POSTGRADO
ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

"MODELO PREDICTIVO PARA LA SELECCIÓN DE POSTULANTES DESTACADOS A UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR"

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CONTROL DE GESTIÓN

Estudiante: Felipe Bugueño

Profesor Guía: Jaime Miranda

Santiago, Mayo 2017

Dedicado a mis abuelos…

It’s a Long Way to the Top (If You Wanna Rock ‘N’ Roll)

AC/DC

AGRADECIMIENTOS

Agradezco profundamente a mis abuelos, en especial a los que ya no están con nosotros, quienes inculcaron en mí las convicciones necesarias para desarrollarme como persona y profesional. Gracias por cada palabra de aliento, historia y consejo que han compartido conmigo.

Fue muy importante en el proceso de desarrollo de esta tesis el apoyo de toda mi familia, mi pareja, mis amigos cercanos y mis jefaturas en el trabajo por la comprensión en los distintos deberes que debí cumplir para llegar a este punto. Agradezco a mi profesor guía, Jaime Miranda, por su compromiso y corrección en cada etapa del arduo camino de esta tesis de investigación. No sólo me guió a través del desarrollo metodológico si no también fue un guía de desarrollo profesional y académico.

Para terminar agradezco a los diferentes funcionarios de la Facultad de Economía y Negocios de la Universidad de Chile, entre ellos al Director de Escuela de Sistemas de Información y Auditoría, Freddy Coronado, al ex Director de Escuela de Sistemas de Información y Auditoría, Ariel La Paz, al Secretario de Estudios, César Ortega, a la Directora de NexoColegios, Constanza Acuña, y al Jefe de la Unidad de Sistemas de Infotecnología, Rodrigo López, quienes dedicaron su tiempo a la facilitación de información y comprensión del problema resuelto en esta tesis.

Índice de Contenido

Capítulo 1 – Introducción

1.1 Resumen Ejecutivo

La tesis que se presenta a continuación se sitúa en el contexto del proceso de selección universitaria en Chile, en particular en el momento en que las universidades compiten por atraer a los alumnos con mejores resultados. En nuestro país ocurrieron cambios en la legislación como la Ley de Financiamiento de la Educación, el cambio en el cálculo del puntaje de notas de enseñanza media (NEM), un aumento sostenido del ingreso universitario por estudiantes secundarios (González, 2002), además del fenómeno de la deserción universitaria que se mantiene tanto en el mundo (Haselgrove, 1994; Levitz, 2000; Yorke, 1997) como en nuestro país alrededor del 22% (Centro de Estudios MINEDUC, 2012). Estos hechos han generado un alto nivel de competencia por los alumnos con buenos puntajes en su postulación. Por lo anterior, las herramientas de selección de estudiantes secundarios toman un rol importante. Para seleccionar estudiantes secundarios se buscan aquellos que puedan entregar un buen desempeño universitario a futuro, es decir que el principal desafío entonces es identificar a estos estudiantes secundarios que potencialmente serán buenos estudiantes universitarios. La literatura provee diversos estudios sobre las variables que influencian en el desempeño académico, de estos estudios las primeras variables utilizadas fueron el coeficiente intelectual (Terman, 1916), luego otras variables como la personalidad y contexto (Freeman, 1970), y también las notas secundarias (Rubin, 1977). Aparecen posteriormente teorías sobre la inteligencia emocional (Goleman, 1998) y como las relaciones interpersonales tienen un efecto en el desempeño (Mayer, 1990). De estas teorías se generan validaciones empíricas con correlaciones altas entre desempeño y características personales de los estudiantes (Chickering, 1991; Parker, 2004; Mestre, 2006). También existen estudios sobre el contexto social (House, 2000), las expectativas del alumno (Thomas, 2002) y el nivel de adaptación al medio (Lowis, 2008). Así como también la importancia de la preocupación de los padres (Shurkin, 1992; Christenson, 2010; Al-Alwan, 2014). Los resultados en educación han sido estudiados desde muchos enfoques. Uno de estos enfoque es desde la Minería de Datos donde se encuentran bastos estudios que predicen determinados resultados en situaciones relacionadas con la educación (Luan, 2002; Romero, 2007; Peña, 2014). Entre ellos se encuentran la aceptación de candidatos (Acikkar, 2009; Padmapriya, 2012), la predicción de tasas de graduación (Nadeshwar, 2011) y por supuesto la clasificación de desempeño (Pitman, 2008; Kabakchieva, 2013). El objetivo de esta tesis es desarrollar un modelo predictivo utilizando Minería de Datos, para determinar el perfil de los postulantes destacados para los estudiantes secundarios que postulan a la Facultad de Economía y Negocios de la Universidad de Chile. La metodología utilizada corresponde al

Knowledge Discovery in Databases – también conocida por sus siglas como KDD, planteada por Fayyad, Piatetsky-Shapiro y Smyth (1996). Se utilizan cuatro enfoques de solución con modelos de predicción con diferente nivel de complejidad, dentro de ellos se utilizan técnicas de selección de variables, el uso de cluster y técnicas de combinación de modelos. Los resultados encontrados por esta tesis tienen implicaciones en las variables utilizadas para identificar buenos alumnos. Los modelos de predicción alcanzan hasta un 80,4% de predicción, con un ratio de identificación de los postulantes destacados de hasta un 100% a un 75%. El mejor modelo es llamado en esta tesis como un enfoque Secuencial Combinado, y en particular utiliza técnicas de selección de variables y técnicas de combinación para obtener una predicción de un 80,4% con un ratio de predicción en postulantes destacados de un 89,4%. En cuanto a las variables que destacan por su influencia en la predicción son las notas del colegio, el sexo femenino, la edad de 19 años o menos, el puntaje de lenguaje y el puntaje NEM, así como también si el alumno trabaja, si egresó del colegio el mismo año que dio la Prueba de Selección Universitaria, si utiliza un crédito para financiar sus estudios y las diferencias de puntajes con el promedio de su colegio. Las implicancias de estos resultados cuestionan la actual importancia de ciertas variables de selección, de ellas la más importante es el puntaje de matemática pues es el foco para buscar a los mejores postulantes secundarios. Los resultados demuestran que el puntaje de matemática sólo debe ser lo suficientemente alto para ingresar, y luego de ello otras variables como el puntaje de lenguaje cobran relevancia. Los resultados demuestran que es posible realizar una predicción eficaz utilizando la metodología del KKD, así como también integrar este modelo predictivo al actual proceso de selección de la Facultad de Economía y Negocios para apoyar y mejorar el actual proceso de atracción de postulantes secundarios a la universidad. A su vez, se recomiendan nuevas variables para la selección de postulantes destacados y la atracción de ellos. Es entonces importante destacar que en nuestro país, las pruebas de selección proveen poca información sobre los postulantes, ya que se observan muchos estudios con resultados importantes cuando se utilizan variables adicionales en la selección, tales como pruebas de pre – ubicación universitaria, pruebas de talento o psicológicas. Se debe continuar profundizando en los mejores mecanismos para seleccionar estudiantes como país y su vez continuar el desarrollo de herramientas por las universidades para atraer a los mejores estudiantes.

1.2 Objetivos de Investigación

En esta tesis se realizará un estudio aplicado sobre la selección de estudiantes secundarios en el contexto universitario de nuestro país. Los objetivos generales y particulares de esta tesis son:

1. **Objetivo general:** Desarrollar un modelo predictivo que determina el perfil de los postulantes destacados para los estudiantes secundarios que postulan a la Facultad de Economía y Negocios de la Universidad de Chile.
2. **Objetivos específicos:**
   - **a.** Definir un perfil de selección de postulantes destacados.
   - **b.** Seleccionar datos y variables que sean explicativas para el perfil de un postulante destacado.
   - **c.** Procesar los datos y variables para que tengan coherencia, completitud y formato apropiado.
   - **d.** Transformar los datos para que sean útiles según los requerimientos de cada modelo predictivo.
   - **e.** Aplicar técnicas de minería de datos con el fin de predecir el perfil de selección de los postulantes destacados.
   - **f.** Evaluar los resultados de la predicción y su aplicación en el proceso de selección.

1.3 Contexto y motivación

En el contexto actual de nuestro país se ha desarrollado un fuerte interés por el rol de las universidades y el aporte a la sociedad que cada institución de educación superior entrega. Las universidades ya sean públicas o privadas compiten por los estudiantes que postulan a ellas. De hecho, hasta principios del año 2016 la competencia era tan ardua que los estudiantes con mejores resultados eran contactados inmediatamente después de la publicación de resultados. En el momento del llamado de las universidades, estos estudiantes recibían invitaciones a postular a determinada carrera, así como también ofrecimientos de apoyo económico directo en becas y otros mecanismos de financiamiento para incentivar a estos estudiantes de puntajes destacados a seleccionar determinada casa de estudios. Esto se evidencia en varias notas de prensa como en El Mercurio, artículo de nombre "Universidades compiten por mejores puntajes: ofrecen hasta carreras gratuitas" (2009). El interés de las universidades está basado en que los buenos estudiantes podrían generarles mejores resultados académicos, sujeto a la misión educacional que se plantea cada institución. En nuestro país se realiza un estudio en detalle y clasificación de este efecto (Améstica, 2014). Entre muchos otros, la autora lo cataloga como una ventaja de inicio por algunas universidades. Esta ventaja competitiva se origina en las universidades que tienen mayor capacidad de captación de estudiantes vía incentivos y llamados atractivos. Ahora bien, debido al contexto país y diversas demandas sociales, se implementó una nueva Ley de Educación Superior. La Ley de Calidad y Gratuidad en la Educación Superior trae consigo cambios importantes que exacerban el fenómeno de competencia entre las universidades.

Los principales cambios que ocurrieron en nuestro país tienen que ver con el financiamiento y la medición utilizada para en el proceso de ingreso a la universidad. En cuanto al financiamiento de los estudiantes que ingresaron a la universidad en el año 2016, la ley de educación plantea apoyo económico directo a las universidades por cada estudiante que se matriculara. Existen entonces condiciones para ello, algunas socioeconómicas tales como que los estudiantes se encuentren entre los deciles de ingreso económico familiar primero y sexto. También condiciones de desempeño académico, como tasa de ramos aprobados. Para estos estudiantes el Ministerio de Educación financiaría la carrera universitaria, además de introducir otros cambios importantes en el proceso de selección universitaria, los que son descritos en la Ley de Calidad y Gratuidad en la educación superior.

En cuanto a los cambios en el proceso de selección de las universidades, estas reciben a los estudiantes después de rendir exámenes globales, Prueba de Selección Universitaria (PSU), que evalúan las aptitudes y asignan un puntaje de acuerdo a la cantidad de respuestas correctas e incorrectas. Cada universidad pondera diferente cada uno de estos puntajes por prueba. Al

conjunto de pruebas PSU, se adhiere un puntaje adicional que se denomina Notas de Enseñanza Media (NEM). El NEM representa el desempeño escolar del estudiante a lo largo de sus años de educación secundaria a través del promedio de notas de todos los años cursados.

Hasta el año 2015 el puntaje NEM dependía de una tabla de conversión de nota promedio a puntaje entre 150 y 850 puntos. La tabla de conversión consistía en transformar directamente el promedio general del estudiante en su enseñanza media a un puntaje fijo. Dentro de los cambios de la ley de educación se cambia el mecanismo de asignación de puntaje NEM y se comienza a utilizar un ranking en orden descendente en que la mayor nota del colegio obtiene el máximo puntaje.

Los cambios descritos anteriormente se tradujeron en parte, en tres desafíos para atraer estudiantes destacados hacia determinada universidad. Primero, altera directamente la métrica de cálculo de la variable puntaje NEM y por tanto su significado, pues en asignación por ranking ahora el mejor estudiante de cada colegio alcanzaría el mejor puntaje NEM. Es decir, exagera el puntaje de estudiantes, los que algunas veces no necesariamente por una nota alta tienen un puntaje NEM mayor. Siendo entonces, en algunos casos, el puntaje NEM una señalización inflada de buen desempeño.

El segundo desafío es que ahora, los estudiantes con puntaje suficiente para postular a las diferentes universidades líderes del país, y que se encuentra entre el primer y sexto decil socioeconómico, encuentran respaldo económico en el Ministerio de Educación. Por ende, no son igual de atractivos los anteriores incentivos ofrecidos hasta el año 2015 por las universidades relativos al arancel. Es decir se agudiza la competencia por estos estudiantes entre las universidades que reciban pagos de arancel desde el Ministerio de Educación, por ende el trabajo de los equipos de captación universitaria cambia y se agudiza a su vez en cuanto a atraer a los estudiantes secundarios. Por ejemplo ya no se les puede atraer con una beca de arancel pero tal vez si otros beneficios. Lo anterior, ya sea en una labor de apoyo al proceso formativo académico en el mediano y largo plazo o bien en el momento de entrega de resultados de la PSU.

Un tercer desafío que se mantiene presente en las universidades tiene que ver con la reducción de la deserción. La deserción es entendida como la pronta salida de un estudiante antes del término del plan de estudio (Bean, 1980) ya sea por decisión propia o por bajo desempeño académico. Esto último es entendido como la aprobación de los ramos contenidos en el plan de estudios de la carrera a la que el estudiante se matricula. En cuanto a estudios sobre deserción internacionales, en Estados Unidos los primeros años de abandono de los estudiantes en el sector universitario puede llegar al 20% (Haselgrove, 1994; Levitz, 1989,2000; Yorke, 1997) o superior. Mientras que en Chile según el Centro de Estudios del Mineduc en el año 2010, las

instituciones de educación superior alcanzan una deserción de 22%, mientras que los institutos profesionales y de formación técnica bordean el 36% (Centro de Estudios MINEDUC, 2012). Siendo una clave para reducir esto la captación de mejores estudiantes al comienzo de la carrera. Desde la disciplina econométrica, en nuestro país se proyecta que para el año 2022 cerca del 45% de los estudiantes secundarios que completen sus estudios alcanzaran la educación superior en nuestro país (González, 2002). Al momento del estudio la participación universitaria de los estudiantes secundarios era de 25% con gastos por deserción sobre los $47.000 MM (González, 2002). Mientras que para algunos casos según un estudio (Miranda & Vásquez, 2015) basado en la minería de datos, el 97% de la deserción se concentra en los 3 primeros años. En cuanto a estudios de deserción universitaria (Tinto &Cullen, 1975; Bean, 1980) muestran que muchos factores influencian el mal desempeño académico, siendo uno importante la brecha en la base de formación deficiente que traen los estudiantes en algunos casos.

Dado los desafíos anteriores sobre cambios en puntaje NEM, aumento de competencia en la captación universitaria y la reducción de la deserción se acrecienta la necesidad de identificar prontamente a los estudiantes secundarios candidatos a la universidad .La identificación de estudiantes y posterior invitación a postular permite a las universidades conseguir los mejores postulantes (Améstica, 2014) traduciéndose en una ventaja competitiva.

Esta necesidad de las casas de estudios tiene características particulares. La primera característica es que el momento crítico de comunicación con los postulantes es entre la entrega de resultados de la PSU y la postulación. Esto es un periodo muy breve de entre 3 y 5 días corridos. Por lo tanto la selección de a quien invitar a la universidad debe ser rápida. La segunda característica es que la selección en si misma depende de los objetivos de cada casa de estudios, por ejemplo atraer determinado perfil de estudiante. Por lo tanto las invitaciones por llamados y otros medios deben ser dirigidas acorde a los objetivos del equipo de captación universitaria. La tercera característica es que la orientación del perfil de estudiante anterior debe ser identificada con la información de los estudiantes disponibles en ese determinado momento. Esta información es provista por el DEMRE para todas las universidades. Tomando en cuenta estos tres elementos, vemos que las necesidades de captación de los estudiantes previa postulación requieren de un proceso rápido de identificación de perfiles en un ambiente de información en particular.

En cuanto a la definición del perfil de estudiantes de educación superior se presentan estudios de diversas ramas. Las características e información de los estudiantes utilizada para clasificarlos se ha complejizado a lo largo del tiempo, en un comienzo era el foco estaba en puntajes de evaluación en relación a la lógica matemática y mediciones del IQ como se ve en Terman, Lewis, Oden y Melita (1947), luego aparecen estudios del contexto de los estudiantes y

sus padres en el estudio de Shurkin y Brown (1992). Posteriormente aparecen estudios de clasificación sobre comportamiento y manejo de las emociones en los estudiantes y sus efectos en el desempeño (Payne, 1985; Meyer, 1990; Salovey, 1990; Goleman, 1998). Finalmente aparecen estudios que clasifican en forma integral el desempeño de los estudiantes considerando resultados de exámenes lógicos matemáticos, contexto del estudiante y comportamientos (Mestre, 2006; Parker, 2004) evidenciados por el estudiante.

Como explican los autores del párrafo anterior, si el nivel de formación y otras características de los estudiantes secundarios pueden ser identificados en el proceso de captación, es entonces posible utilizar estas características para dirigir los esfuerzos de captación a los estudiantes secundarios según un objetivo en particular, por ejemplo apuntar a estudiantes con mayor probabilidad de buen desempeño académico y menor probabilidad de deserción. Cuando hablamos de un perfil, es posible entonces buscar a los estudiantes que puedan desenvolverse de mejor forma en la facultad en sus diferentes ámbitos, ya sea académico, profesional y extra-curricular. Por lo tanto, son estudiantes que como el principal input del proceso formativo, sean capaces de facilitar este mismo proceso con sus características o potencial de desarrollarlas. Entre las principales variables sobre desempeño de estudiantes en la educación superior destacan las notas de los ramos cursados puesto que definen qué estudiante adquiere o no los conocimientos suficiente para luego ser evaluado en el examen de grado final conducente al título profesional.

Entenderemos que el perfil anteriormente mencionado dependerá de los objetivos particulares de cada universidad. Esto es acorde a que el proceso de formación después de ingresar a la universidad varía según cada casa de estudios. Los diferentes procesos de formación consideran comúnmente una nivelación inicial para sus estudiantes que incluye asignaturas preparatorias en verano, luego asignaturas globales de pensamiento lógico y matemático, incluyendo materias tal cómo cálculo, algebra, lógica, semántica, comunicación y teorías según la carrera en general. Es decir que deben ser considerados las características del proceso formativo a la hora de plantear una clasificación del perfil de los estudiantes secundarios que postulan a la universidad. Cabe destacar entonces que la selección y deserción están ligadas con el proceso de educación formativo.

Estudios importantes en relación a clasificación según desempeño universitario se explican utilizando variables de evaluaciones lógicas, comportamientos y contexto de los estudiantes. Vemos estudios sobre identificar el buen o bajo rendimiento de un estudiante y su posible abandono (Lowis, 2008), adaptación a la educación superior y características de estos estudiantes (Thomas, 2002) e inclusive estudios sobre el contrato psicológico entre profesor y estudiantes superiores y los principios de buena enseñanza (Chickering, 1991) que se aplican de

cara a mejorar desempeño. Luego, si existe entonces la información sobre las características del estudiante al momento de ejecutar la captación de estudiantes es posible clasificar al estudiante en base a estas variables. Desafíos y problemas complejos como identificar a un tipo de estudiante con antelación, generan la oportunidad para que disciplinas aplicadas respondan a mejorar la posición competitiva en cuanto a captación de estudiantes.

