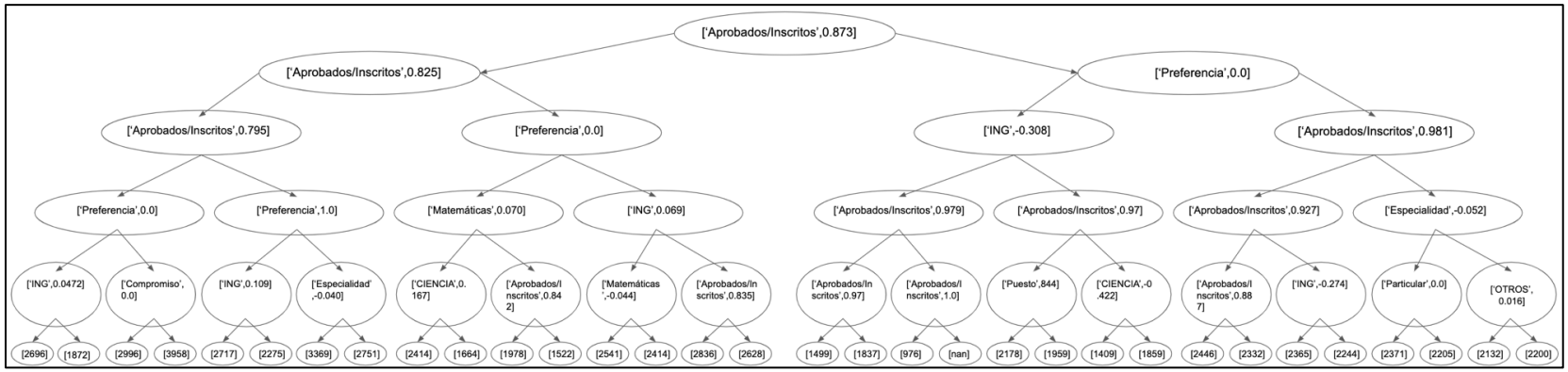


Figura 0-5:Resultado Base de datos completa con éxito académico en función de la duración de la carrera.

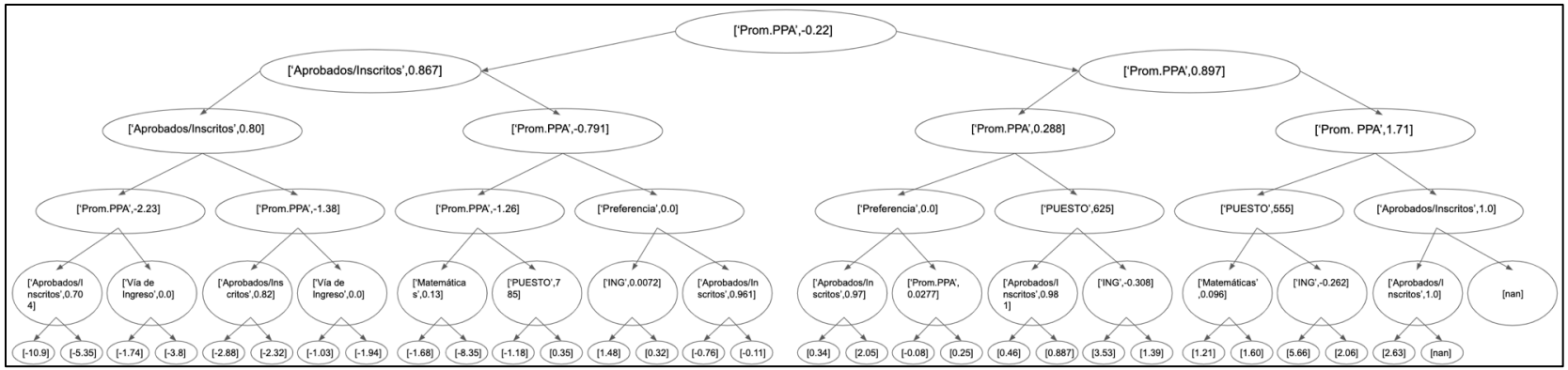


Figura 0-6:Resultado base de datos completa con éxito académico como función de las notas y la duración.