La metodología de estudio a aplicar en este trabajo será el Knowledge Discovery in Database (Fayyad, 1990) desde una base de datos sobre los estudiantes mediante la minería de datos. Para esta tesis nos concentraremos en un modelo de minería de datos para apoyar el proceso de selección de estudiantes secundarios dirigido a los futuros estudiantes universitarios, utilizando el contexto, información de estudiantes y lineamientos de la casa de estudios, además de apoyar desde la perspectiva del modelo de clasificación de estudiantes, y el entendimiento de los cambios introducidos por la reforma educacional.

La minería de datos permite generar modelos de predicción capaces de determinar si un estudiante secundario que postulanta a la educación superior será un postulante destacado o no. Para obtener el resultado anterior es necesario considerar tanto los desempeños académicos universitarios de los estudiantes como sus desempeños e información de educación secundaria cuando, hace años atrás, postularon a la universidad. El desempeño académico universitario es utilizado para generar el perfil de un estudiante, como destacado o no. El proceso de minería de datos permite aprender de la información de educación secundaria de los estudiantes con un perfil universitario destacado, para luego utilizar los resultados de la minería de datos para predecir si un nuevo postulante a la educación universitaria es un postulante destacado o no. Por lo tanto, al elaborar un modelo de clasificación podríamos obtener resultados valiosos para dirigir las políticas de captación de estudiantes hacia los objetivos particulares de la casa de estudios. Siendo esta aplicación una propuesta con un impacto económico según explica Laudon y Laudon (2010) en cuanto a reducir los costos de transacción (Coase, 1937; Williamson, 1985) y de reducción en la asimetría de información (Akerlof, 1970) entre la casa de estudios y los estudiantes que postulan a ella. En el contexto de estudio de esta tesis estos costos corresponden a los costos de participación de mercado como localizar y comunicarse con futuros estudiantes. Sin embargo la aplicación de este modelo predictivo, debe cumplir con las características descritas, sobre rapidez, disponibilidad de la información y orientación hacia los objetivos de captación.

Esta tesis tiene la siguiente estructura. En el Capítulo 1 revisado en los párrafos anteriores, se revisó la introducción que describe la motivación del problema de esta tesis. La estructura de este documento es como sigue: En el Capítulo 2 que prosigue se describe el problema a resolver en forma práctica, puesto que resulta importante en cuanto al fenómeno temporal y la aplicación

de esta tesis. En la sección del Capítulo 3, se realiza una revisión de la literatura sobre desempeño y predicción del desempeño de estudiantes universitarios, luego de modelado de perfiles y minería de datos en educación. Posteriormente el Capítulo 4 se describe la revisión de la metodología aplicada en este trabajo y las técnicas de minería de datos. Además en el Capítulo 5 se describe en detalle el enfoque de solución de esta tesis puesto que se incluye la creación de un perfil y tres enfoques de solución de modelos predictivos. En el Capítulo 6 se presenta el caso de estudio en la Facultad de Economía y Negocios de la Universidad de Chile entre los años 2004-2010 y la caracterización del perfil generado para que luego en el Capítulo 7 se detalle el experimento realizado en software. Continuando con el Capítulo 8 que describe los principales resultados de los experimentos en base a los indicadores de evaluación y las principales variables explicativas. Para finalmente en el Capítulo 9 se realiza una revisión de las conclusiones y discusiones que aporta esta tesis con respecto a la selección de estudiantes, el modelo de predicción generado y sus resultados.

Capítulo 2 – Descripción del Problema

El problema a resolver por esta tesis consiste en seleccionar postulantes destacados a FEN de acuerdo a un perfil asociado al desempeño académico al final de la carrera universitaria. Es decir, se predice el perfil de desempeño académico al momento de su postulación universitaria.

Como fue discutido en el Capítulo 1, el proceso de selección de las universidades adscritas al proceso por selección PSU utiliza postulaciones en orden de preferencia. La postulación es en base a un puntaje ponderado de los resultados de las pruebas del estudiante. Esto quiere decir, que en realidad las universidades no pueden escoger a un estudiante en particular, sino que es por un proceso de llenado de cupos que inicia con el máximo puntaje ponderado y termina cuando el cupo es llenado. Pesé a no poder seleccionar en particular un estudiante secundario, la competencia de las universidades por seleccionar a los mejores estudiantes secundarios es cada vez mayor. Estas compiten por ser atractivas para los estudiantes secundarios en dos etapas, la primera es antes del proceso de selección y la segunda etapa ocurre justo en el proceso de selección. La diferencia es que en el primero no se conocen los resultados de la PSU, mientras que en el segundo son conocidos. En las etapas previas al proceso de selección se busca atraer estudiantes secundarios realizando actividades variadas tales como visitas a colegios, realización de ferias, coberturas de prensa, difusión en medios de comunicación y otros medios tradicionales de publicidad. Estos son costos de participación de mercado para localizar y comunicarse con futuros estudiantes. La segunda etapa ocurre en el proceso de selección, después de conocer los resultados de las pruebas PSU pero antes de postular a las universidades. Este es un breve periodo de tres a cinco días, en que las universidades contactan a los estudiantes con resultados atractivos en sus pruebas PSU. A los estudiantes contactados se les ofrecen beneficios y becas de arancel en caso de postular en primeras prioridades a determinada carrera y ser seleccionado. En esta segunda etapa la cantidad de estudiantes llamados es limitada por el breve tiempo y la falta de herramientas para identificar buenos candidatos. El proceso de selección vía PSU que utiliza FEN se muestra en la Figura Nº 1.

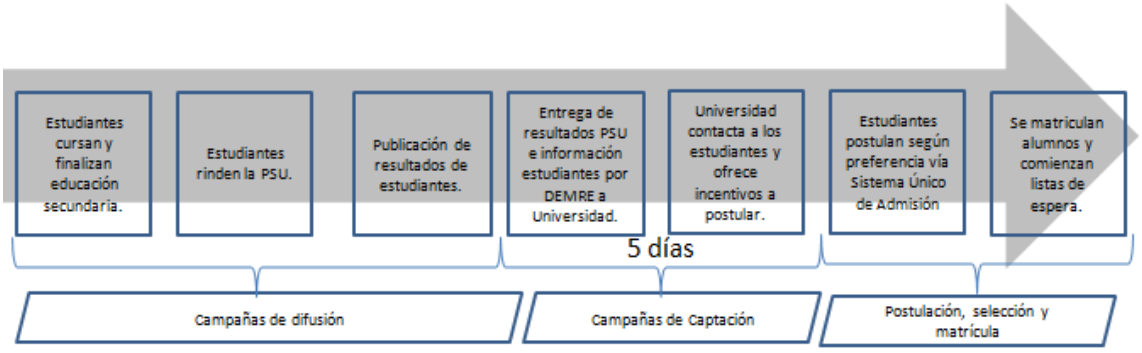


Figura Nº1. Proceso de selección de postulantes y campañas de difusión.

Por lo tanto, el problema que describe esta tesis guarda relación con la segunda etapa de selección de estudiantes. Esto quiere decir, que el problema principal en la selección es identificar a un postulante destacado para ser contactado e invitarlo a postular a una carrera en particular en menos de cinco días. Pero para identificarlo mediante un modelo predictivo, se debe cumplir con las características de rapidez, disponibilidad de la información y orientación hacia los objetivos de captación.

Por su parte, el problema de decisión si contactar a un postulante para ofrecerle beneficios o no hacerlo depende del indicador utilizado para determinar que postulante es atractivo para la casa de estudio. De ahora en adelante se refiere esta tesis como perfil al indicador de cuan atractivo es el postulante para una determinada universidad. Actualmente las universidades utilizan como principal indicador para asignar un perfil a los postulantes los mismos resultados de la PSU. Es decir, contactar primero a los estudiantes con el mayor puntaje en las pruebas. Pero se evidencia que las pruebas de selección por si solas no son un indicador claro y preciso sobre el desempeño académico universitario futuro (Freeman, 1970; Goldman, 1976; Rubin, 1977). Existen entonces, otras variables que sumadas a las pruebas de selección permiten una mejor predicción del desempeño académico universitario (Eskew, 1988; Chickering, 1991; Lowis, 2008). Por lo tanto el principal problema es descubrir este perfil de Postulante Destacado en la breve etapa de postulación, para hacerlo se describe la aplicación del modelo predictivo generado en esta tesis en la Figura Nº 2.

Figura Nº 2. Proceso de selección de postulantes basado en modelos predictivos.

2.1 Construcción del perfil de selección

Esta tesis centrará sus esfuerzos en identificar los estudiantes que ingresan vía PSU, ya que son estos los estudiantes que postulan utilizando las pruebas de selección y los estudiantes contenidos en la base de datos que envía DEMRE a todas las universidades al momento de postulación de los estudiantes. Además la naturaleza de este ingreso no es comparable con las otras, pues los otros tipos de ingresos responden a mandatos legales particulares. Además el foco de atención son los estudiantes a captar por el proceso de búsqueda de los postulantes, por lo que se requiere que estos postulantes puedan llegar al estado de titulado, egresado o regular hasta el quinto año.

En la búsqueda de un indicador más adecuado para seleccionar postulantes e invitarlos a tomar determinada universidad en primera prioridad, esta tesis se enfoca en el desempeño académico universitario futuro. Esto quiere decir que se contactarían a los postulantes secundarios que tengan mayor probabilidad de alcanzar un desempeño académico alto. Para predecir este desempeño académico en nuevos candidatos se utiliza la información de candidatos anteriores. Estos candidatos anteriores, fueron seleccionados vía PSU en el pasado convirtiéndose en estudiantes de FEN y demostraron determinado desempeño académico con el paso de los años de carrera. Esta tesis centra su atención en el desempeño al final de la carrera universitaria. Por lo tanto, si una carrera tiene cinco años de duración el modelo predictivo generado utiliza datos de al menos cinco años hacia atrás. Prediciendo la calidad de un nuevo postulante, en base a los resultados académicos de postulantes de al menos cinco años atrás.

La Figura Nº 3 muestra la lógica de construcción de este perfil en base a los diferentes etapas y grupos de datos, es importante destacar que el perfil es creado al obtener un resultado académico en el Periodo n+5 siendo n el periodo de postulación anterior. La generación del modelo predictivo se realiza utilizando los datos de postulación del Periodo n más el perfil del desempeño académico del Periodo n+5. Luego de entrenado y generado el modelo de predicción se aplica a los nuevos estudiantes en el Periodo n+6 que quieren ingresar a la universidad. La predicción los marca como Postulantes Destacados, a quienes se predice podrán obtener un desempeño académico alto o Postulantes no Destacados a los estudiantes secundarios que no tienen mayor probabilidad de obtener un desempeño destacado.

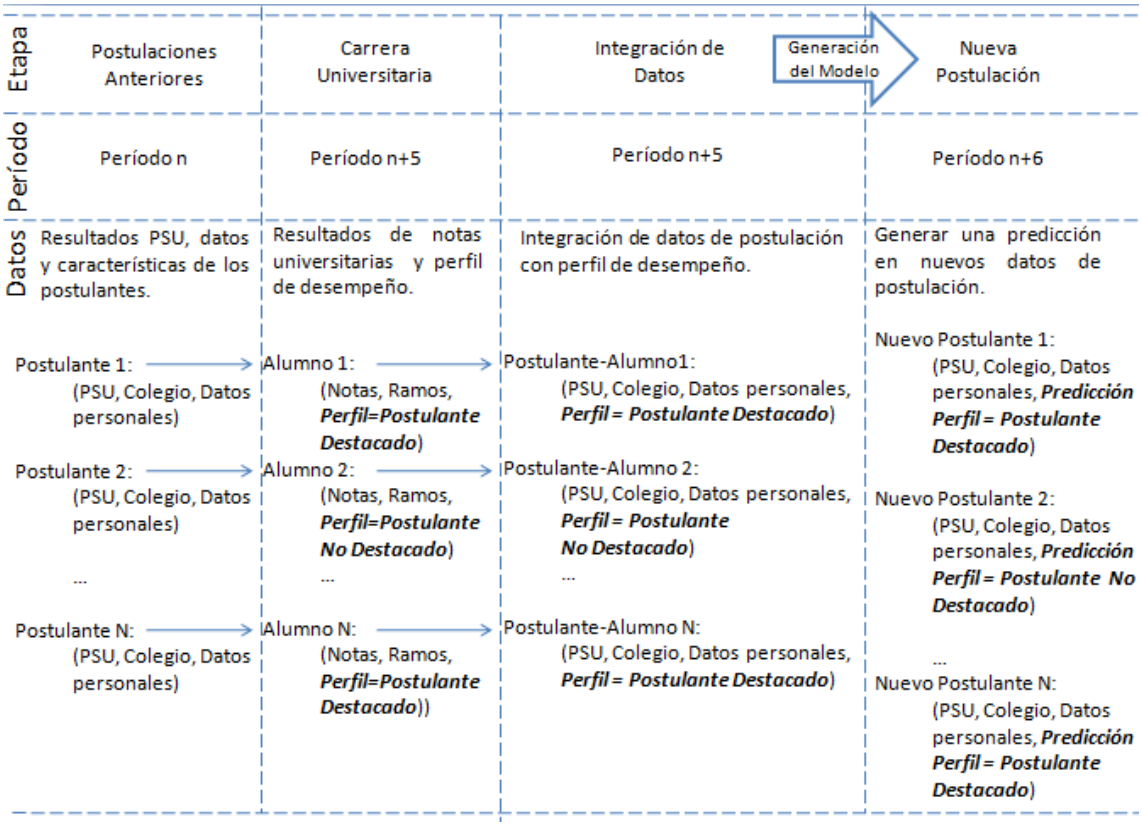


Figura Nº 3: Construcción del perfil de selección

Para la generación de este perfil se utilizan las directrices de la dirección de escuela que indicó que es necesario evitar el sesgo que se produce en las notas de los estudiantes al principio de su carrera, al primer y segundo año. Este sesgo ocurre debido a la adaptación de los estudiantes al medio universitario. Este sesgo tiene que ver con las diferencias de los estudiantes para adaptarse al medio y los diferentes trasfondos académicos además de las diferentes exigencias con las que cada colegio prepara a sus estudiantes.

Por lo tanto el perfil de selección generado para esta tesis corresponde a los estudiantes que al quinto año de carrera obtienen el tipo de Egresado o Titulado con nota mayor a 5,5 en el promedio ponderado por unidades docentes de sus tres últimos años. Es decir, no se considera el promedio de primer y segundo año

Capítulo 3 – Revisión de la Literatura

En este capítulo se realizara una revisión bibliográfica de las investigaciones relacionadas con la predicción del desempeño académico en candidatos de estudiantes secundarios para la educación superior.

En la primera sección se revisarán los modelos de predicción de los estudiantes universitarios en relación a su desempeño. En la segunda sección se revisan las variables y características de estudiantes más importantes que influyen en el desempeño académico. Mientras que en la última parte se revisarán los trabajos de minería de datos y modelos de clasificación que se han aplicado en educación.

3.1 Predicción de desempeño académico

El desempeño académico de los estudiantes universitarios se entiende como el resultado medible de cursar el plan de estudios de una carrera universitaria. Ahora bien la medición de desempeño académico se encuentra asociada a una escala numérica. Los trabajos de investigación realizados en predicción de estudiantes apuntan a dos grandes áreas, las que son identificar a los estudiantes con bajo desempeño que pueden desertar o bien, identificar y potenciar a los estudiantes que más destacan por su desempeño.

El concepto de predecir desempeño toma fuerzas a comienzos del siglo XX donde la predicción de resultados académicos se centraba únicamente en estudios de coeficiente intelectual y lógica matemática visto en Terman et al. (1945). A comienzos de los años 1970 comenzaron numerosos estudios sobre desempeño académico, de ellos vemos un estudio conducido en la Universidad de Londres, donde se concluye que la predicción del desempeño académico de los estudiantes es mucho mayor al incluir variables adicionales a los resultados de pruebas de admisión (Freeman, 1970). El autor apuntó a desviar la atención a la importancia que se le daba en el campo de estudio a los exámenes de admisión en aquel entonces, y comenzar a considerar otras variables. Las variables que consideró incluyen test de personalidad, información del contexto de los estudiantes, autobiografías y otros reportes. En particular sobre las notas obtenidas en la secundaria por los estudiantes vemos en un estudio realizado en estudiantes universitarios en Canadá, que la relación entre las notas universitarias finales depende en forma cuadrática de las notas de la educación secundaria (Rubin, 1977). Lo anterior implica que las notas de la educación secundaria son una variable más importante en los estudiantes con notas secundarias bajas que en sus pares con notas altas. Ahora bien, a mediados de los años 1970 se generó mucha controversia académica sobre la validez de las notas como variable explicativa de desempeño y con ello, la validez de las pruebas de ingreso universitaria. Las discusiones se

aclararon en torno a la comprensión del ruido que se genera en las notas del promedio universitario gracias al estudio de Goldman y Slaughter (1976). En este estudio se profundiza las relaciones de correlación y regresiones entre admisión y notas académicas para confirmar que es posible predecir el GPA (Notas promedio en inglés) para un estudiante. Pero esta predicción contiene ruido en el resultado debido a la composición de notas dentro del promedio. Esto es debido a que el GPA como promedio general de un estudiante universitario proviene de diferentes cursos que responden a diferentes ramas de la ciencia y que además existen diferentes puntuaciones a igual nivel de desempeño entre diferentes cursos de la misma cátedra. Por su parte en un estudio conducido en los estudiantes de la Universidad de Illinois (House, 2000) en base a los exámenes de admisión concluye que tanto los exámenes de admisión como las mediciones del background o contexto social de los estudiantes se relacionaban significativamente con los grados de desempeño de los estudiantes.

Uno de los principales estudios relacionado directamente con el caso de estudio de esta tesis es el conducido por Eskew y Faley (1988) en el primer año de una carrera de negocios y contabilidad en Estados Unidos. El autor concluye que la varianza en los desempeños académicos al primer año es explicada en parte por la experiencia académica previa, educación secundaria y datos sobre las características del estudiante. Las variables que utilizó para este estudio se pueden ver a continuación en la Tabla Nº 1.

Variable	**Definición**
Notas examen	Medida de desempeño del curso igual al total de puntos obtenidos en cuatro exámenes del curso.
Aptitud Académica	Medida de aptitud académica igual a la suma de puntajes matemáticos y verbales del test de aptitud escolástica divida por 10.
Notas Secundaria	Notas de la educación secundaria igual a la suma del promedio de matemática y lenguaje.
Notas Bachillerato	Medida del desempeño académico igual al cumulo de las notas de grado.
Número de pruebas	Cantidad de pruebas tomadas, en representación del esfuerzo o motivación.
Experiencia Contable	Previa experiencia en contabilidad, con un formato dummy de 1 ó 0.
Previa experiencia universitaria	Medida de la experiencia universitaria anterior, medida como el total de horas completadas en cursos matemática y estadística.

Horas universitarias anteriores	Medida de las horas de los cursos completados en semestres anteriores.

Tabla Nº 1. Variables utilizadas para explicar desempeño. Fuente: Eskew y Faley (1988)

En un estudio realizado por dos años en una institución de educación superior por Lowis y Castley (2008), se desarrolló un instrumento para identificar el bajo o buen rendimiento de un estudiante y su posible abandono. Este estudio se relaciona con las mismas teorías descritas anteriormente. El autor centra su medición en la adaptación y su proceso en los estudiantes secundarios de cara a la educación superior. Esto es basándose en las expectativas de ingreso contra la realidad universitaria. Se toma una medida de las primeras expectativas de la educación superior que traen consigo los estudiantes secundarios (Thomas, 2002) y posteriormente se comparan esas expectativas con la experiencia real. El diseño del experimento también consideró el concepto de un contrato psicológico, tomada desde el campo de las relaciones laborales, y los siete principios de la buena enseñanza de pregrado (Chickering, 1991), que favorecen el buen desempeño. En la Tabla Nº 2 podemos ver una breve definición de estos principios.

Principio	**Definición**
Nivel de contacto de estudiantes y staff	Trata de la cantidad de horas e instancias en que los estudiantes comparten con los profesores
Reprocidad y cooperación de los estudiantes	Trata del comportamiento de los estudiantes en términos de cooperación o competencia.
Aprendizaje activo	Se centra en las actividades de estudio no obligatorias y búsqueda de materiales adicionales.
Feedback puntual	Trata de la guía y retro alimentación que le den los profesores a un estudiante.
Conciencia del tiempo necesario en la tarea	Cuan consiente es un estudiante de que demorará realizando determinadas tareas
Expectaciones altas	Se relaciona con las expectativas que traen los estudiantes al comenzar la universidad, en relación con los otros principios.
Respeto por la diversidad de talentos y formas de aprendizaje	Grado en que se favorecen diferentes niveles de aprendizaje y la pluralidad entre los estudiantes.

Tabla Nº2. Principios de buena enseñanza de pregrado. Fuente: (Chickering, 1991).

El experimento conducido por Chickering (1991) consistía en dos fases, en la primera fase se les realiza un set de 7 preguntas de rango (1-4) basados en los 7 principios. Las dos etapas ocurren en la primera semana de clases y en la novena semana. En la segunda fase se les envía un cuestionario por email para ser respondido por los estudiantes. Luego se describe la estadística de los resultados y la correlación entre el puntaje de las respuestas (alto) se correlaciona en gran medida con las notas de los estudiantes (altas). Los resultados en general muestran que en la fase uno, los puntajes de experiencia en la semana 9 fueron menores que las expectativas en la semana 1, indicando una baja comprensión de los requerimientos de la educación superior. Un aspecto clave del desempeño actual es el entendimiento de cuánto tiempo es necesario para estudiar y cuanto es dedicado a hacerlo. Por su parte las mujeres tienen una apreciación más realista del tiempo de estudio, y al efectivamente realizarlo, obtienen mejores notas que los hombres. Los estudiantes que se retiran son los que no se relacionan con otros estudiantes, no participan de actividades de aprendizaje, o bien no utilizan el suficiente tiempo en estudio. Un programa de intervención de las entrevistas fue instituido con estudiantes baja puntuación. En la fase dos de entrevistas, un tema recurrente fue la falta de logro en el primer periodo académico. Los estudiantes necesitaban sentir que estaban bien encaminados con feedback en los primeros trabajos. Entre las principales conclusiones encontradas por Chickering (1991) es que a ojos de los estudiantes las clases con muchos estudiantes no son generalmente del gusto de los estudiantes, por falta de interacción, feedback y discusión. Se piensan más útiles los seminarios y workshops. A los estudiantes les gustaría más contacto con profesores al principio para aconsejarlos y darles seguridad. Por lo tanto, en palabras de Chickering (1991), los estudiantes requieren más ayuda al alentar horas de estudio independiente, pues es una nueva tarea para ellos. Se destaca también que estudiantes que formaron grupos de estudio por cuenta propia dijeron que se beneficiaron de él, también que estudiantes mayores prefieren compañeros de su misma edad. En algunos casos estudiantes sienten que existen posibilidades de mayor preparación para los exámenes finales. Según los resultados anteriores se concluye que el proceso educativo es muy complejo y las variables de comportamiento de los estudiantes explican en buena medida el desempeño alcanzando.

Como vemos el estudio de Chickering (1991) pone en práctica muchas de las teorías de predicción de desempeño que utilizan variables más complejas y más integras. En los últimos años el foco esta puesto en predecir efectivamente las notas de graduación en contextos más complejos. Entre los estudios realizados se encuentran predicciones por sobre un 90% de correlación de la nota de graduación de estudiantes universitarios (Tekin, 2014), utilizando la información académica disponible al primer, segundo y tercer año. En este caso el autor pone foco en las mediciones de notas de los estudiantes en diferentes cursos.

3.1.1 Clasificación de estudiantes universitarios

En las investigaciones sobre la predicción de desempeño es común utilizar una forma de clasificación de estudiantes para luego realizar una predicción sobre esta clasificación, por ejemplo clasificar entre estudiantes de alto desempeño y estudiantes de bajo desempeño. Se revisan a continuación estudios sobre clasificación de estudiantes con respecto a su desempeño.

La clasificación de estudiantes apunta a separar a los estudiantes en diferentes grupos. Es decir se definen un modelo o perfil de estudiante que describe los resultados, intereses, metas o preferencias de un usuario o grupo de usuarios en particular. Estos perfiles comúnmente son utilizados para ofrecer contenido y servicios personalizados a los usuarios de esta información (Widyantoro, 1999). Cabe destacar que las clases al momento de predecir pueden ser numéricas o bien de tipo binomiales. Estos modelados apuntan a transformar estas clases numéricas en clases binomiales, por ejemplo aprueba o reprueba.

Las formas de clasificar estudiantes son variadas, pero existen ciertas tendencias. Es así como Chrysafiadi y Virvou (2013) plantea una revisión de los distintos acercamientos en cuanto a la clasificación de estudiantes utilizados por la literatura, incluyendo una vasta gama de estudios en educación y e-learning. En palabras del autor existen tres principales preguntas que deben ser respondidas para la clasificación de estudiantes. Las tres grandes preguntas son: que características se desean clasificar, cómo clasificarlas y finalmente cómo se va a usar el modelo o usuario modelado. En cuanto a las clasificaciones de estudiantes, estos pueden apuntar a descubrir grupos similares, estereotipos, restricciones y causalidades entre las variables Los estudiantes transfieren el conocimiento a través de experiencias por medio de modelos mentales que son usados para asimilar nueva información según explican Vizcaíno, Olivas y Prieto (2000) y esta transferencia es parte importante del proceso de aprendizaje. El aprendizaje es una actividad social, según Vizcaíno et al (2000) y está asociado directamente con la conexión con otros seres humanos como profesores, compañeros, familia, amigos; por ende el aprendizaje es contextual. En cuanto a las herramientas utilizadas para clasificar, Chrysafiadi y Virvou (2013) las separa en ocho grandes tipos pero que corresponden a técnicas de minerías de datos.

Las clasificaciones en los estudiantes permiten estudiar de mejor forma las observaciones sobre su conocimiento, sus preferencias de aprendizaje y también sus creencias. En los estudios sobre estudiantes universitarios que realiza Chrysafiadi y Virvou (2013) apuntan a que comúnmente se mezclan técnicas de clasificación, es decir se modela tanto conocimiento como también preferencias de aprendizaje y creencias.

3.2 Variables que impactan en el desempeño académico

Las principales variables que explican el desempeño académico han evolucionado bastante a lo largo del tiempo. Los primeros esquemas causales se enfocan en relacionar el desempeño académico casi únicamente con su nivel de coeficiente intelectual y lógica matemática (Terman, 1916). Este autor en sus estudios argumenta que los estudiantes con mayor nivel de coeficiente intelectual, asociado a variables genéticas, obtenían mejores resultados académicos y que además era posible detectarlo usando evaluaciones de coeficiente intelectual y lógica matemática en pruebas con puntajes. Con el pasar de los años, los nuevos estudios y teorías a mediados del siglo XX comenzaron a demostrar que hay muchas más variables que explican el desempeño de un estudiante. De hecho, en estudios que comparan grupos de niños con alto índice de coeficiente intelectual versus otros niños de coeficientes promedios y a lo largo del tiempo se obtienen resultados de desempeño académico similares entre ambos grupos por Terman et al. (1947). Es decir que los dos grupos tenían estudiantes brillantes que luego tenían un futuro laboral exitoso como también estudiantes con un desempeño promedio que luego tenían un futuro laboral promedio o mucho más sombrío. En relación con los estudios descritos anteriormente, desde ramas como la psicología también existen estudios que concluyen que características particulares de un estudiante pueden generar mejores desempeños académicos que solamente resultados de pruebas de admisión basadas en lógica matemática. En un estudio entre estudiantes de la universidad de Pensilvania se les realizo a los estudiantes test de coeficientes intelectuales como también test de personalidad para medir su nivel de autodisciplina. Posteriormente al comprar los desempeños académicos de los estudiantes se observó que los estudiantes con mayor nivel de autodisciplina obtienen mejores desempeños académicos que sus pares con mayor coeficiente intelectual según el estudio de Duckworth y Seligman (2005), especialmente en el 10% superior.

Sobre las variables de estudiantes se observan dos grandes tipos por Chrysafiadi y Virvou (2013). El primer tipo tiene que ver con características fijas (Como la lengua, nacionalidad, etnia, etc.). Mientras que las segundas características son dinámicas (Como el desempeño, la cantidad de conocimiento, comportamiento, etc.). Las dinámicas presentan desafíos pues existe una mayor complejidad de relacionar la variable independiente con estas características dinámicas. Los estudios comentados hasta ahora evidencian que existen habilidades intrínsecas que permiten desarrollar características cognitivas para luego alcanzar un mejor desempeño académico.

Profundizando el análisis más allá del coeficiente intelectual, muchas ramas de la investigación pusieron sus esfuerzos en caracterizar otras variables explicativas. Una de ellas es el contexto familiar, en particular la atención y esfuerzo de los padres los que tienen una correlación importante con el desempeño académico (Shurkin, 1992) y hasta el futuro laboral exitoso.

Actualmente ya existen variados estudios que apuntan a la importancia de los padres en los resultados de desempeño educacionales desde el colegio como el de Christenson y Huebner (2010), entre ellos destacamos un modelo teórico probado con regresiones econométricas sobre desempeño académico que trata el involucramiento de los padres con tres grandes efectos en los estudiantes (Al-Alwan, 2014). Los efectos son en el comportamiento del estudiante, en el compromiso y estabilidad emocional para responder a las actividades académicas, y finalmente el compromiso cognitivo. Se puede apreciar esta relación en la Figura Nº 4.

Figura Nº 4. Modelo de involucramiento de los padres, efectos y desempeño académico.
Fuente: (Al-Alwan, 2014)

Además de las variables explicativas sobre el puntaje obtenido en pruebas de coeficiente intelectual, puntajes en pruebas de lógica matemática y contexto familiar, comienzan a aparecer desde mediados de los años 1980 variables que responden a características intrínsecas del estudiante.

Desde la psicología y sociología comenzaron proliferar explicaciones sobre desempeño académico más allá del coeficiente intelectual. Estas nuevas variables ponen foco en el comportamiento de los estudiantes y explicación de las emociones que se traducen en comportamientos que pueden apoyar el desempeño académico (Payne, 1985). La decantación de estos estudios aparece alrededor de 1990 con teorías sobre la inteligencia emocional. Esta inteligencia es la que envuelve las apreciaciones y uso de las emociones para alcanzar objetivos y formar relaciones por Mayer et al. (1990), y luego en extensión a esto la inteligencia emocional se utiliza cómo un marco de habilidades que contribuyen apreciar y expresar emociones según explica Salovey y Mayer (1990), y como explica Goleman (1998) es gracias al uso de estas habilidades que muchos individuos, y estudiantes también, son brillantes en sus áreas de desempeño.

Cabe destacar que diversos investigadores al plantear nuevas variables que explican el desempeño, plantean a su vez mecanismos para registrar y obtener estas variables. Así

comienzan a aparecer muchos test para identificar los tipos de inteligencia, entre ellos Schutte et al. (1998) genera un estudio para medir la inteligencia emocional. Estos test se han realizado también en estudios en campus universitarios, con foco en conocer el comportamiento y niveles de inteligencia emocional para conocer los efectos en el desempeño académico de los estudiantes. Dentro de este tipo de estudios encontramos que se ha predicho el desempeño desde la inteligencia emocional como el estudio de Parker et al. (2004) en estudiantes de primer año universitario. Es este estudio se comparan estudiantes con buenos resultados académicos, es decir sobre el 80% de los mejores estudiantes versus quienes estaban bajo el 59%. Se observa que su posición en el desempeño académico superior o inferior tiene una fuerte relación con varias de sus dimensiones de inteligencia emocional. Particularmente esta tesis se relaciona con la transición que ocurre y la capacidad de un estudiante de adaptarse a su nuevo ambiente universitario y obtener así un buen desempeño, según los conocimientos y habilidades que posea desde su previo ambiente escolar. Mestre et al. (2006) presenta un estudio en que ya en los cursos secundarios se evidencia que los estudiantes con mayores habilidades emocionales tenían mejores indicadores de adaptación social y académica. Es decir que los estudiantes tienen mayores habilidades de adaptación y de inteligencia emocional logran obtener mejores desempeños académicos en la universidad al inicio de la carrera universitaria, y con ellos una ventaja con respecto a sus pares.

Pittman (2008) quien utiliza evaluaciones de distribución de los atributos para destacar los más importantes. Logra destacar variables entre las 103 variables que utiliza. Sus tipos de atributos son principalmente cuatro: a) descripción de apoyo financiero, b) Indicadores de performance académico, c) descripción y experiencia de profesores y finalmente d) categoría de los atributos. Busca entonces los atributos que indiquen retención y luego los que atributos que indican que un estudiante se va de la carrera. Sus principales resultados son que no se encuentran variables para el Año 1 y Año 2. Ahora bien en los resultados para el Año 3, se encuentran predicciones de un 69% y encuentra diferentes reglas de asociaciones para el desempeño, abandono de carreras, foco en variables de apoyo financiero y otros comportamientos como vida de campus y educación de los padres. Los indicadores de performance en educación y experiencia de profesores no resultan ser buenas variables. En un estudio sobre admisión de la universidad, (Acikkar, 2009) realiza una predicción en cuanto a si un candidato será aceptado o no en la universidad de educación física según su desempeño en pruebas físicas, GPA (Graduation Point Average), resultado en prueba nacional de selección y test de placement. Los datos que utiliza para su estudio son los resultados de pruebas físicas de los candidatos en donde cada estudiante obtiene un puntaje por cada prueba, la prueba nacional de selección de esos candidatos, los resultados del test de placement del gobierno, GPA y las notas de especialización del colegio. En otro estudio pero en particular para estudiantes mujeres

(Pradmapiya, 2012) los datos utilizados en son la data del colegio de artes para mujeres del gobierno, la información personal (hijos, ingresos, padres, nacimiento), información pre universitaria (lugar, perfil y notas), y también los datos universitarios (Rama, notas, intereses).

En un estudio sobre admisión universitaria (Kabakchieva, 2013) en cuanto a las variables explicativas importantes se encuentra que la variable con el mínimo error es Puntaje de Admisión. Las rreglas de clasificación encontradas se concentran, entre número de fallos, puntajes admisión, semestre y universidad y notas de colegio. Arboles de decisión pone en primer nivel el número de fallos, puntaje de admisión, semestre en segundo, y en tercero la universidad, carrera y género. Los resultados muestran el puntaje de admisión y el número de fallos en los exámenes de primer año están en los factores más importantes.

3.3 Minería de Datos y el desempeño académico

El estudio del desempeño académico de estudiantes universitarios también ha sido revisado por las técnicas econométricas y el uso de técnicas de minería de datos. En estas técnicas se revisan tanto los factores que se traducen en un determinado desempeño o comportamiento, cómo también realizan predicciones utilizando esta información. Las relaciones de información oculta que entregan las técnicas de minería de datos pueden ser utilizadas para predecir la solicitud de cursos, las tasas de deserción, resultados anormales en exámenes y también predecir desempeño según explica el estudio de Yukselturk, Ozekes, y Turel (2014). Ahora bien la aplicación de estas técnicas en los procesos educacionales implica un tratamiento especial de Minería de datos para cada problema y contexto, esto por sus características únicas, donde cada casa educacional y contexto dentro de ella es diferente (Li & Zaiane, 2004; Pahl, 2003). Se han identificado variables con un grado de relevancia estadística importante según diversos investigadores la que es información valiosa para entender el proceso educativo y la medición de resultados a cabalidad.

Los trabajos de minería de datos pueden ser clasificados según el propósito de la técnica utilizada y hacia quien es orientada Zorrilla et al. (2005) en el contexto de la educación superior, tenemos que pueden ser de tres grandes tipos. El primer tipo son las técnicas orientadas hacia los estudiantes. Estas permiten recomendar a estudiantes recursos, actividades o tareas para mejorar su desempeño en un curso en particular o tema. El segundo tipo son las técnicas orientadas a los educadores. Estas permiten dar el feedback adecuado a profesores, clasificar estudiantes, errores comunes, organización de contenidos y mejorar adaptación de los profesores. El tercer tipo es orientado a la administración universitaria. Algunos ejemplos de

esto guardan relación con mejorar eficiencia del lugar, comprender el comportamiento del usuario, como también una mejor organización educacional y programas académicos.

Para continuar con la clasificación de estos trabajos y la comprensión de los objetivos que persiguen, las técnicas son usadas para apoyar tanto el proceso de aprendizaje en sala y también el aprendizaje basado en web. En el desarrollo de estos estudios es común utilizar datos desde Sistemas de administración docentes y base de datos de plataformas online como explica Arruabarrena, López, Pérez y Vadillo (2002) con el objetivo de mejorar el aprendizaje cómo también técnicas de evaluación formativa. Un estudio bastante completo sobre la literatura de Minería de datos en educación divide los trabajos realizados entre 1995 y 2005 por Romero y Ventura (2007), en dos grandes sistemas educacionales. El sistema tradicional y los sistemas a distancia. Cada uno de estos sistemas educacionales presenta diferentes problemas y desafíos estudiados. Los sistemas tradicionales son caracterizados principalmente por realizarse en la sala de clases, con lecturas dictadas por un profesor con diferentes materiales audiovisuales de apoyo. Los estudios realizados en sistemas con profesores en sala, apuntan fundamentalmente a cuatro objetivos según Romero y Ventura (2007) tenemos que son: a) Predecir la toma de ramos desde los estudiantes, b) Revisiones curricular que afectan a estudiantes, c) Seleccionar estudiantes débiles a atender clases extras y por último d) Análisis extenso de características de estudiantes.

En cambio, los sistemas educacionales a distancia no tienen una sala de clases física. Los estudios realizados se concentran en plataformas web que permiten acceso a los estudiantes y profesores para interactuar y dar soporte al proceso educativo. Los dos principales tópicos objetivos según Romero y Ventura (2007), que se han resuelto con técnicas de Minería de datos son dos: a) Suman otros canales de comunicación para modelar comportamiento al navegar a través de la pagina web o sistema interactivo, y finalmente b) Descubrir patrones para asociar comentarios y eventos web que sean de interés para los estudiantes, académicos y miembros administrativos. La principal diferencia entre ellos es la capacidad de interactuar de los estudiantes con el sitio web y su contenido. Ahora bien dejaremos de lado los estudios relacionados con sistemas educacionales a distancia dado que en esta tesis nos centraremos en el sistema tradicional de educación.

3.3.1 Técnicas de minería de datos utilizadas en predicción

Ahora bien, revisaremos a continuación las principales técnicas de Minería de datos utilizadas en estudios relacionados con sistemas tradicionales de educación que se centran en los resultados de los estudiantes.