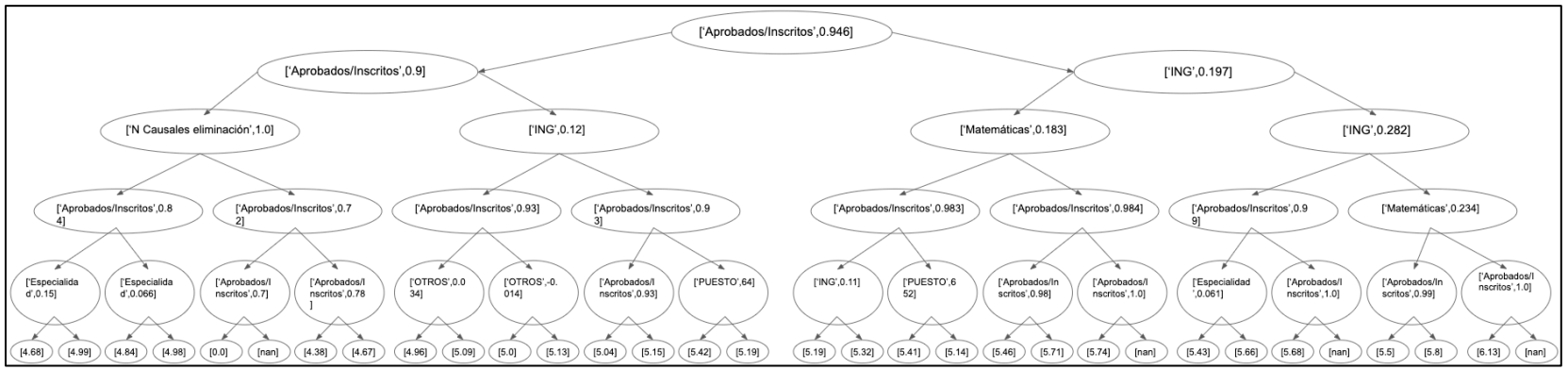


Figura 0-7:Resultado sólo alumnos que realizaron actividad extra programática y éxito académico como función de las notas.

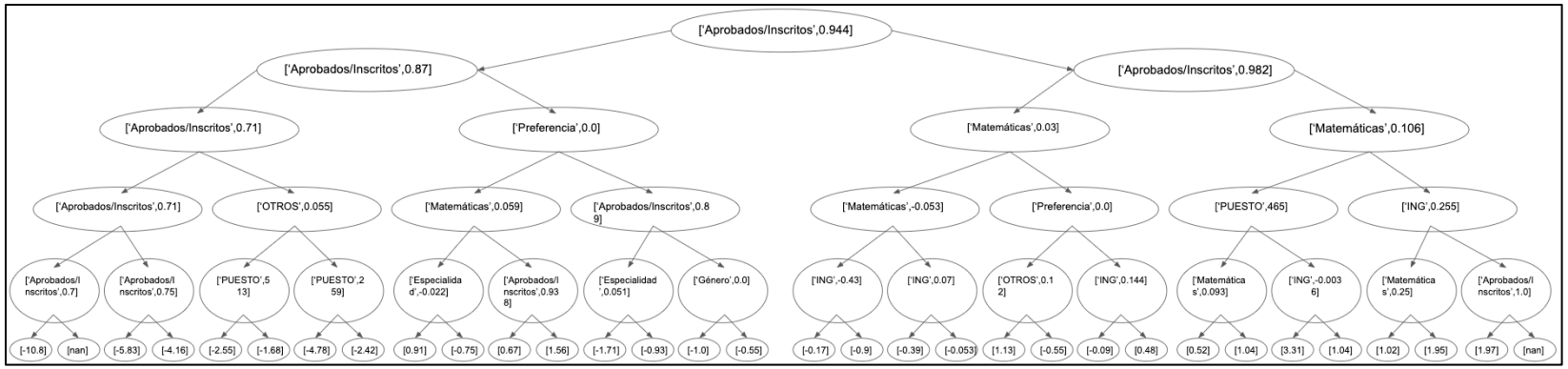


Figura 0-8:Resultado alumnos que sólo realizaron actividad extra programática con éxito académico como función de las notas y la duración de la carrera.