Cabe destacar que, antes de aplicar técnicas de minería de datos es útil aplicar técnicas de visualización y estadística de datos masivos, en este sentido Romero y Ventura (2007) destacan que la mayoría de los estudios desarrolla un análisis estadístico previo a la aplicación de Minería de Datos. Pues responde a la metodología de análisis Hipótesis-Datos y están presentes en diferentes partes del proceso de Minería de datos. Siendo los primeros acercamientos de estudios al desempeño de estudiantes, los casos en cuestión son por ejemplo para el uso de herramientas particulares por los estudiantes como es el estudio de Zaiane, Xin y Han (1998). Cómo también consultas SQL que buscan proveer y permiten interpretar resultados de grandes cantidades de datos en los sistemas de docencia administrativos como en Heiner et al. (2004).

La minería de datos en educación, de acuerdo a Luan (2002), persigue descubrir patrones y hacer predicciones que caracterizar las conductas y los logros de los estudiantes, el conocimiento de dominios contenidos, evaluaciones, funcionalidades educativas y aplicaciones. Los últimos resultados de estudios apuntan en su mayoría, aproximadamente en un 98% (Peña-Ayala, 2014) de los trabajos en EDM son posteriores a 2000. Se destaca que trabajos sobre predicción fueron un 60% y descriptivos un 40%. Por lo tanto, existen trabajos que apoyan la predicción de resultados académicos, los que serán revisados para entender sus variables a continuación.

Las técnicas de clustering, clasificación y detección de outlier también son ampliamente usadas en las investigaciones que utilizan Minería de datos. La detección de elementos que sobresalen a una muestra en particular brinda resultados importantes en diferentes campos de estudios. Existen diversos algoritmos de clustering como: aglomeración jerarquica y k-means, en un estudio universitario amplio (Ayers, 2009) fue posible identificar grupos de estudiantes con perfiles de habilidades diferentes gracias a estas técnicas. K-means ha sido utilizado para descubrir patrones que caractericen los trabajos de estudiantes fuertes y débiles en cuanto a sus desempeños académicos individuales. Las técnicas de Asociación de Reglas también han sido bastante utilizas. En aplicaciones a grandes grupos de estudiantes se utilizó para recomendar materiales de aprendizaje por Markellou et al. (2005), adecuados según las formas de aprendizaje y estudio de cada estudiante. También se conoce el uso de Redes Bayesianas, por ejemplo en un estudio de por Huang et. al (2007) se utiliza para predecir respuestas o habilidades y son aplicadas en test de tipos de personalidad según comportamiento.

Por su parte (Nadeshwar, 2011), enfoca su estudio en el uso de recursos para estudiantes de alto riesgo, para aumentar su probabilidad de terminar la universidad y así disminuir programas de retención que se enfocan en variables de desempeño. Nadeshwar (2011) basa su estudio en trabajos como el de (Baker 2004) quien usa redes neuronales y Support Vector Machine (SVM) para estudiar tasas de graduación. Así como también se basa en el trabajo de Pitman (2008) en

su estudio las principales hipótesis son las siguientes. Su primera hipótesis trata sobre el apoyo financiero. H1: La ayuda financiera es importante para el desempeño académico y la retención del estudiante, al momento de este estudio la literatura era ambigua. La segunda hipótesis tiene que ver con las notas de los estudiantes. H2: El desempeño académico (GPA) tiene una alta relación con la retención del estudiante. Mientras que la tercera hipótesis es sobre el cuerpo académico de la universidad. H3: La Experiencia de profesores, en este caso, los profesores part time tienen impacto negativo con graduación. Pitman (2008) Aplica una serie de técnicas de minería de datos para luego comprobar sus resultados usando Cross Validation (Kohavi, 1995) en donde divide la base de datos en 5, luego con 1/5 de cada subsets entrena con el restante 4/5, luego de 5 rondas se guardan las medias de recall y falsas alarmas. En cuanto a los estudios particulares relacionados con la admisión de la universidad, (Acikkar, 2009) realiza una predicción en cuanto a si un candidato será aceptado o no en la universidad en un programa donde se admiten 25 hombres y 15 mujeres, siendo bastante menor la clase positiva como caracteriza todos estos casos de estudios. Se utiliza 5-fold cross validation y el proceso de aplicación de técnicas de minería de datos en base al modelo SVM combinado con Grid Search. Podemos ver una descripción grafica de este proceso a continuación en la Figura Nº 5.

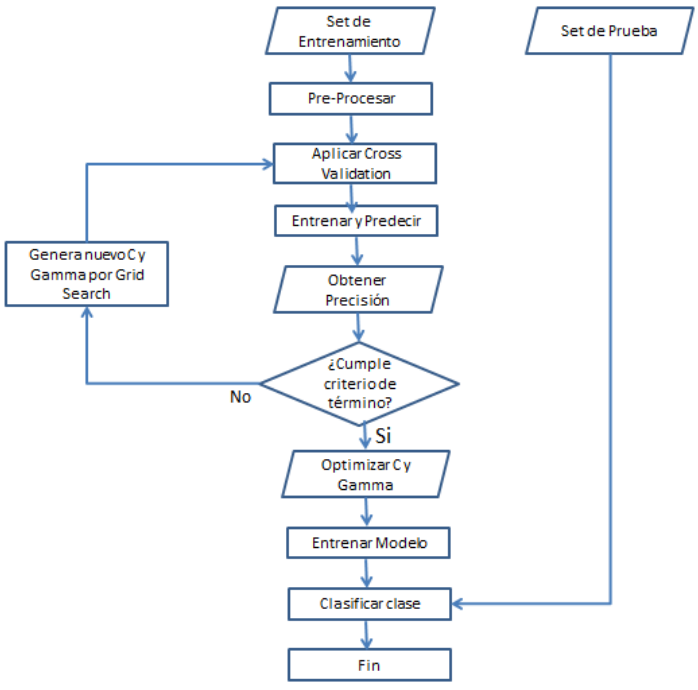


Figura Nº 5. Modelo SVM con Grid Search. Fuente: (Acikkar, 2009).

En sus resultados obtiene una performance de predicción cercano a un 90%. Además el autor destaca que SVM no puede predecir sobre candidatos con puntajes cercanos al puntaje de admisión que cambian cada año, y que al aumentar la diferencia entre el puntaje corte y los

candidatos SVM mejora su desempeño. Siendo lo anterior relevante para esta tesis y será revisado en el caso de estudio en los capítulos siguientes.

En otro estudio (Pradmapiya, 2012) pero en particular para estudiantes mujeres, se busca predecir las mejores candidatas que buscan estudios avanzados basados en su información personal, pre universitario y graduados. Entonces la clase es tomar un postgrado y utiliza la información disponible para predecir si tomará el postgrado o no. Las Técnicas utilizadas son Arboles de Decisión y Clasificador Bayesiano. Sus resultados muestran que Arboles de decisiones tienen un accuracy de 93%, con alto accuracy la clase predominante. El otro modelo predice mejor la clase débil. Entonces la data personal y universitaria resulta importante para el estudio de estas decisiones de admisión universitaria. En un estudio en particular sobre las necesidades de admisión de una universidad, para la aplicación de minería de datos, Kabakchieva (2013) logra clasificar estudiantes de acuerdo a sus resultados de performance universitaria basados en sus características pre-universitarias. Utiliza para lo anterior un aprendizaje supervisado, pues el target de los modelos es conocido. Entonces genera clases posibles para clasificar la performance de un estudiante: Excelente, Muy bueno, bueno, medio y malo. Los datos que utiliza son las campañas de admisión, los datos pre-universitarios personales, escolares y exámenes de admisión, datos de la universidad, como carrera y datos como promedio en el primer año, y también performance universitaria promedio (Clase – según parámetros de gobierno Búlgaro). En este estudio se utilizan 10330 estudiantes de Bulgaria de 3 Universidades en periodo de campaña, 2007 y 2009. En cuanto a las técnicas de DM utiliza los más populares como: arboles de decisiones, clasificadores bayesianos y dos técnicas de Reglas de Asociación. Cada clasificador se testea dos veces, utilizando Cross Validation y Split de Porcentaje. Split utiliza dos tercios de la base de datos como entrenamiento y prueba sus resultados en el tercio restante. Sus resultados en cuanto a la predicción el autor presenta que el desempeño de los modelos es muy pobre en los estudiantes con excelente desempeño académico con resultados bajo un 20% de aciertos. Mientras que para los malos estudiantes obtiene resultados cercanos a un 80% de aciertos, para los estudiantes de desempeño bueno a muy bueno obtiene modelos con resultados entre 60% y 75% de aciertos.

La literatura revisada apuntan a probar el modelo con nuevos datos o set de datos que se apartan para usarlos solo para medir el desempeño del modelo, según Baesens et al. (2003) se destaca que es mejor utilizar datos en posteriores periodos de tiempo que parte de la muestra que no participó del training. A su vez, y de acuerdo con la literatura en general, Baesens et al. (2003) explica que es valioso utilizar sólo en Training 70% de la data y en el Test un 30 %.

Capítulo 4–Metodología

En este capítulo se describe la metodología que se utilizara en esta tesis. La metodología corresponde al Knowledge Discovery in Databases – también conocida por sus siglas como KDD (Fayyad et al, 1996). Esta metodología busca encontrar conocimiento, a través del análisis exhaustivo de los datos utilizando técnicas de conocidas como minería de datos.

El proceso de extracción de conocimiento toma los datos y permite ejecutar predicciones en múltiples campos en forma acertada como ilustra a través de muchos ejemplo Provost y Fawcett (2013). Todos los estudios revisados en el Capítulo Nº 3 han sido realizados utilizando el proceso de KDD. Este proceso construye sus resultados en orden de desarrollar predicciones analizando exhaustivamente los datos a través del ensayo y error en las raíces de la estadística y la ciencia basada en computadoras, o en inglés Computer Science.

El proceso KDD se caracteriza por una serie de cinco grandes etapas sucesivas entre sí según la metodología planteada por Fayyad et al. (1996), en que los autores plantean el objetivo de preparar los datos lo mejor posible para ejecutar correctamente las técnicas de minería de datos, como también aumentar la probabilidad de éxito de estas técnicas. El proceso consiste en dar el trato correcto a los datos con el fin de obtener conclusiones suficientemente confiables, el primer paso consiste en seleccionar los datos apropiados los que pueden ser mediante técnicas de estudio estadístico o bien en base al criterio clínico y experto de quien realiza el estudio. La segunda etapa es el pre procesamiento en que se revisan por completo la robustez de los datos. La tercera etapa consiste en la transformación, que es donde se realizan cambios en el tipo de datos para ser procesado por determinada técnica de minería de datos. La cuarta etapa es ejecutar el proceso de minería de datos, normalmente utilizando un software que contiene estas técnicas. La quinta etapa es evaluar los resultados del estudio y además brindarles la apropiada evaluación pues el resultado está sujeto al tipo de técnica de minería de datos. Podemos ver a continuación un diagrama en la Figura Nº 6 que describe el proceso.

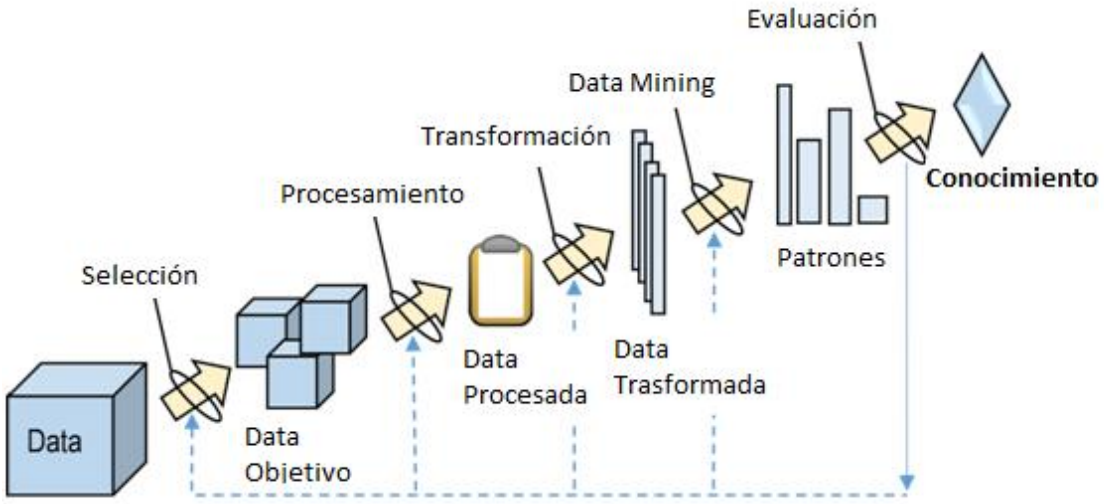


Figura Nº 6. Etapas del proceso de extracción del conocimiento. Fuente: Fayyad et al. (1996).

Si bien es cierto que las predicciones no son exactas, y que por tanto es necesario reconocer los límites de una predicción es también necesario reconocer el tremendo potencial que esto tiene. Es sabido que las predicciones del clima no son exactas, de hecho actualmente bordean el 85% de exactitud en los mejores escenarios, según el estudio de Mass et al. (2002), no deja de ser útil conocer estos resultados y continuar profundizando en ellos.

De la misma forma esta generación de conocimiento sobre patrones tiene efectos importantes en las ventas de muchas compañías tales como Target, quien popularizó el concepto de predicción en ventas aumentando entre 15% y 30% sus ingresos (Pole, 2010), a su vez el casino Harrah's Las Vegas puede predecir cuánto un cliente gastará en cuanto a largo plazo (Loveman, 2003). También existen muchos casos prominentes en finanzas como reconocidos inversionistas y fondos de inversión que basan sus decisiones en algoritmos predictivos como los discutidos anteriormente, se estimado que el 40% de London Stock Exchange es conducido por sistemas de algoritmos (Brugler, 2014). Incluso los gigantescos bancos Chase y Citigroup utilizan modelos predictivos forense para predecir quien refinanciará sus hipotecas por Bumacov y Ashta (2011) o bien generar modelos de Credit Scoring, es decir puntuar a sus clientes según su capacidad de pago en los préstamos que adquieren revisado por Huang et al. (2007). Las aplicaciones de técnicas de minería de datos son tan variadas que pasan por modelos predictivos en salud, por ejemplo predecir un aumento en casos de influenza en un hospital 7 o 10 días antes que lo haga el Center for Disease Control, utilizando las búsquedas de resfríos y otras en Google según estudió Ginsberg et al. (2009). Por último podemos mencionar trabajos importantes en educación que permiten predecir diversos casos de abandono universitario para brindarle apoyo a estos estudiantes (Peña-Ayala 2014; Miranda & Vásquez, 2015).

4.1 Selección

La primera etapa consiste en seleccionar las variables y registros con los que se ejecutara el proceso de descubrir conocimiento. Según vimos en la literatura existen técnicas de diversos tipos para realizar esta selección. Estas técnicas pueden ser estadísticas y matemáticas, procesos de detección de ganancia de información e incluso la aplicación de modelos de minería de datos para seleccionar las variables o orientar al investigador. Las variables deben cumplir requisitos tales como disposición de los datos, pocos errores y existe su medición antes de que ocurra la situación que se estudie. Se estudian los datos para obtener información respecto a los tipos de datos, dispersión, outliers, sesgos en los datos, a las categorías de datos y acumulación de los estudiantes en ellos.

Para apoyar la selección de variables se utilizarán los criterios de Information Gain y Decisión Tree según explica Provost y Fawcett (2013), así como también Select Backward Elimination. El proceso de Information Gain calcula cuánta información aporta cada variable respecto de la clase a predecir. Por su parte el Decisión Tree es una técnica de minería de datos en sí misma, pero que por sus características puede ser utilizada como una guía al investigador. Inclusive puede ser utilizada en modelos automáticos. Se describe esto en detalle en la sección de minería de datos. Las técnicas de Backward Elimination es un algoritmo en sí mismos que encierran un proceso de mejora continua.

4.1.1 Backward Elimination

Esta técnica permite obtener la mejor selección posible de atributos de acuerdo al desempeño del modelo (Guyon, 2003). El proceso iterativo comienza con un set completo de todos los atributos considerados en la etapa de selección. En cada iteración el proceso remueve un atributo de entre todos los seleccionados en un principio. Por cada atributo que es removido se estima una performance utilizando un proceso como Split Validation o Cross Validation. Solo el atributo que arroja el desempeño más pobre es removido de la selección inicial. Luego de removido comienza nuevamente el proceso con otra ronda de eliminación. Para detener este proceso de eliminaciones iterativas se define un parámetro de detención. El comportamiento del parámetro de detención especifica cuándo se debe abortar el proceso de iteración. Las opciones de comportamiento de este parámetro son:

1. Detención al disminuir performance: La iteración ocurre mientras no exista ninguna baja en el desempeño del modelo predictivo.

2. Detención con disminución límite: La iteración ocurre hasta que la baja en el desempeño es menor que un determinado límite.

3. Detención con disminución significativa: La iteración ocurre hasta que la disminución en el desempeño cae sobre un nivel porcentual definido.

4.2 Pre procesamiento

El objetivo es eliminar ruidos o perturbaciones en los datos que puedan guiar a errores al modelo. El trabajo realizado con los datos en el procesamiento de datos, incluye: limpieza, identificación del usuario, identificar la sesión, completar referencias, información transaccional, transformar datos, enriquecimiento, integrar datos y reducción. Se consideraron los siguientes criterios:

a. Coherencia: Esto quiere decir que los datos tengan relación lógica con el contexto de parámetro. En este sentido se identifican los outliers, es decir los casos que son distantes del resto de la muestra.
b. Llenado: Que no existan variables con campos vacíos, en ingles se conocen como missing values.
c. Formato: Se valida que los registros sean acordes a la codificación esperada y formato adecuado para ser considerado un registro valido.

Ahora bien para asegurar el cumplimiento de estos criterios la literatura plantea diferentes tratamientos a los datos con problemas. Los que son reemplazar por moda, reemplazar por un modelo predictivo y reducir registros.

a. Reducir registros, quiere decir que se remueven de la muestra de investigación las instancias que no tengan todos los datos necesarios. De esta manera los datos serán reales en su totalidad.
b. Reemplazar por un modelo predictivo, quiere decir que se generará un modelo para predecir que el valor en particular que debería tomar la variable faltante o errónea. Estos modelos pueden ser técnicas de minería de datos en sí mismos.
c. Reemplazar por moda y medio, quiere decir que según el tipo de dato, para los datos numéricos un valor vacio puede ser reemplazado por la media. Mientras que para los datos categóricos pueden ser reemplazados por la moda.

4.3 Transformación

Las técnicas de Minería de datos utilizan modelos predictivos que requieren datos en determinados formatos y tipo para ser utilizados. En particular requieren que los datos sean variables numéricas. Pero existen registros que por su naturaleza no son medibles, por lo tanto

son categóricas. Estas variables compuestas por texto son transformadas para poder ser utilizadas en el proceso de minería de datos.

Otra tarea importante en el proceso de transformación es la normalización de los datos para poder obtener mejores resultados. La normalización de los datos soluciona problemas entre rangos amplios, por ejemplo puntajes de pruebas, ingreso familiar y edad. Las diferencias de escala entre los números distorsionan los resultados de los modelos debido al ruido de estas en la asignación de pesos en las variables, por ejemplo de notas entre 1 y 7 mientras que el puntaje es hasta 850 puntos. Para resolver esta problema se aplica una normalización de los datos es decir, que se transformar a un rango entre 0 y 1.

Para convertir las variables no numéricas a numéricas se tienen dos opciones. Las primeras son las variables binomiales, estas son variables que tienen solo dos posibles valores. Por lo tanto para transformarles se les asigna un valor numérico a cada categoría. El segundo tipo de transformación ocurre con las variables polinomiales, estas son variables que tienen más de dos categorías. Para transformarlas se llevan a número generando atributos binomiales por el total de categorías menos uno.

4.4 Minería de Datos

En esta etapa es cuando ocurre la aplicación de la técnica de modelado, clasificación o predicción. Se pone en práctica entonces el objetivo del proceso KDD sobre extraer conocimiento desde los datos. En particular las técnicas aplicadas pueden ser de regresión, caracterización, asociación, clusterización y clasificación (Klosgen, 2012). Las actividades descritas permiten generar un modelo matemático para categorizar cada instancia en las clases o grupos de estudio predefinidos. En cuanto al tipo de técnica a utilizar depende del objetivo del estudio, esto ya que se apuntan las regresiones a predecir números reales como variables, mientras que en los modelos de categorización la variable dependiente es de tipo binomial o polinomial.

Actualmente existen muchas técnicas de minería para ejecutar las actividades que se mencionan anteriormente, y se conocen como Máquinas de Aprendizaje. Cabe destacar que gracias a la creación de software orientado al usuario es cada vez más fácil generar modelos predictivos utilizando estas técnicas. Este tipo de software contienen las técnicas de minería de datos programadas en librerías que pueden ser utilizadas fácilmente. Las técnicas incluidas como Maquinas de Aprendizaje son Decision Tree (DT), Support Vector Machine (SVM) y Logistic Regression (LR).

4.4.1 Clasificadores

Las técnicas de minería de datos descritas en las secciones anteriores utilizan un clasificador estándar para operar conocido como confidence. Esta clasificador se conoce también como un umbral de clasificación. Este límite o punto crítico es el valor que determina si una instancia pertenece a una clase u otra.

Normalmente este estándar de confidence esta fijo en 50%, es decir que el umbral para pertenecer a una clase corresponde a un 0.5 y si una instancia los supera podría ser clasificada como esa clase. Esta medida estándar de confidence puede ser modificada. Esto permite gran libertad de maniobra al investigador de minería de datos, ya que el uso de un confidence puede ser manipulado si se quiere ser más restrictivo o bien más permisivo con las clasificaciones en la predicción.

Para la definición de estos confidence se deja a criterio clínico del investigador o también según la evaluación de del análisis o curva ROC. Que es las representaciones grafica de la precisión comparando verdaderos positivas con falsos positivas. La diagonal de la curva ROC se convierte en una barrera que divide a una buena clasificación de una mala. Entonces es posible evaluar el confidence de determinada técnica de minera de datos por los cambios de puntos en la curva ROC, es decir que sube los resultados de bajo de la curva por sobre ella.

4.4.2 Técnicas de Aprendizaje

Support Vector Machine

Esta técnica es ampliamente aceptada y utilizada por la literatura. Consiste en un aprendizaje supervisado basado en análisis de regresiones y algoritmos de clasificación. Esta técnica fue presentada por Vapnik y Chervonenkis (1964) y se ha profundizado su aplicación desde entonces, así como también las capacidades del hardware y software que soportan esta técnica.

Esta técnica trata sobre encontrar una función de separación representada por un hiperplano en el espacio $\mathbb{R}^2$, ahora bien, se trata de un hiperplano lo más ancho posible, es decir que se maximiza el margen de separación. Cabe destacar que el margen es definido como la distancia entre un par canónico de paralelas, es decir que el margen es igual a dos veces el mínimo entre los puntos del hiperplano de separación según Vapnik y Chervonenkis (1964). Los puntos de separación son en realidad vectores y se les conoce como vectores de soporte. Ahora bien en un espacio $\mathbb{R}^2$ el problema de separar dos puntos finitos es que pueden estar separados por infinitos hiperplanos, es por ello que SVM incluye la identificación lineal óptima en base a la capacidad de generalización y el mínimo error (Vercellis, 2009) de clasificación. Es gracias a los vectores y la capacidad de generalización y error mínimo de clasificación que SVM entrega

reglas de clasificación. Matemáticamente se requiere encontrar el par (w, b) que clasifique correctamente los vectores x_i, y se define w como el vector de coeficientes del hiperplano y b como el intercepto. Se define el hiperplano como:

$$wx_i = b \qquad (1)$$

Los hiperplanos paralelos canónicos se definen como:

$$wx_i - b - 1 = 0$$

$$wx_i - b + 1 = 0 \quad (2)$$

Donde $\| w \| = \sum_{j \epsilon N} w_j^2$, Si el objetivo de SVM es encontrar, entre todos los hiperplanos canónicos que clasifican correctamente los datos, es aquel con mínimo $\| w \|^2$ el que se conoce como margen. Entonces el margen de separación δ se define como:

$$\delta = \frac{2}{\|w\|} \qquad (3)$$

Es interesante notar que la minimización de $\| w \|^2$ es equivalente a encontrar el hiperplano separador (Weber, 2012) para el cual la distancia entre las dos clases del conjunto de datos de entrenamiento, medida a lo largo de una línea perpendicular al hiperplano, es maximizada.

Para poder determinar los coeficientes de w y el intercepto b, el hiperplano que optimiza la diferencia puede ser representado en un problema de maximización de margen como se aprecia en las siguientes formulas que describen el modelo (4).

$$Min_{w,b} \frac{1}{2} \| w \|^2$$

$$\text{s. a. } y_i(w' x_i - b \geq 1, \qquad i \varepsilon M \quad (4)$$

Ahora bien en los casos en que los m puntos no son linealmente separables, se resuelve el siguiente problema:

$$Min_{w,b,e} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} e_i$$

$$\text{s.a. } y_i(w' x_i - b \geq 1 - e_i, \qquad i \varepsilon M$$

$$e_i \geq 0 \qquad i \varepsilon M \qquad (5)$$

Donde C refleja el trade off entre la generalización y el error. Para ilustrar lo anterior, y además evidenciar la diferencia con una regresión en la Figura Nº 7 podemos ver como SVM divide los

registros en dos categorías. Ubicándose a la mayor distancia posible de ellos, y además el hiperplano en cuestión minimiza el error de clasificación. Los puntos en negro son categorizados como Postulantes Destacados mientras que los puntos en blanco son Postulantes no Destacados.

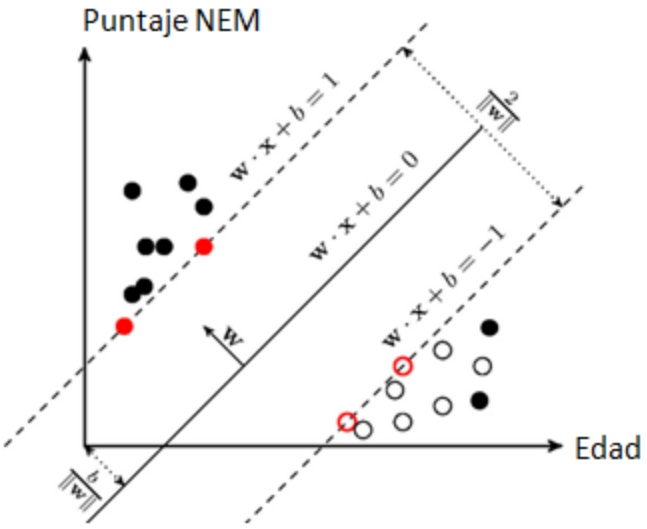


Figura Nº 7. Representación de la aplicación de SVM.

Árbol de Decisión

Esta técnica corresponde a un modelo de clasificación introducido por Quinlan (1986). Lleva su nombre porque la representación es muy característica por ser gráficamente similar a un árbol, ya que tiene ramas y hojas. Esto lo hacen muy comprensible a simple vista.

Esta técnica se basa en teorías de decisiones para realizar clasificaciones a la base de datos. Estos árboles distribuyen los registros a través de sus ramas y hojas utilizando técnicas estadísticas en base al concepto de entropía. Las raíces o nodos, son los puntos donde se toma una decisión para clasificar

En la Figura Nº 8 podemos ver una representación gráfica de un árbol de decisión desde la aplicación en esta tesis, se puede ver claramente la progresión de las decisiones a través de los nodos, y sus extensiones de ramas hacia las hojas. Las hojas contienen una probabilidad de pertenecer a una clase u otra para distribuir las clases.

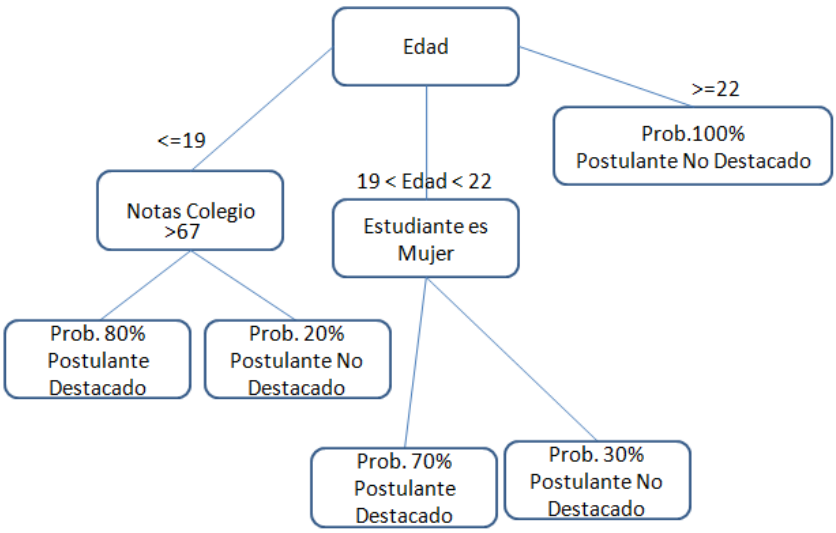


Figura Nº 8. Representación de un árbol de decisión.

Este proceso de segmentación de las clases en hojas pone en práctica los criterios de entropía, las reglas de separación así como también los criterios de poda y prepoda. Se agrupan entonces las instancias en las hojas del árbol utilizando el criterio de entropía. Los criterios de entropía hacen referencia a la distancia entre un punto y otro en el plano, es decir que tan diferente es una clase de la otra. Entonces se calcula la diversidad entre los grupos de clasificación. Estas diferencias pueden ser medidas a través de la entropía, en particular en una raíz o nodo se calcula como sigue:

$$Entropia\ (q) = -\sum_{l=1}^{L} p_l(log_2 p_l) \quad (6)$$

A su vez para cada nodo se cálculo el índice de Gini. Esta índice mide la homogeneidad entre las clasificaciones de las instancias para distinguir la diferencia entre los grupos que se estén estudiando. Se calcula como sigue:

$$Gini(q) = 1 - \sum_{l=1}^{L} p_l^2 \quad (7)$$

Por su parte el índice de clasificación errónea también actúa como una de las reglas de separación, midiendo una proporción de las instancias mal clasificadas cuando todas las instancias del nodo o raíz son asignadas a la clase con mayor cantidad. Matemáticamente podemos ver que se calcula como:

$$ErrorClas(q) = 1 - {}^{max} pl \quad (8)$$

Los árboles de decisión necesitan una definición para controlar su crecimiento en cuanto al número de hojas. Estos se entienden como criterios de detención, para detener el crecimiento de uno de estos elementos dentro del árbol. La aplicación de estos es necesaria para contra polar el

algoritmo recursivo. El nombre de estas técnicas de detención son pre-poda y poda del árbol. El crecimiento de un árbol excesivo es controlado por la pre-poda mientras que la cantidad de raíces generadas es controlada por la poda.

Desde la literatura se rescatan los criterios para la detección del crecimiento de árbol, es decir la aplicación de la pre-poda y poda. Estos criterios son tres. El primer criterio hace referencia al tamaño del nodo, es decir que sea consistente con la cantidad de instancia y el algoritmo determinara si el tamaño es inferior al límite definido. El segundo criterio es el nivel de pureza, esto es la proporción de instancias que pertenecen a la misma clase en el mismo nodo. Mientras mayor sea la cantidad de instancias de la misma clase mayor es la pureza. Se puede establecer entonces que el algoritmo se detenga en determinado nivel de pureza. El tercer punto hace referencia al mejoramiento, en el árbol se continua la segmentación mientras mejore el desempeño del modelo. Para esta tesis la evaluación será a través de la precisión.

Regresión Logística

La regresión logística es un tipo de regresión muy utilizada para predecir el resultado de una variable dependiente categórica. También se destaca su uso en el cálculo de probabilidades para predecir un evento en función de otros. Para describir la Regresión Logística podemos representarlo matemáticamente como sigue.

De acuerdo a la regresión logística, se hablará de la probabilidad de obtener determinado y dado determinado x se representará como $P(y|x)$, siendo x un vector que condiciona el resultado se tiene que la probabilidad posterior es como sigue en la formula (9).

$$P(y = 0|x) = \frac{e^{w'x}}{1+e^{w'x}} \qquad (9)$$

El algoritmo entonces encuentra los coeficientes apropiados para w' en forma iterativa, bajo el método de máxima verosimilitud. Por su naturaleza de cálculo al igual que las regresiones pueden verse afectadas por problemas de multicolinealidad y sesgo. Se itera para identificar el mejor conjunto de valores para w'. Ahora bien regresiones logísticas al igual que SVM puede identificar el costo del error de las clasificaciones erróneas, esto le permite minimizar el costo del error.

4.4.3 Clusterización

Una muestra de datos puede tener agrupaciones características en sí misma que permiten separar el grupo de muestra inicial en dos o más grupos más pequeños. Esto permite generar un modelo predictivo para cada cluster y así adaptarse mejor (Agrawal, 1998) a cada grupo para obtener mejores predicciones. El proceso de clusterización consiste en separar una base de datos en dos

o más grupos diferentes. Este proceso es ejecutado sin supervisión es decir que antes de terminado el proceso de clusterización no se conocen las variables finales de asignación de cluster o grupos para las instancias. El objetivo es crear grupos cercanos entre sí, es decir similares, pero sin saber de antemano cuales serán y que características tienen. Cada uno de estos grupos o cluster comienza a agrupar las instancias de estudio en torno a un centroide.

La asignación entonces de un cluster cada instancia puede ser expresado como un problema de minimizar la distancia entre las asignaciones, es decir que las instancias cercanas entre si se encontraran en el mismo cluster. La distancia d puede ser calculada como la distancia euclidiana.

$$d(x, y) = \sqrt{\sum_{i=1}(x_i - y_i)^2} \quad (10)$$

Utilizando el concepto de distancia euclidiana matemáticamente un algoritmo de clusterización identifica n segmentos en un conjunto de de X observaciones. Para cada grupo se identifica un centroide, como el punto alrededor del cual se agrupan las observaciones. Se define y_j como la asignación de una observación a un centroide, x_i la asignación de una observación como centroide y d_{lj} como la distancia entre la i-ésima y la j-ésima observación. La identificación de los centroides se obtiene a través de resolver el siguiente problema de programación líneal:

$$\min_{x} \sum_{i,j \epsilon I} y_l d_{lj}$$

$$s.a. \sum_{j \epsilon I} y_{lj} = 1, \quad i \epsilon I$$

$$. \sum_{i \epsilon I} x_l = N$$

$$y_{ij} \leq x_{i,} \qquad i \epsilon I$$

$$x_i, y_j \epsilon [0,1] \qquad (11)$$

En el problema de programación lineal se puede observar que dado un número de N segmentaciones se deben asignar un centroide a todas las observaciones. En la asignación se busca minimizar la distancia entre el centroide y las observaciones asignadas.

4.4.4 Técnicas de Combinación de modelos

Bagging

El concepto de Bagging aplica al área de Minería de datos para combinar predicciones de una clase desde el mismo tipo modelo con diferente data de aprendizaje. Es también utilizado para alcanzar estabilidad en la predicción cuando se aplican modelos complejos a grupos de datos relativamente pequeños. La forma de agrupación depende del tipo de predicción. Cuando se predice una clase, Bagging se comporta con votos por cada modelo, estos pueden ser ponderados según sea el caso, en la sección inmediatamente siguiente se revisa un algoritmo especializado en selección por votación. Un método para entregar una sola predicción desde variados modelos es utilizar las predicciones de cada modelo y aplicar un sistema de votación. Es decir que la clasificación final es la que ha sido predicha más veces por los diferentes modelos. Por otro lado, cuando se predice una variable continua se realiza un promedio de la predicción de los modelos. Si se tiene una tarea de Minería de datos para construir un modelo y el set de datos utilizado para entregar, donde se observa la clasificación es relativamente pequeño, se puede repetidamente rellenar una sub muestra con reemplazos para el nuevo set de entrenamiento y así entrenar el modelo con las muestras sucesivas. Como ejemplo árboles de decisión. En la práctica árboles muy diferentes serán construidos de estas diferentes muestras, lo que ilustra el problema de inestabilidad en la predicción. La Figura Nº 9 presenta este proceso.

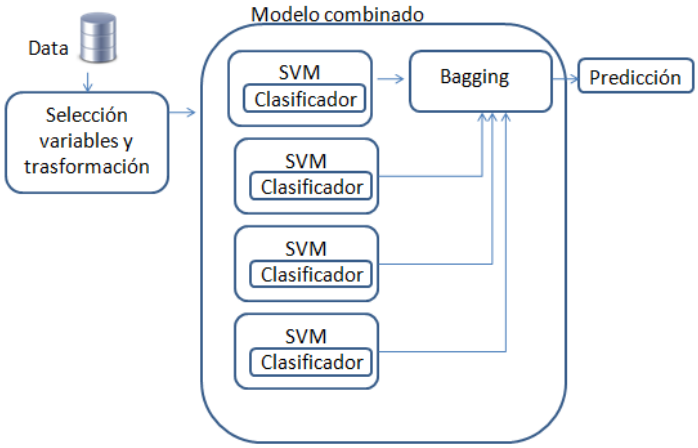


Figura Nº 9. Modelo combinado utilizando Bagging

Esta combinación de modelos resulta tener más flexibilidad, pero esto en teoría podría generar un sobre ajuste del modelo sobre los datos de entrenamiento que un modelo sencillo. Pero en la práctica las técnicas de combinación de modelos reducen los problemas de over-fitting en el entrenamiento de los modelos. Empíricamente la combinación de modelos entrega mejores resultados cuando hay una diversidad significativa entre los modelos, muchos métodos de combinación buscan promover la diversidad entre los modelos que combinan. Sin embargo,

utilizar una combinación de fuertes técnicas de aprendizaje de Minería de datos ha mostrado ser más efectivo que usar técnicas que buscan simplificar los modelos para promover la diversidad.

Stacking

Los procesos de Minería de datos tienen diferentes tipos de modelos de aprendizaje asociados como fueron revisados en la sección 4.4.1, cada tipo de modelo tiene fortalezas y debilidad. Para aprovechar en su totalidad las fortalezas de cada modelo es posible combinarlos entre sí. El proceso de Stacking es una forma de combinar múltiples modelos que introduce el concepto de meta-aprendizaje. A diferencia de Bagging se utiliza para combinar modelos de diferente tipo. El procedimiento para hacerlo es como sigue:

1. Separar el set de entrenamiento en dos set diferentes.

2. Entrenar diferentes modelos en la primera parte del set de datos.

3. Probar el desempeño de esos modelos en la segunda parte del set de datos.

4. Se utilizan las predicciones del paso 3 como los inputs y las respuestas correctas como los outputs para entrenar un aprendizaje de alto nivel.