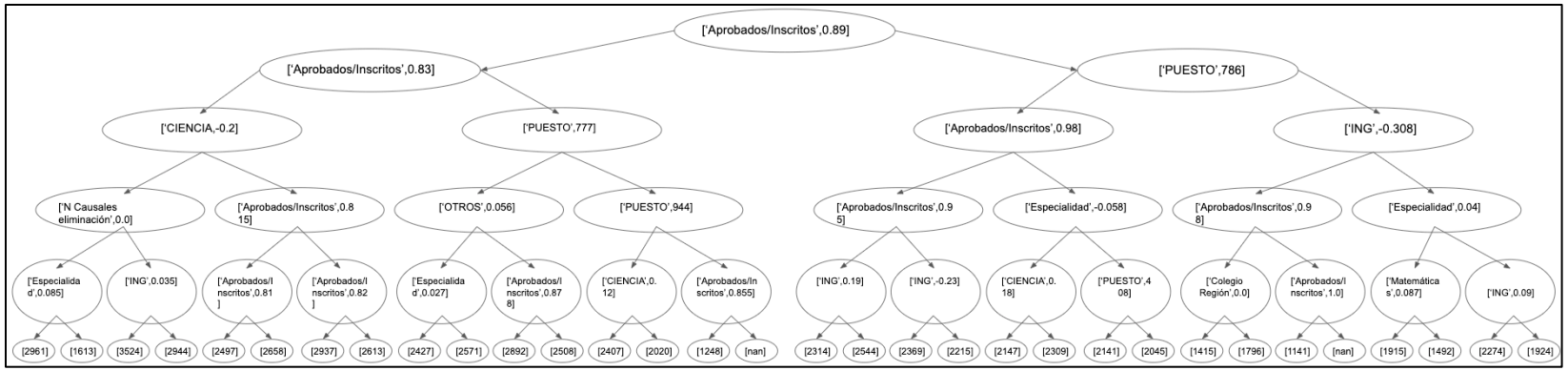


Figura 0-9:Resultado alumnos que sólo realizaron actividad extra programática con éxito académico como función de la duración de la carrera.

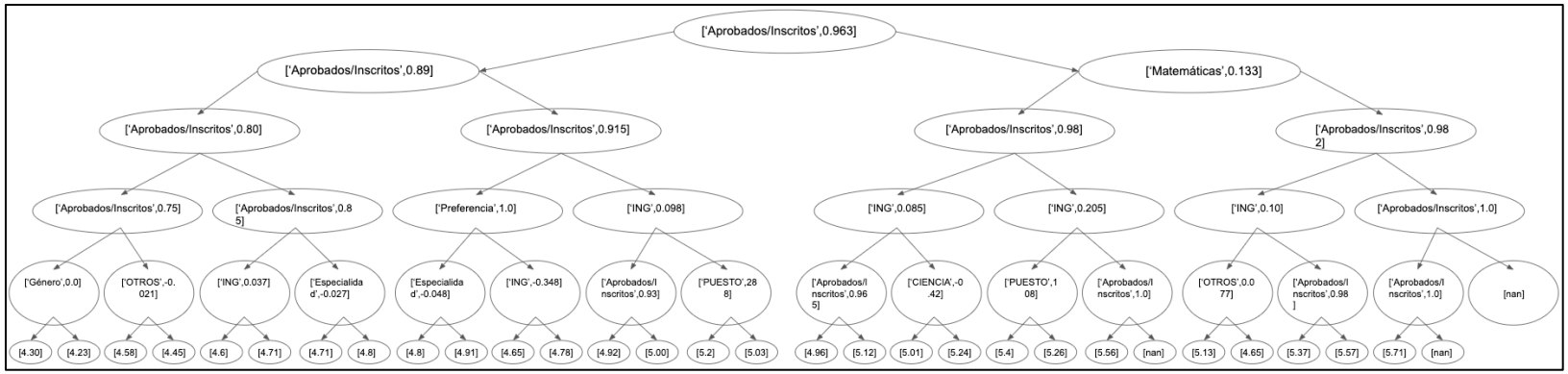


Figura 0-10:Resultado alumnos que no realizaron actividades extra programáticas con éxito académico como función de las notas.

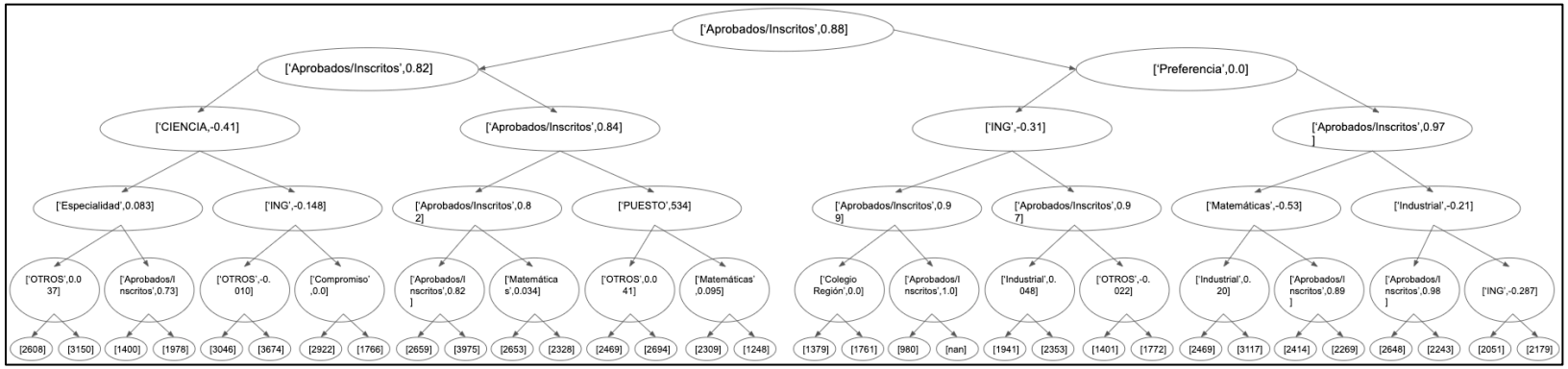


Figura 0-11:Resultado alumnos que no realizaron actividades extra programáticas con éxito académico como función de la duración de la carrera.