La diferencia entre estos pasos y el proceso de validación cruzada es que el enfoque no es una obtener una sola mejor predicción. El enfoque es obtener el mejor desempeño aprendiendo de los diferentes modelos al mismo tiempo. Es decir que se combina el desempeño de los modelos en la etapa de prueba en vez de elegir uno de ellos. Para producir este aprendizaje desde muchos modelos con resultados más pobres es necesaria mayor computación de resultados y evaluaciones que la predicción de un solo modelo, siendo muy demandante en términos de recursos de software.

El resultado del entrenamiento de modelos combinados puede ser representado como una hipótesis. Esta hipótesis sin embargo, no necesariamente está contenida en base del espacio de hipótesis del modelo en que fue construida, es decir puede alcanzar una nueva hipótesis que aplique a nuevos datos previniendo over-fitting.

Los resultados de la combinación de modelos por Stacking son entonces hipótesis que pueden ser representadas bajo el teorema de Bayes. Recordando este teorema, sea $\{ A_1, A_2, \ldots, A_n \}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la siguiente Fórmula (12):

$$P(A_i|B) = \frac{P(B|A_i)*P(A_i)}{P(B)} \qquad (12)$$

Así también puede ser utilizado para calcular la influencia de diferentes variables en un determinado resultado. Por lo anterior, el algoritmo que se utiliza para agrupar los resultados de modelos de diferente tipo es el de Bayes Ingenuo, o en ingles Naive Bayes. Bayes Ingenuo es un clasificador en base a simple probabilística aplicando el teorema de bayes con fuerte independencia de supuesto, por eso lleva el nombre de ingenuo. Naive Bayes entonces es representado por:

$$Y_{NaiveBayes} = \underset{y_j \epsilon V}{Max} P(y_j) \prod_i P(x_i|y_j) \qquad (13)$$

Es decir, que asume la presencia o ausencia de una clase en particular (por ejemplo, una variable como la edad) no está relacionada con la presencia de otra variable (por ejemplo, el tipo de egreso del colegio). Considera todas las variables en forma independiente. Este proceso calcula entonces la media y las varianzas de cada una de las variables necesarias para la clasificación. La Figura Nº 10 presenta el flujo de combinación de modelos para la predicción.

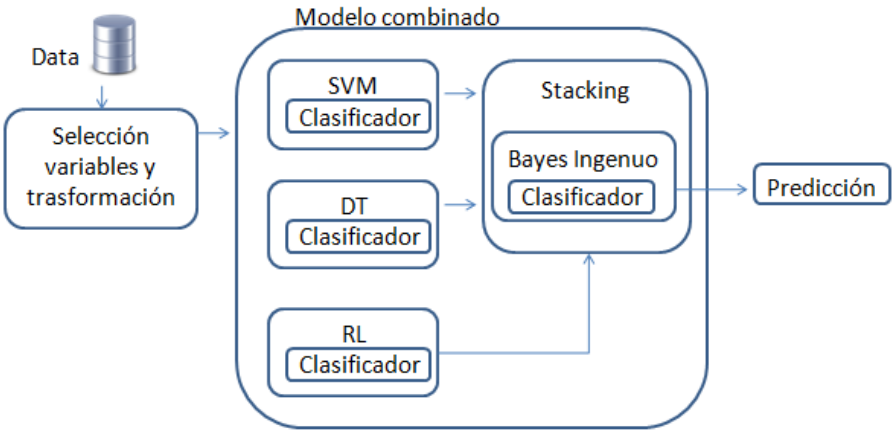


Figura Nº 10. Modelo combinado utilizando Stacking

4.5 Interpretación y evaluación

Esta etapa corresponde al análisis de los resultados del modelo. En esta etapa se observan los desempeños del modelo, principales resultados y patrones o findings que agreguen valor según el objetivo del estudio. Es importante destacar, que el proceso revisado de KDD puede volver a

ser ejecutado desde el comienzo o bien corregir una de las partes del proceso en pos de los objetivos del estudio.

El conocimiento del investigador tiene un papel muy importante. Debido a que el juicio experto debe aplicarse caso a caso para evaluar si los patrones extraídos son útiles según el contexto del estudio, y para esta tesis, principalmente en la aplicación de los resultados obtenidos.

Lo anterior, se relaciona como vimos con que el desempeño de un modelo depende del objetivo del investigador. En cuanto a desempeño, lógicamente un modelo predictivo tiene buen desempeño cuando efectivamente es capaz de registrar con antelación un resultado antes de que ocurra realmente. En esta tesis, es predecir que es un Postulante Destacado antes de que el estudiante demuestre que lo sea. Si la predicción es errónea puede ser catalogada en dos tipos. Se predice Postulante No Destacado cuando en realidad era un Postulante Destacado, fenómeno conocido como Error tipo I. La segunda categoría es cuando se predice Postulante Destacado para su Captación cuando en realidad era un Postulante No Destacado, fenómeno conocido como Error tipo II.

4.5.1 Indicadores de Desempeño

Ahora bien para analizar el desempeño se definen métricas que son capaces de evaluar que tanto predice o no un modelo. Las métricas evalúan en términos numéricos el desempeño de los modelos de predicción. Para medir el desempeño del modelo se utilizaran la precisión de la predicción, matriz de confusión, la validación cruzada. La cual es una herramienta ampliamente validada y utilizada en aplicaciones de Minería de datos.

Esta tesis se desarrolló en torno de una variable binomial o dicotómica, Postulante Destacado o Postulante No Destacado. Es decir que existen dos clases las que es bastante común asignarles tipos positivo y negativo de acuerdo al objetivo del estudio. La matriz de confusión permite observar cuando una clasificación positiva es acertada y se conoce como True Positive (TP), así como también cuando una clasificación positiva en realidad era una instancia negativa, se conoce como False Positive (FP). Esto también aplica a las clases negativas, por su parte cuando se predice una clase negativa y realmente era negativa se conoce como True Negative (TP). Caso contrario cuando se predice una clase negativa pero en realidad corresponde a una clase positiva se conoce como False Negative (FN).

Precisión de la Predicción

El primer indicador corresponde al desempeño general del modelo de predicción, este considera los errores en la predicción de una clase tan importantes como la predicción de otra clase. En términos de costo, quiere decir que la métrica no consideran los costos de los diferentes errores. Está métrica tiene dos grandes componentes, la primera es la Precisión de la Predicción (en

inglés Accuracy) mientras que el error de clasificación es la segunda herramienta generalmente utilizada. La precisión y el error son herramientas que permiten medir entonces cuanto se equivoca el modelo predictivo, en dos ejes de exactitud y costo.

Las operaciones matemáticas entre los indicadores de acertividad por tipo de clasificación por instancia y error anteriores permiten análisis y generar nuevas métricas de evaluación. El índice más común y típicamente utilizado es el de precisión, la Precisión de la Predicción se calcula como muestra la Formula Nº 14.

$$Precisión = \frac{TP+TN}{TP+FP+TN+FN} \quad (14)$$

Matriz de Confusión

La matriz de confusión es una tabla resumen de los resultados de desempeño de las predicciones del modelo implementado en determinada técnica de Minería de datos. La tabla tiene dos filas y dos columnas. Las filas representan las instancias que el modelo predijo en una clase u otra, mientras que las columnas representan las instancias que realmente existen en la base de datos utilizada.

Esta mide entonces la acertividad en general del modelo para ambas clases, como se aprecia tanto para los True Positive como para los True Negative. Existe otro indicador de interés para esta tesis. Como es posible combinar estos elementos, se utilizará el ratio conocido como verdadero positivo. Esta ratio permite evaluar la acertividad pero enfocándose en la clase de interés del estudio, es decir la positiva. Se calcula como sigue:

$$Ratio\ Verdadero\ Postitivo = \frac{TP}{TP+FP} \quad (15)$$

Es decir que es importante la clasificación general del modelo pero también es muy útil saber si el modelo apunta al objetivo de predecir la clase de interés para la aplicación de la técnica de minería de datos en la realidad.

Validación Cruzada

En las técnicas de minería de datos se busca optimizar el modelo para que se ajuste a los datos de entrenamiento tan bien como sea posible. Pero si probamos el modelo en data independiente, es decir que no ha sido utilizada para el entrenamiento de los datos, puede que el modelo no se desempeño tan bien como en los datos en los que fue generado. Esto se conoce como over-fitting. El proceso de validación cruzada permite el ajuste del modelo a datos de prueba

hipotéticos. Este es especialmente útil cuando los datos de prueba no real no están presentes cuando se genera el modelo.

La validación cruzada consiste dividir los datos de entrada (set) en n set más pequeños de igual tamaño. De estos n set se retiene uno de ellos como set de prueba y validación. Los restantes n-1 sets son utilizados como datos de entrenamiento del modelo predictivo. El proceso de validación cruzada se repite n veces en donde cada set es utilizado una vez como set de prueba. Los k resultados de las k iteraciones son luego promediados para producir una estimación única. Este proceso se aprecia en la Figura Nº 11.

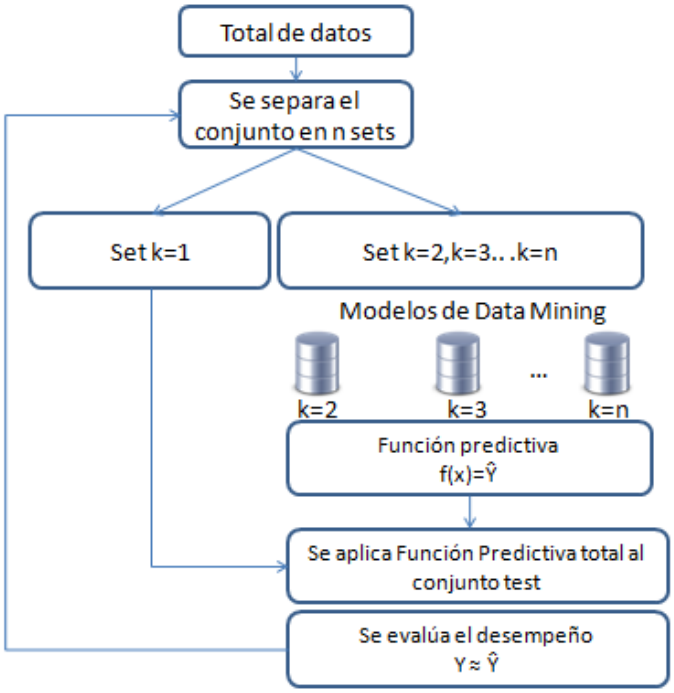


Figura Nº 11. Diagrama explicativo Validación Cruzada

Capítulo 5 – Enfoque de Solución

El enfoque de solución de esta tesis es la aplicación de la metodología descrita en el Capitulo 4, Knowledge Discovery in Databases – KDD para predecir a través de modelos de minería de datos si cumple con determinado perfil o no. Ahora bien, se utilizan cuatro enfoques en la aplicación de estos modelos. El primer enfoque es singular para cada modelo, donde se generan predicciones utilizando los resultados de cada modelo por separado. El segundo enfoque es secuencial, es decir se utilizan los resultados de un segundo modelo después de aplicados los resultados de un primer modelo. El tercer enfoque de modelos mezcla los resultados de modelos del mismo tipo. Finalmente el cuarto enfoque por su parte, considera modelos combinados, aplicando combinatorias de modelos y técnicas de minería de datos secuenciales para mejorar los resultados de la predicción.

5.1 Enfoques para la creación de modelos predictivos

Primer enfoque: Modelo Singular

El primer enfoque consiste en la construcción de un modelo predictivo basado en una técnica de predicción utilizando la metodología del KDD. Continuando con la aplicación del proceso de KDD corresponde entonces ejecutar las técnicas de Minería de datos para extraer la información oculta entre los datos. El objetivo de esta tesis es predecir a través de un modelo predictivo de postulantes destacados previo al momento de su real postulación a la universidad. Por lo tanto se generan 3 modelos singulares en este enfoque. Un modelo para Support Vector Machine (SVM), Regresión Logística (RL) y Árbol de Decisión (DT). Esto quiere decir que se ejecuta una vez el proceso en forma separada para cada modelo y sus resultados son evaluados a través de los indicadores de desempeño revisados en el Capítulo 4.

Segundo enfoque: Modelos secuenciales

El segundo enfoque corresponde a la aplicación en conjunto de dos o más técnicas de minería de datos en serie para mejorar los indiciadores de desempeño. En particular se utilizan como primer modelo técnicas de Decision Tree, Clusterización y Sequential Backward Elimination para luego aplicar técnicas como SVM y RL utilizando Grid Search, es decir ajustando los parámetros de las técnicas para obtener mejores resultados de predicción. Las combinaciones de estas técnicas entregan 48 modelos posibles. En anexo 5 se encuentra descrito cada modelo en serie.

Las técnicas de selección de variables y cluster pueden utilizarse como no utilizarse en un modelo en serie. Por lo tanto ocurre una combinación de modelos predictivos en serie diferentes. Las técnicas de Minería de datos que se realizan en serie en estos modelos, son: la aplicación de cluster o no, el uso de árbol de decisión para seleccionar variables o no, aplicar Sequential Backward Elimination o no, la aplicación de un clasificador o no y finalmente tres técnicas de aprendizaje no supervisado. A su vez estas técnicas tienen sus propios parámetros, los que agregan mayor cantidad de combinaciones al cambiarlos entre sí. Se detallan a continuación en la Tabla Nº 3 las diferentes combinaciones de estos modelos híbridos.

Etapa	Técnica	Aplicación	Descripción
Primera etapa	Selección de Variables	Sequential Backward Elimination	Selección de variables utilizando Sequential Barckward Elimination
		Sin SBE	No se aplica.
	Selección de Variables	Decision Tree	Selección de variables utilizando DT previo

		Sin Decision Tree	No se aplica.
	Cluster	Cluster	El proceso se ejecuta identificando n-centroides y asigna las instancias a estos centroides.
		Sin Cluster	No se aplica.
	Clasificador	Clasificador	El proceso se ejecuta asignándoles costos diferentes al error tipo I y error tipo II, estos costos son los First y Second Cost.
		Sin Clasificador	No se aplica.
Segunda etapa	Aprendizaje	Support Vector Machine	Técnica basada en vectores, donde C corresponde al error de clasificación.
		Árboles de Decisión	Técnica de árboles de decisión, la profundidad indicada cuantos nodos o hojas puede tener una rama.
		Regresión logística	Técnica basada en regresiones, donde C corresponde al error de clasificación.

Tabla Nº 3. Modelo de Enfoque secuencial.

Para describir la secuencia de aplicación de las técnicas de Minería de datos en el enfoque secuencial se utiliza la siguiente Figura Nº 12.

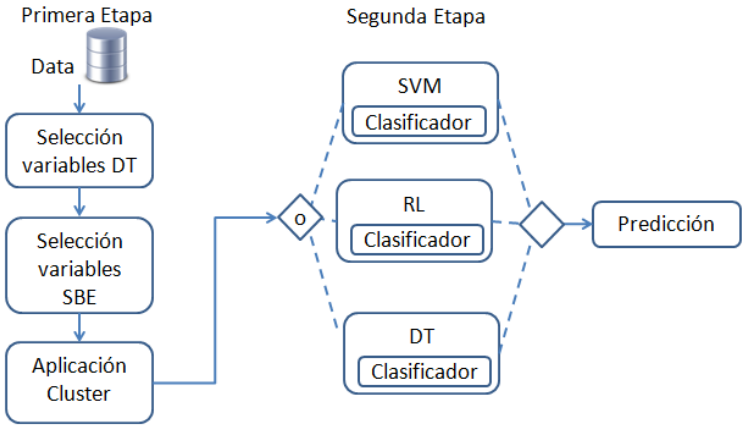


Figura Nº 12. Secuencia de aplicación en Enfoque Secuencial

Tercer enfoque: Modelos Combinados

El tercer enfoque de solución consiste en combinar los resultados de múltiples modelos del mismo tipo utilizando Bagging. Esto permite combinar los resultados de predicción de modelos del mismo tipo para obtener mejores resultados, mediante votación. Es decir que cada resultado de predicción se convierte en un voto y luego se cuentan para obtener una predicción. Como se revisó en el Capítulo 4, esta técnica favorece las predicciones en modelos del mismo tipo que arrojan predicciones diferentes. Se combinan entonces cada uno de tipos de los modelos del enfoque singular por separado, es decir una combinación de resultados de SVM, una combinación de RL y una combinación de RL. La Figura Nº 13 ejemplifica este enfoque de solución para SVM.

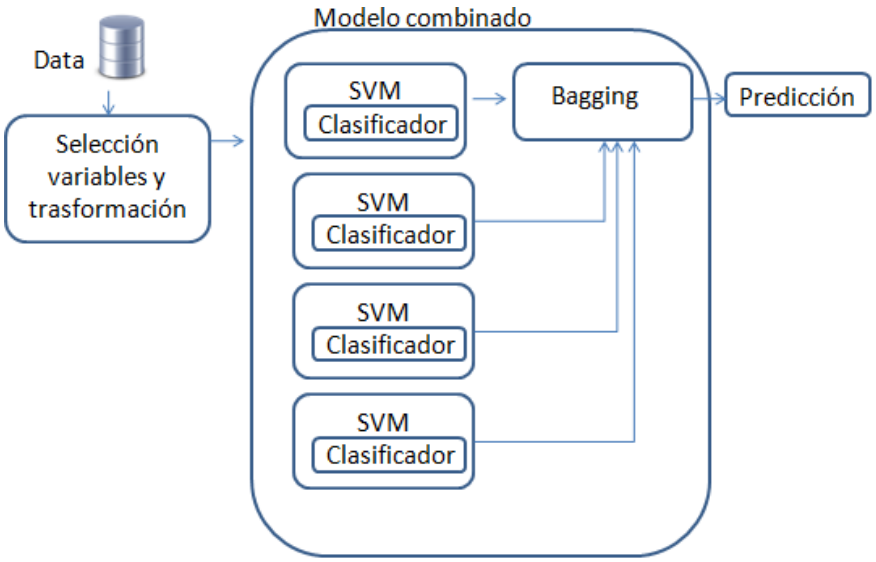


Figura Nº 13. Secuencia de aplicación en Enfoque de Modelos Combinados

Cuarto enfoque: Modelos Combinados Secuenciales

El cuarto enfoque consiste en ponderar las predicciones de diferentes modelos, del mismo o diferente tipo para obtener mejores resultados de predicción. Además de la aplicación de modelos en serie se acumulan los resultados de diferentes modelos del mismo tipo utilizando Bagging así como también se combinan los resultados de modelos diferentes tipos mediante Stacking en un proceso de aprendizaje combinado mediante un modelo Naive Bayes. Cabe destacar que el proceso de Stacking puede combinar modelos creados utilizando Bagging. Es decir se utiliza la combinación del tercer enfoque y el modelo secuencial del segundo enfoque, esto es mediante una etapa de combinación utilizando Stacking. En la tabla Nº 4 se detalla la aplicación de este enfoque.