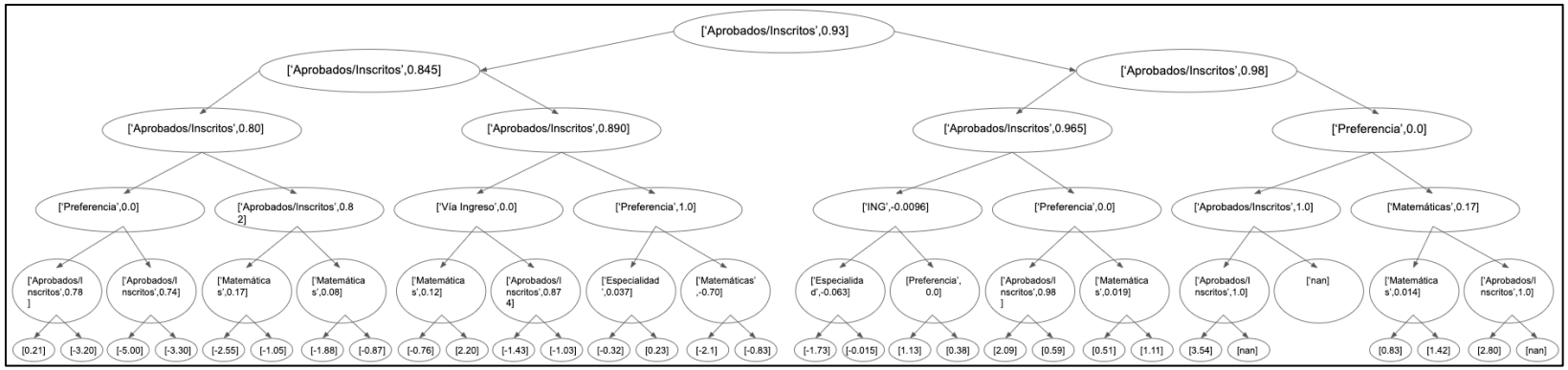
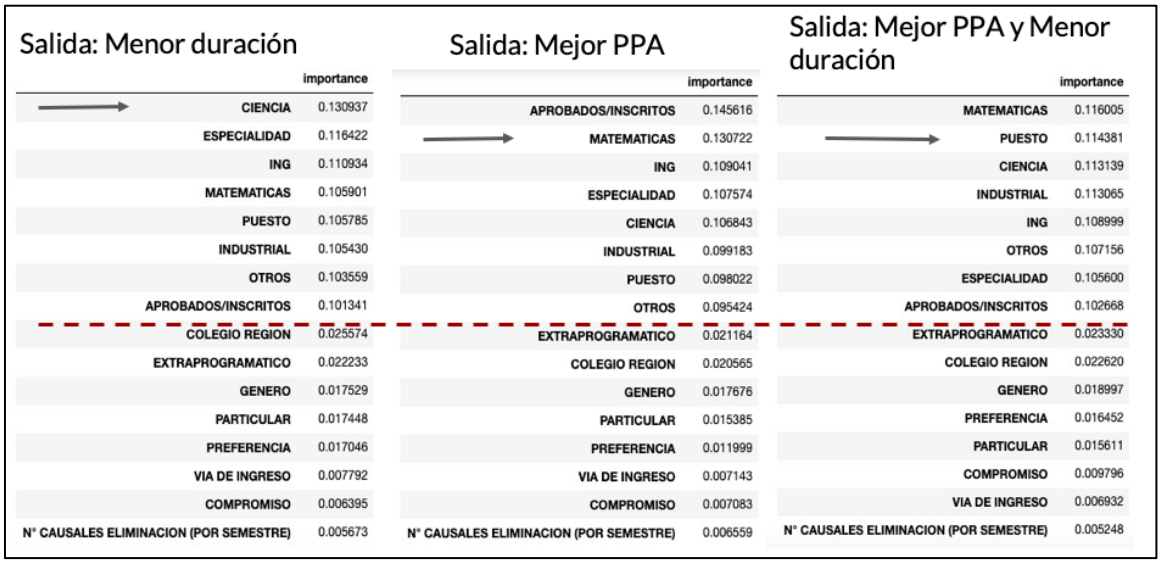


Figura 0-12:Resultado alumnos que no realizaron actividades extra programáticas con éxito académico como función de las notas y la duración de la carrera.

Apéndice D: Resultados Random Forest

A continuación se presentarán los resultados de *"Random Forest"* para las distintas muestras de la base de datos. Los resultados de la muestra que contiene a todos los alumnos se presentan en la figura 9-13. Luego los resultados de la muestra con sólo alumnos que realizaron alguna actividad extra programática en la universidad se presentan en la figura 9-14. Por último en la figura 9-15, se encuentra los resultados de la muestra de alumnos no que no realizaron actividades extra programáticas en la universidad.

En todas las figuras se presentan las variables ordenadas según su importancia relativa. Por lo tanto la suma de todas da 1, lo que indica que cada una de estas está entre 0 y 1. Se destacó con una línea punteada la clara diferencia de importancia entre las variables académicas y las no académicas. Esta diferencia es cercana a un diferencia de hasta 5 veces en grados de importancia relativa.



Salida: Menor duración

	importance
CIENCIA	0.130937
ESPECIALIDAD	0.116422
ING	0.110934
MATEMATICAS	0.105901
PUESTO	0.105785
INDUSTRIAL	0.105430
OTROS	0.103559
APROBADOS/INSCRITOS	0.101341
COLEGIO REGION	0.025574
EXTRAPROGRAMATICO	0.022233
GENERO	0.017529
PARTICULAR	0.017448
PREFERENCIA	0.017046
VIA DE INGRESO	0.007792
COMPROMISO	0.006395
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.005673

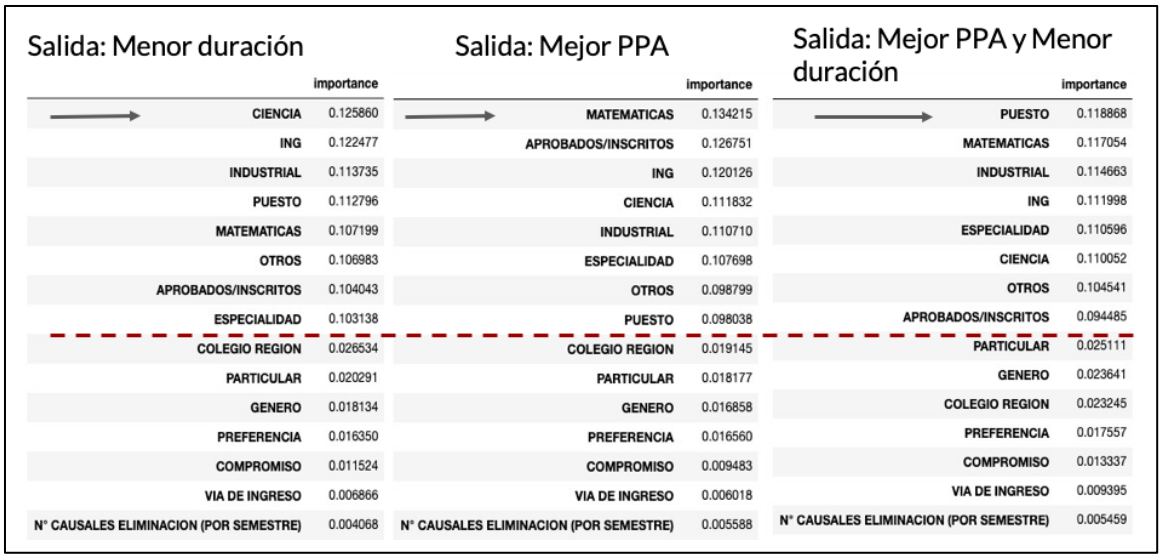
Salida: Mejor PPA

	importance
APROBADOS/INSCRITOS	0.145616
MATEMATICAS	0.130722
ING	0.109041
ESPECIALIDAD	0.107574
CIENCIA	0.106843
INDUSTRIAL	0.099183
PUESTO	0.098022
OTROS	0.095424
EXTRAPROGRAMATICO	0.021164
COLEGIO REGION	0.020565
GENERO	0.017676
PARTICULAR	0.015385
PREFERENCIA	0.011999
VIA DE INGRESO	0.007143
COMPROMISO	0.007083
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.006559

Salida: Mejor PPA y Menor duración

	importance
MATEMATICAS	0.116005
PUESTO	0.114381
CIENCIA	0.113139
INDUSTRIAL	0.113065
ING	0.108999
OTROS	0.107156
ESPECIALIDAD	0.105600
APROBADOS/INSCRITOS	0.102668
EXTRAPROGRAMATICO	0.023330
COLEGIO REGION	0.022620
GENERO	0.018997
PREFERENCIA	0.016452
PARTICULAR	0.015611
COMPROMISO	0.009796
VIA DE INGRESO	0.006932
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.005248

Figura 0-13:Resultados de la muestra completa.



Salida: Menor duración

	importance
CIENCIA	0.125860
ING	0.122477
INDUSTRIAL	0.113735
PUESTO	0.112796
MATEMATICAS	0.107199
OTROS	0.106983
APROBADOS/INSCRITOS	0.104043
ESPECIALIDAD	0.103138
COLEGIO REGION	0.026534
PARTICULAR	0.020291
GENERO	0.018134
PREFERENCIA	0.016350
COMPROMISO	0.011524
VIA DE INGRESO	0.006866
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.004068