Etapa	Técnica DM	Aplicación	Descripción
Primera Etapa	Selección y Cluster		
Segunda Etapa	Clasificador y Aprendizaje		
Combinación	-Stacking	Aplicar Stacking	Se combinan modelos utilizando Stacking. Se pueden combinar uno, dos o tres modelos de la segunda etapa.
		Sin Stacking.	No se aplica.
	-Bagging	-Support Vector Machine	Se agrupan resultados de SVM utilizando Bagging.
		-Árboles de Decisión	Se agrupan resultados de DT utilizando Bagging.
		-Regresión logística	Se agrupan resultados de RL utilizando Bagging

Tabla Nº 4. Técnicas utilizadas en cuarto enfoque de solución de modelos combinados secuenciales.

El cuarto enfoque de solución contempla un entrenamiento de los datos muy exigente en cuanto a la capacidad de procesamiento computacional y al uso de recursos. Para explicar la secuencia y lógica de aplicación de las técnicas de Minería de datos de este enfoque se presenta la Figura Nº 14, que ha modo de ejemplo muestra la combinación mediante Stacking de tres modelos que utilizan Bagging.

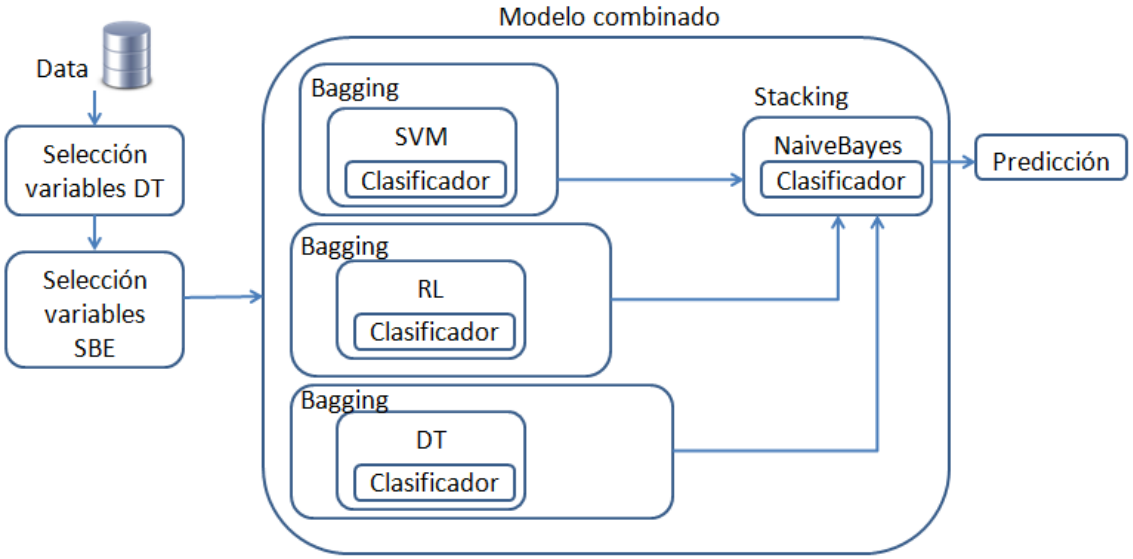


Figura Nº 14. Secuencia de aplicación para modelos combinados secuenciales.

Capítulo 6 – Caso de estudio

6.1 Descripción del caso

El experimento realizado en esta tesis fue realizado en la Facultad de Economía y Negocios (FEN) de la Universidad de Chile utilizando los datos de los estudiantes entre 2004 y 2010. Las carreras de esta facultad se agrupan bajo dos escuelas que son La Escuela de Economía y Administración con la carrera de Ingeniería Comercial la cual tiene dos menciones de Administración y Economía, y a su vez la Escuela de Sistemas de Información y Auditoria que tiene dos carreras Ingeniera en Información y Control de Gestión y Contador Auditor. En esta tesis se considerarán los estudiantes de las tres carreras, dado que se apunta a un perfil de ingreso como Facultad, en estudios posteriores es posible continuar segmentando para obtener un perfil particular de ingreso.

Las tres carreras tienen una malla curricular de 10 semestres en 5 años que contempla materias de diversa índole de negocios. Las tres carreras participan del proceso normal de selección universitaria vía PSU, pero también existen otros tipos de ingreso de estudiantes a FEN. El foco de esta tesis será en los estudiantes que ingresan vía PSU, señalizando sus puntajes, desempeño escolar y datos socio-económicos. Existen distintos tipos de cupos de ingreso según tipo, estos ingresos pueden ser por:

1. BEA 5% Superior: Corresponde a los mejores estudiantes del colegio y su generación
2. Deportivo: Son cupos especiales para deportistas destacados
3. PSU: Son los cupos normales para estudiantes por ingreso de selección universitaria.
4. Transferencia: Son estudiantes que ingresan utilizando su característica de ex estudiante en otra casa de estudio o carrera.
5. Convenio Isla de Pascua: Estudiantes provenientes de la isla de pascua.

Los estudiantes que cursan o cursaron la universidad en FEN tienen estados académicos. Estos estados académicos corresponden a su situación en cuanto al plan de estudio que están cursando. El sistema de administración docente identifica los siguientes estados:

1. Titulado: Son los estudiantes que obtienen el grado académico y todos los ramos.
2. Egresado: Es el estudiante con los ramos aprobados y en condiciones de dar el examen de grado.
3. Regular: Son los estudiantes activos académicamente.
4. Postergado: El estudiante está matriculado pero congelado académicamente.
5. Eliminado: El estudiante fue removido del plan de estudios por motivos académicos o disciplinarios.
6. Renuncia: El estudiante deja el plan de estudios por voluntad propia.

7. Transferido: El estudiante renuncia al programa y se cambia a otro de la misma facultad.

Los datos utilizados en esta tesis son de estudiantes con estados académicos de Titulado, Egresado, Eliminado o Regular.

6.2 Análisis de la admisión por PSU

Se revisa a continuación las características de los estudiantes que han pasado por el proceso de admisión vía PSU, con el objeto se comprender su variabilidad o no en el tiempo. En este proceso se otorga mayor ranking de ingreso a los estudiantes con los puntajes más altos luego de aplicar un ponderador de puntaje definido. El resultado de esa ponderación permite generar un ranking entre todos los postulantes y se ingresa a la carrera postulada desde el estudiante con más ranking hasta el menor, proceso que ocurre hasta llenar el cupo universitario por carrera. La universidad de Chile exige para postular vía PSU un puntaje ponderado mínimo de 600 puntos para todas sus carreras, independientemente de lo que pueda exigir en sus sistemas especiales de admisión (Fuente http://www.uchile.cl, consultada el 22 de abril de 2017).

La asignación de ingresos según ranking ordena los estudiantes por su ponderación lo que permite obtener información estadística relevante a la hora de hablar de ingresos vía PSU. Los conceptos de primer matriculado y último seleccionado son bastantes útiles para evidencia el rango de puntajes ponderados en los que se encuentran los estudiantes de una carrera. Se hace distinción con el concepto de último matriculado, puesto que ocurre después de la asignación de carrera por ranking inicial. Estos ocurren por estudiantes que se retractan y abren cupo, esta proceso se conoce coloquialmente como lista de espera. En el caso de FEN, podemos ver los tramos para los años del estudio en la Tabla Nº 5 según carrera.

	Ingeniería Comercial			**IICG y Contador Auditor**		
Año	**Primer Matriculado**	**Último Seleccionado**	**Último Matriculado**	**Primer Matriculado**	**Último Seleccionado**	**Último Matriculado**
2004	801,3	680,3	680,3	778,2	654,5	654,5
2005	810	681,5	679	745,4	647,4	647,4
2006	795,1	690,6	690,2	750,0	662,4	662,6
2007	801,9	698,1	698,1	790,1	675,4	675,4
2008	816,9	702	702	768,2	677,8	677,8
2009	796,6	708,5	708,1	758,4	684,0	684,0
2010	823,8	707,1	707,1	803,6	673,0	673,0

Tabla Nº 5. Primer Matriculado, último matriculado y último seleccionado por año en FEN.
Fuente: <http://www.uchile.cl>, consultada el 10 de mayo de 2017.

Mientras que si observamos los resultados de las pruebas de admisión y promedios de la enseñanza media de los estudiantes, podemos ver que se mantienen constantes a lo largo de los años de esta estudio, con leves alzas año a año o manteniendo el resultado del año anterior. Lo anterior es fundamental para sustentar que las aplicaciones de minería de datos sean útiles en futuras generaciones de estudiantes dado que la tendencia se mantiene constante. La Tabla Nº 6 muestra los resultados de puntajes presentados según tipo de prueba.

Año	Cantidad de Estudiantes	Promedio Notas Colegio	Puntaje Lenguaje	Puntaje Matemática
2004	252	6,25	683	717
2005	255	6,31	671	713
2006	275	6,36	665	715
2007	263	6,40	676	726
2008	260	6,40	682	729
2009	251	6,42	689	736
2010	293	6,35	687	755
Total	**1849**			
Promedio	264	6,35	679	727
Varianza	228	0,34	73	216
Coeficiente de Variación	0,86	0,005	0,10	0,29

Tabla Nº 6. Cantidad de estudiantes, promedio notas y puntajes, varianza y coeficiente de variación según año de proceso.

Se puede concluir que la admisión vía PSU se mantiene constante con leves alzas en cuanto al puntaje de ingreso ponderado, es decir puntaje del último seleccionado. Se destaca que ingeniería comercial mantiene sobre IICG y Contador Auditor en promedio 27,7 puntos del puntaje ponderado. Por su parte, los promedios de los estudiantes y, los puntajes de lenguajes y matemática se mantienen estables en los años de estudios de esta tesis.

6.3 Análisis estadístico del perfil de selección

En esta sección se describe el análisis estadístico del perfil de selección generado y sus variables. Para esta tesis se generó una clasificación de los estudiantes universitarios según el enfoque temporal descrito en el Capítulo 2. Estos estudiantes en su momento fueron postulantes, y se clasifican en relación al desempeño y su estado de situación académico. Luego se utiliza esta clasificación para entrenar los modelos de predicción según los tres enfoques revisados en el Capítulo 5. Los modelos de predicción son aplicados posteriormente en los nuevos

postulantes universitarios, para predecir si corresponden a un Postulante Destacado o no. En la tabla Nº 7 podemos ver los promedios a lo largo de los años contemplados en esta tesis.

Año Ingreso	Promedio 1er Año	Promedio 2do Año	Promedio 3er Año	Promedio 4to Año	Promedio 5to Año	Promedio entre 3er y 5to Año
2004	4,7	4,8	4,8	5,0	5,4	4,7
2005	4,8	4,6	4,7	5,0	5,4	4,7
2006	4,9	4,8	4,9	5,0	5,4	5,0
2007	4,9	4,8	4,8	5,0	5,4	4,9
2008	5,0	4,8	4,9	5,0	5,5	4,9
2009	5,0	4,9	4,9	5,1	5,5	5,0
2010	4,8	4,8	4,8	5,1	5,6	4,9
Total general	**4,9**	**4,8**	**4,8**	**5,0**	**5,5**	**4,9**

Tabla Nº 7. Promedios ponderados por año de estudios y año de ingreso.

En la tabla Nº7 se aprecia que los promedios de primer y segundo año son más bajos que los de cuarto y quinto año. Por su parte es importante aclarar que estos son promedios ponderados por el peso de cada ramo en cuanto a las unidades docentes o académicas que representa. Por lo anterior el promedio ponderado entre tercer año y quinto año no es igual al promedio simple.

Los estudiantes con sobre 5,5 en el promedio ponderado entre tercer y quinto año se convierten entonces en la clase positiva de esta tesis y el objetivo es detectar a estos estudiantes al momento de postular. Son estos los que llamaremos Postulantes Destacado según indica la dirección de escuela del caso de estudio, además se observa que corresponden al quintil superior. Los estudiantes (instancias) que no cumplan este perfil de selección serán marcados como Postulante No Destacado. Los Postulantes No Destacados son también los que desertan en forma académica e involuntaria de la carrera, es decir son expulsados por desempeño insuficiente. No se consideran los estudiantes que desertan voluntariamente. Podemos ver a continuación la distribución del total de estudiantes según su perfil de Postulante. La tabla Nº 8 muestra la cantidad y proporción de postulantes destacados recibidos por año, así como también los postulantes no destacados.

Año Ingreso	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado / Total
2004	58	194	252	23,0%
2005	51	204	255	20,0%
2006	44	231	275	16,0%
2007	57	206	263	21,7%

2008	70	190	260	26,9%
2009	65	186	251	25,9%
2010	66	227	293	22,5%
Total general	**411**	**1438**	**1849**	**22,2%**

Tabla Nº 8. Cantidad y proporción de Postulantes Destacados por año.

Los Postulantes Destacados fueron marcados utilizando su promedio de ramos universitarios en los últimos tres años, pero esta información no es lo único que los separa de los postulantes no destacados como se analiza a continuación en las variables de postulación universitaria.

Análisis de variables de postulación universitaria y desempeño académico

A continuación se revisan las diferencias de puntajes y promedio del colegio entre los estudiantes marcados como Postulante Destacado y los Postulante No Destacados. Las tablas Nº 9 resume las diferencias entre los promedios y puntajes de estos estudiantes al postular, esta diferencia fue calculada con la resta simple entre el puntaje promedio de los postulantes destacados menos el promedio de los postulantes no destacados.

	Puntaje NEM Promedio		**Promedio de Notas**	
Año Ingreso	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**
2004	711	660	6,4	6,2
2005	717	675	6,5	6,3
2006	719	689	6,5	6,3
2007	728	695	6,5	6,4
2008	733	690	6,6	6,3
2009	726	698	6,5	6,4
2010	730	682	6,5	6,3
Total general	**724**	**684**	**6,5**	**6,3**

	Promedio de Puntaje Lenguaje		**Promedio de Puntaje de Matemática**	
Año Ingreso	**Postulante Destacado**	**Postulante No**	**Postulante Destacado**	**Postulante No**

		Destacado		Destacado
2004	700	678	718	717
2005	686	668	708	714
2006	681	662	713	715
2007	701	669	727	725
2008	691	678	721	732
2009	698	685	731	737
2010	711	680	748	757
Total general	**696**	**674**	**725**	**728**

Tablas Nº 9. Continuación Diferencias de notas y puntajes promedio entre Postulantes Destacados y Postulantes No Destacados.

Se destaca de la tabla Nº 9 que los postulantes destacados tienen en promedio dos puntos más en su notas de enseñanza media, luego, eso se traduce en cerca de 40 puntos de puntaje NEM de ventaja de los postulantes destacados. Así como también, tienen una clara tendencia de presentar un puntaje de lenguaje en promedio de 23 puntos superior. El puntaje de matemática no muestra una tendencia clara.

De la misma forma, se calcula el promedio de notas y puntaje para cada colegio por cada año. Asociando cada postulante con su colegio, se presenta a continuación la diferencia promedio entre los colegios de un postulante destacado y otro que no. Las tablas Nº 10 resumen los resultados.

Año Ingreso	Promedio de Notas Colegio Postulante Destacado	Promedio de Notas Colegio Postulante No Destacado	Promedio de NEM Colegio Postulante Destacado	Promedio de NEM Colegio Postulante No Destacado
2004	5,85	5,83	591	589
2005	5,89	5,88	599	596
2006	5,76	5,86	584	594
2007	5,82	5,88	592	597
2008	5,87	5,86	597	592
2009	5,89	5,85	598	591
2010	5,92	5,85	604	596
Total general	**5,86**	**5,86**	**595**	**594**

Año Ingreso	Puntaje Lenguaje Postulante Destacado	Puntaje Lenguaje Postulante No Destacado	Puntaje Matemática Postulante Destacado	Puntaje Matemática Postulante No Destacado
2004	601	597	589	593
2005	592	593	599	600
2006	592	589	596	595
2007	607	604	611	610
2008	611	605	618	616
2009	615	609	620	616
2010	629	628	642	643
Total general	**608**	**604**	**612**	**611**

Tablas Nº 10. Notas y puntajes promedios entre los colegios de Postulantes Destacados y No Destacados.

En esta caso las diferencias no son tan pronunciadas como en el análisis por estudiante presentado anteriormente, esto se debe a que dos estudiantes pueden provenir del mismo colegio pero uno es marcado como destacado y el otro como no destacado.

Utilizando los dos tipos de información anterior, tanto el promedio del estudiante como el promedio del colegio, se genera una distancia simple restando cada promedio de estudiante con el promedio de su colegio. Luego se comparan las distancias de los postulantes destacados con los postulantes no destacados. Se observan los resultados en la Tablas Nº11.

Año Ingreso	Diferencias Notas Promedio Estudiante con Colegio Postulante Destacado	Diferencias Notas Promedio Estudiante con Colegio Postulante No Destacado	Diferencias NEM Estudiante con Colegio NEM Postulante Destacado	Diferencias NEM Estudiante con Colegio NEM Postulante No Destacado	Diferencias Puntaje Lenguaje Estudiante con Puntaje Lenguaje Colegio Postulante Destacado	Diferencias Puntaje Lenguaje Estudiante con Puntaje Lenguaje Colegio Postulante No Destacado
2004	0,6	0,4	120,7	70,9	99,4	80,3
2005	0,6	0,4	118,4	78,9	93,9	74,5
2006	0,7	0,5	134,9	95,0	89,1	73,0
2007	0,7	0,5	136,8	98,0	94,3	65,2
2008	0,7	0,5	136,2	97,8	80,9	73,4
2009	0,6	0,5	128,2	107,1	83,3	76,4
2010	0,6	0,4	126,1	86,3	82,4	52,1
Total	**0,6**	**0,5**	**128,9**	**90,4**	**88,5**	**70,3**

general

Tablas Nº 11. Diferencias entre variables del estudiante y variables del colegio.

En este caso se observa la misma tendencia que al comparar a los estudiantes por separado. Es decir que la distancia entre los estudiantes y su propio colegio de los Postulantes Destacados es tener por sobre al menos un decimal de nota de enseñanza media, con un puntaje NEM cerca de 40 puntos superior y un puntaje de lenguaje de al menos 18 puntos más alto en promedio. Las diferencias con las pruebas de matemáticas, ciencias e historia no son de niveles relevantes.

Ahora bien, si comparamos el desempeño académico universitario año a año de los Postulantes Destacados se observa que se mantienen en promedio de nota universitaria ponderada por unidades docentes (UD) por encima del desempeño de los Postulantes No Destacados. Se agregan también los cálculos de las diferencias del promedio para los dos primeros años, como también las diferencias de los promedios de notas ponderadas por UD de los últimos tres años. Se observa esto en la Tabla Nº 12.

	Promedio de 1er y 2do Año		**Promedio de 3er, 4to y 5to año**		**Promedio de 1er Año**		**Promedio de 2do Año**	
Año Ingreso	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**
2004	5,29	4,52	5,74	4,48	5,24	4,48	5,33	4,29
2005	5,23	4,57	5,70	4,51	5,33	4,64	5,12	4,36
2006	5,37	4,72	5,71	4,88	5,39	4,76	5,35	4,60
2007	5,37	4,64	5,76	4,64	5,39	4,69	5,35	4,45
2008	5,37	4,70	5,77	4,62	5,41	4,82	5,34	4,50
2009	5,37	4,76	5,72	4,85	5,38	4,82	5,35	4,62
2010	5,34	4,56	5,74	4,54	5,34	4,61	5,34	4,44
Total general	**5,34**	**4,64**	**5,74**	**4,65**	**5,35**	**4,69**	**5,32**	**4,47**

	Promedio de 3er Año		**Promedio de 4to Año**		**Promedio de 5to Año**	
Rótulos de fila	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**	**Postulante Destacado**	**Postulante No Destacado**
2004	5,45	4,11	5,63	4,13	6,10	4,43
2005	5,37	4,15	5,60	4,15	5,81	4,45

2006	5,46	4,60	5,55	4,61	6,07	4,93
2007	5,45	4,30	5,50	4,31	6,09	4,66
2008	5,56	4,25	5,65	4,42	6,02	4,64
2009	5,46	4,52	5,60	4,67	6,05	4,83
2010	5,40	4,24	5,68	4,17	6,05	4,45
Total general	**5,45**	**4,31**	**5,61**	**4,35**	**6,03**	**4,63**

Tablas Nº 12. Diferencias de promedios universitarios ponderador por unidades docentes entre Postulantes Destacados y Postulantes No Destacados.