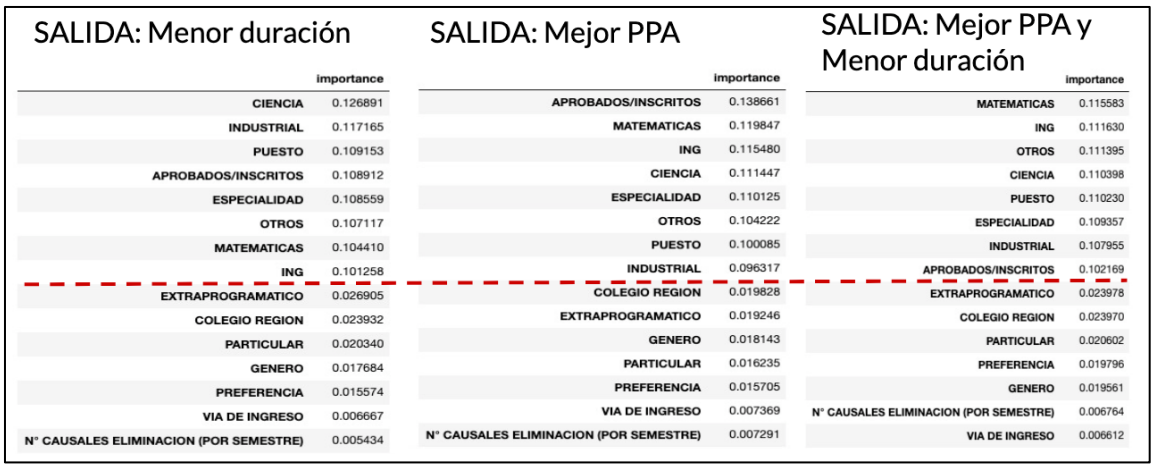
Salida: Mejor PPA

	importance
MATEMATICAS	0.134215
APROBADOS/INSCRITOS	0.126751
ING	0.120126
CIENCIA	0.111832
INDUSTRIAL	0.110710
ESPECIALIDAD	0.107698
OTROS	0.098799
PUESTO	0.098038
COLEGIO REGION	0.019145
PARTICULAR	0.018177
GENERO	0.016858
PREFERENCIA	0.016560
COMPROMISO	0.009483
VIA DE INGRESO	0.006018
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.005588

Salida: Mejor PPA y Menor duración

	importance
PUESTO	0.118868
MATEMATICAS	0.117054
INDUSTRIAL	0.114663
ING	0.111998
ESPECIALIDAD	0.110596
CIENCIA	0.110052
OTROS	0.104541
APROBADOS/INSCRITOS	0.094485
PARTICULAR	0.025111
GENERO	0.023641
COLEGIO REGION	0.023245
PREFERENCIA	0.017557
COMPROMISO	0.013337
VIA DE INGRESO	0.009395
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.005459

Figura 0-14:Resultados de la muestra de alumnos que realizaron actividades extra programáticas.



SALIDA: Menor duración

	importance
CIENCIA	0.126891
INDUSTRIAL	0.117165
PUESTO	0.109153
APROBADOS/INSCRITOS	0.108912
ESPECIALIDAD	0.108559
OTROS	0.107117
MATEMATICAS	0.104410
ING	0.101258
EXTRAPROGRAMATICO	0.026905
COLEGIO REGION	0.023932
PARTICULAR	0.020340
GENERO	0.017684
PREFERENCIA	0.015574
VIA DE INGRESO	0.006667
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.005434

SALIDA: Mejor PPA

	importance
APROBADOS/INSCRITOS	0.138661
MATEMATICAS	0.119847
ING	0.115480
CIENCIA	0.111447
ESPECIALIDAD	0.110125
OTROS	0.104222
PUESTO	0.100085
INDUSTRIAL	0.096317
COLEGIO REGION	0.019828
EXTRAPROGRAMATICO	0.019246
GENERO	0.018143
PARTICULAR	0.016235
PREFERENCIA	0.015705
VIA DE INGRESO	0.007369
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.007291

SALIDA: Mejor PPA y Menor duración

	importance
MATEMATICAS	0.115583
ING	0.111630
OTROS	0.111395
CIENCIA	0.110398
PUESTO	0.110230
ESPECIALIDAD	0.109357
INDUSTRIAL	0.107955
APROBADOS/INSCRITOS	0.102169
EXTRAPROGRAMATICO	0.023978
COLEGIO REGION	0.023970
PARTICULAR	0.020602
PREFERENCIA	0.019796
GENERO	0.019561
N° CAUSALES ELIMINACION (POR SEMESTRE)	0.006764
VIA DE INGRESO	0.006612

Figura 0-15:Resultados de la muestra de alumnos que no realizaron actividades extra programáticas.