Se observa una marcada tendencia de los estudiantes destacados de mantener un promedio ponderado más alto. La tendencia es un poco menor los dos primeros años debido al periodo de ajuste al medio universitario, pero aún así es sobre seis décimas en promedio. Pero luego, en los últimos tres años se incrementan por sobre un punto las diferencias entre los Postulantes Destacados y los no destacados. Esta diferencia ocurre principalmente porque muchos postulantes no destacados reprueban asignaturas o cambian de estado académico, como ser eliminados.

Análisis de características del colegio de los postulantes

Las características del colegio resultan importantes para su estudio puesto que los colegios forman a los estudiantes, esta formación según fue revisado en el Capítulo 3 no sólo importa en el conocimiento académico si no también en las herramientas para adaptarse al nuevo medio universitario. En cuanto a las variables relacionadas con el contexto del colegio del estudiante encontramos la el gráfico Nº 1, que muestra los porcentajes de la cantidad de instancias por región (En la tabla por número de región) entre Postulantes Destacados menos los Postulantes No Destacados.

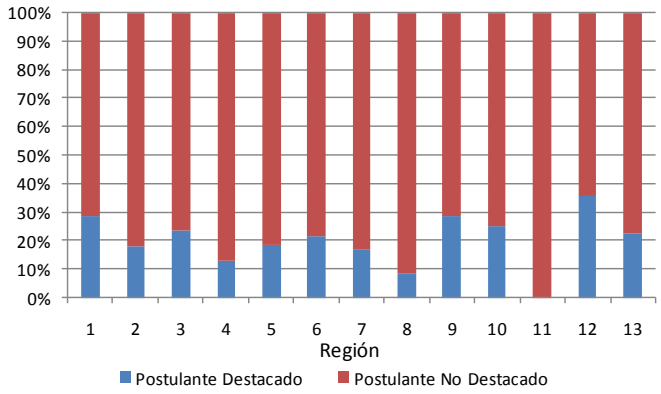


Gráfico Nº 1. Porcentajes de postulantes destacados y no destacados por región.

Las diferencias por región se acentúan en la región metropolitana, doceava, novena y primera región, si se considera que el porcentaje de postulantes destacados para esta muestra de datos es de un 22% en promedio. En la onceava región no se observar postulantes destacados. Se ven diferencias menores entre la séptimo y octava región a favor de los Postulantes No Destacados. Por su parte las diferencias de acuerdo al régimen (sexo) del colegio son en promedio casi un 3% de mayor presencia de Postulantes Destacados para los colegios de régimen Femenino aunque con bastante variación en los últimos años. Lo siguen los colegios mixtos. La Tabla Nº 13, muestra esta relación que se mantienen medianamente constante a lo largo de los años.

Año Ingreso	Régimen Colegio	Postulante Destacado	Postulante No Destacado	Total general	% Postulantes Destacados
2004	Masculino	9	48	57	16%
	Femenino	15	37	52	29%
	Mixto	34	109	143	24%
2005	Masculino	11	47	58	19%
	Femenino	10	34	44	23%
	Mixto	30	123	153	20%
2006	Masculino	6	42	48	13%
	Femenino	8	30	38	21%
	Mixto	30	159	189	16%
2007	Masculino	12	44	56	21%
	Femenino	8	21	29	28%
	Mixto	37	141	178	21%
2008	Masculino	10	44	54	19%
	Femenino	13	22	35	37%
	Mixto	47	124	171	27%
2009	Masculino	10	35	45	22%
	Femenino	3	26	29	10%
	Mixto	52	125	177	29%
2010	Masculino	9	51	60	15%
	Femenino	10	20	30	33%
	Mixto	47	156	203	23%
Total general		**411**	**1438**	**1849**	

Tabla Nº 13. Porcentajes de participación según régimen.

Ahora bien si observamos las diferencias de participación entre postulantes destacados y no destacados en base al grupo de dependencia del colegio, nos encontramos con que los estudiantes de colegios particulares concentran en promedio un 2% más de Postulantes Destacados que los colegios municipales. Sin embargo siendo la proporción general de Postulantes Destacados un 22%, los colegios particulares están sólo apenas por sobre esta proporción. Ahora bien en la distribución interna de Postulantes Destacados se ve que los provenientes de colegios particulares son más del doble respecto del total de la muestra. La tabla Nº 14 evidencia estas diferencias. La tendencia por año puede ser revisada en Anexos.

Dependencia	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado sobre Total Dependencia	% Postulante Destacado sobre Total	% Postulante No Destacado sobre Total
Particular	219	729	948	23%	12%	39%
Subvencionado	112	405	517	22%	6%	22%
Municipal	80	304	384	21%	4%	16%
Total general	**411**	**1438**	**1849**		**22%**	**78%**

Tabla Nº 14. Cantidad de Postulantes Destacados y Postulantes No Destacados según grupo de dependencia del colegio.

Por su parte la rama del colegio, humanista o técnica, no presenta diferencias importantes de participación, es sólo un 1% más orientado a los colegios técnicos. Pero si es un precedente puesto que los colegios técnicos tienen mucha menos participación en la vía de ingreso PSU.

En resumen por tanto, los colegios de la región metropolitana concentran ligeramente más Postulantes Destacados de hasta un 4% más en promedio, los que además aumenta un 2,2% de participación de esta clase en el caso de colegios mixtos y un 3% en el caso de los colegios particulares pagados. Por su parte los colegios de régimen Femeninos son los que más aportan estudiantes destacados en un tendencia que se mantiene en el tiempo excepto en un año. Mientras que los colegios particulares concentran más Postulantes Destacados con sobre 3% los otros colegios, son también los que más estudiantes aportan en general, por lo que se destaca que los Postulantes Destacados de colegios Municipales sean casi un 3% de menor proporción que el total general.

Análisis de variables de contexto familiar y socio demográficas

Según lo indicado por la literatura las variables del ambiente familiar, como la relación con los padres, la educación de ellos o niveles de ingreso; así como las condiciones socio demográficas tienen relación directa con los desempeños académicos. Se describen a continuación las diferencias porcentuales entre los Postulantes Destacados y los Postulantes No Destacados.

La primera variable a comentar es el sexo, que presenta una concentración mucho mayor en las mujeres para la clase positiva, en promedio las mujeres son 14% más Postulante Destacado que los hombres. Además para cada año de ingreso las mujeres Postulante Destacado están al menos un 5% por sobre la tendencia general para su año de ingreso. La tabla Nº 15 explica la distribución por año.

Año Ingreso	Sexo	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado sobre total	% Postulante No Destacado sobre total
2004	Masculino	24	122	146	16%	84%
	Femenino	34	72	106	32%	68%
Total 2004		**58**	**194**	**252**	**23%**	**77%**
2005	Masculino	22	119	141	16%	84%
	Femenino	29	85	114	25%	75%
Total 2005		**51**	**204**	**255**	**20%**	**80%**
2006	Masculino	16	151	167	10%	90%
	Femenino	28	80	108	26%	74%
Total 2006		**44**	**231**	**275**	**16%**	**84%**
2007	Masculino	26	131	157	17%	83%
	Femenino	31	75	106	29%	71%
Total 2007		**57**	**206**	**263**	**22%**	**78%**
2008	Masculino	31	124	155	20%	80%
	Femenino	39	66	105	37%	63%
Total 2008		**70**	**190**	**260**	**27%**	**73%**
2009	Masculino	35	109	144	24%	76%
	Femenino	30	77	107	28%	72%
Total 2009		**65**	**186**	**251**	**26%**	**74%**
2010	Masculino	26	150	176	15%	85%
	Femenino	40	77	117	34%	66%
Total 2010		**66**	**227**	**293**	**23%**	**77%**
Total general		**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 15. Distribución por sexo y tipo de postulante.

Con respecto a la edad, los postulantes con 19 años tienen un 3% más de proporción para la clase positiva. Mientras que los 21 años y mayores, la proporción de clase positiva es muy baja, menos del 10%. La tabla Nº 16 muestra estos resultados, por simplicidad se agruparon edades con pocas instancias. En Anexos se puede observar la tendencia por año de ingreso.

Edad	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado	% Postulante No Destacado
18	8	16	24	33,3%	66,7%
19	278	818	1096	25,4%	74,6%
20	101	341	442	22,9%	77,1%
21	10	112	122	8,2%	91,8%
22	5	63	68	7,4%	92,6%
Entre 23 y 25	9	68	77	11,7%	88,3%
Mayor a 26		20	20	0,0%	100,0%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 16. Diferencias de para las edades entre Postulantes Destacados y no destacados.

El mes de nacimiento expresado como el porcentaje sobre el total, muestra que los meses de enero, abril, mayo y junio están por sobre el porcentaje de Postulantes Destacados de la distribución general. La tabla Nº 17 muestra estos resultados.

Mes de Nacimiento	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado	% Postulante No Destacado
1	44	111	155	28,4%	71,6%
2	23	105	128	18,0%	82,0%
3	35	115	150	23,3%	76,7%
4	36	101	137	26,3%	73,7%
5	37	109	146	25,3%	74,7%
6	34	91	125	27,2%	72,8%
7	25	118	143	17,5%	82,5%
8	28	142	170	16,5%	83,5%
9	41	147	188	21,8%	78,2%
10	40	139	179	22,3%	77,7%
11	30	133	163	18,4%	81,6%
12	38	127	165	23,0%	77,0%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 17. Diferencias porcentuales según mes de nacimiento.

Las variables que se relacionan con el poder adquisitivo toman importancias diferentes. Por su parte las variables de cuantas personas trabajan y cuán grande es el grupo familiar no tienen

alguna diferencia en proporciones interesante. Sobre el financiamiento de la educación superior, antes de resultados de becas, no se aprecian diferencias importantes entre las clases. Sin embargo, la variable de ingreso bruto familiar presenta diferencias en las proporciones entre las clases, se presenta con tramos de ingreso que aumentan en $144.000 por tramo. Se aprecia que en los tramos de más alto ingreso no hay diferencias importantes mientras que en el primer tramo de ingresos si hay diferencias de hasta un 4% menor a la proporción general de Postulantes Destacados. Los tramos de ingreso sexto, séptimo y octavo tienen una proporción de Postulantes Destacados mucho mayor a todos los otros tramos. Se presentan las diferencias en la tabla Nº 18.

Ingreso Bruto Familiar	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado	% Postulante No Destacado
1	44	190	234	18,8%	81,2%
2	91	374	465	19,6%	80,4%
3	51	199	250	20,4%	79,6%
4	34	110	144	23,6%	76,4%
5	26	86	112	23,2%	76,8%
6	45	141	186	24,2%	75,8%
7	17	38	55	30,9%	69,1%
8	42	106	148	28,4%	71,6%
9	5	8	13	38,5%	61,5%
10	2	8	10	20,0%	80,0%
11	4	15	19	21,1%	78,9%
12	50	163	213	23,5%	76,5%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 18. Diferencias en ingreso bruto familiar entre las clases

Además se encuentran hasta un 4% de diferencias en los casos de estudiantes con salud privada como isapre versus FONASA. Ver en la tabla Nº 19. Además la ocupación de los padres, es decir que tienen trabajo, ya sea el padre 1 o padre 2 en la base de datos, se identifica, Postulantes con padres que trabajan con mayor proporción en Postulantes Destacados. Se ven estos resultados en la tabla Nº 20.

Cobertura Salud	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado	% Postulante No Destacado
FONASA	98	386	484	20%	80%
Isapre	284	911	1195	24%	76%
Otro	25	129	154	16%	84%
Dipreca		3	3	0%	100%
Capredena	4	9	13	31%	69%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 19. Cobertura de Salud

Ocupación Padre 1	Postulante Destacado	Postulante No Destacado	Ocupación Padre 2	% Postulante Destacado	% Postulante No Destacado
No trabaja	203	769	972	21%	79%
Trabaja	208	669	877	24%	76%
Ocupación Padre 2				**22%**	**78%**
No trabaja	91	379	470	19%	81%
Trabaja	320	1059	1379	23%	77%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Tabla Nº 20. Ocupación de los padres.

Finalmente al revisar otras variables relativas al colegio del estudiante, se encuentra que los postulantes que no estudian en la misma comuna se identifican en la muestra con un 3% en promedio más con la clase positiva que con la negativa. Los estudiantes que estudian en la misma provincia no tienen variación entre sus postulantes. Ver tabla Nº 21 para observar el detalle.

Estudia misma comuna	Postulante Destacado	Postulante No Destacado	Total	% Postulante Destacado sobre total	% Postulante No Destacado sobre total	% Postulante Destacado sobre Total Destacado	% Postulante No Destacado sobre total No Destacado
No	217	808	1025	21%	79%	53%	56%
Sí	194	630	824	24%	76%	47%	44%
Total	**411**	**1438**	**1849**	**22%**	**78%**	**22%**	**29%**

Tabla Nº 21. Diferencias entre las clases en cuanto a estudiar en la misma comuna.

Capitulo 7 – Construcción del modelo predictivo

7.1 Muestra utilizada

Como se explica en el Capítulo 2, el foco de esta tesis será en los estudiantes que ingresan vía PSU, señalizando sus puntajes, desempeño escolar y datos socio-económicos. El estudio utiliza los datos de 1849 estudiantes que ingresaron vía PSU a la facultad en el periodo de tiempo entre 2004-2010. Se utiliza esta de desfase de 5 años, como explica el Capítulo 5 sobre enfoque de solución, pues es el periodo de tiempo necesario para que un estudiante se titule o egrese en el plazo normal acorde a la malla curricular, es con esta información con la que los estudiantes son marcados como Postulantes Destacados o Postulantes No Destacados. La tabla Nº 22 muestra la distribución entre las clases. El análisis comparativo en detalle sobre estas clases se encuentra en el Capítulo 6 sección 3.

Clase	Cantidad de Instancias	Porcentaje sobre total
Postulante Destacado	411	22%
Postulante No Destacado	1438	78%
Total	**1849**	

Tabla Nº 22. Distribución entre las clases

7.2 Bases de datos utilizadas

Esta tesis utiliza dos grandes bases de datos, la primera es el Sistema de Administración Docente (SAD) mientras que la segunda son los datos del DEMRE. Una pertenece a la Facultad de Economía y Negocios y es utilizada en la administración de la propia facultad mientras que la segunda es entregada por el DEMRE a todas las universidades que participan del proceso de selección universitaria vía PSU.

La base de datos del Sistema de Administración Docente, es vital para el funcionamiento de la facultad ya que es este sistema el que mantiene la principal información sobre los estudiantes y docentes de la facultad. Esta información contiene los cursos, notas, asignaciones de salas, homologación de ramos, desempeño académico, evaluaciones docentes y hasta solicitudes estudiantiles.

La base de datos del DEMRE, es decir el Departamento de Evaluación, Medición y Registro Educacional (DEMRE), el que depende de la Universidad de Chile es responsable de consolidar los registros relacionados con los puntajes de las pruebas de selección, las postulaciones de los estudiantes, la información sobre los colegios de los estudiantes postulantes y también datos sobre la información socio económica sobre los estudiantes y sus familias.

Esta información, cabe destacar, es recopilada en encuestas realizadas a los estudiantes en diferentes instancias en relación a la toma de la prueba de selección universitaria. La información es un su mayoría constante a lo largo de los años de estudio de esta tesis y permite su análisis temporal.

7.3 Variables utilizadas

Las variables utilizadas en esta tesis provienen de la base de datos del DEMRE, exceptuando la variable dependiente. La variable dependiente corresponde a la clase a predecir en esta tesis. Esta clase fue definida como el perfil de captación asignado a los postulantes universitarios, los que podían ser Postulante Destacado y Postulante No Destacado. La construcción de esta clase se encuentra detallada en la sección 6.3 de esta tesis. Esta variable tiene entonces características binomiales que provienen de las dos categorías descritas.

Variables Desempeño Pre – Universitario

Entre estas variables se encuentran las que representan el desempeño académico escolar y las pruebas de selección universitaria.

1) Notas del colegio: Para representar el desempeño académico del colegio se utiliza el promedio de notas de la enseñanza media. Estas notas cabe destacar se convierten en un puntaje que es conocido como NEM.
2) PSU: Por otro lado luego de evaluadas las respuestas de la prueba de selección universitarias estas son asignadas con un puntaje. Quiere decir que se traducen las respuestas a los puntajes de las pruebas de lenguaje, matemática, ciencias y historia y ciencias sociales. Al profundizar los análisis del comportamiento de los Postulantes Destacados y No Destacados se crean tramos para los puntajes de las pruebas que se describen en Anexo.

Variables del Colegio de los Postulantes

Los colegios entregan diferente información al DEMRE para ser considerada y cruzada con los estudiantes que tienen. El DEMRE envía esta información y en esta tesis se utiliza la información disponible sobre:

1) Región: Se utilizan los datos sobre la ubicación del colegio en cuanto a la región del país.
2) Rama Educacional: Es posible observar la rama educacional al que pertenecen los estudiantes. La rama educacional son dos, Técnico y Humanista.
3) Grupo Dependencia: Los colegios tienen una fuente de financiamiento y apoyo económico, y esta información también se encuentra en el DEMRE. Los tipos de

dependencia son tres: municipal, particular subvencionado y particular pagado. La dependencia municipal quiere decir que recibe financiamiento estatal a través de las comunas donde se encuentran. El particular pagado es el que no mantiene vínculos financieros con el estado y todo el financiamiento proviene de los propios apoderados y privados. El particular subvencionado en cambio, recibe tanto aportes estatales como financiamiento de privados, las proporciones de este financiamiento pueden variar de colegio en colegio.

4) Régimen Educacional: El régimen educacional corresponde al género de los estudiantes en el colegio, esto quiere decir que un colegio solo puede aceptar hombres o mujeres por separado es un colegio de régimen único mientras los colegios mixtos se conocen como coeducacionales.
5) Desempeño Promedio: Esta variable es calculada gracias a los datos que envía el DEMRE. Considera el promedio de los puntajes de cada prueba de los estudiantes de ese colegio para un año en particular. Esto es calculado para el promedio de notas del colegio, promedio de puntaje de lenguaje, promedio de puntaje de matemática, promedio de puntaje de ciencia y promedio de puntaje de historias y ciencias sociales.

Variables Socio demográficas

El DEMRE aplica encuestas a los estudiantes que se inscriben y estas permiten describir tres aspectos importantes sobre el estudiante secundario que postula a la universidad, que son: La ubicación geográfica, la caracterización del estudiante y finalmente el contexto familiar.

1) Ubicación geográfica: Corresponde a la región donde vive el estudiante pues describe en parte su contexto cotidiano. Además se agregan variables calculadas tales como si el estudiante estudia en la misma comuna donde vive, comparando la comuna del colegio con la comuna de residencia del estudiante. De la misma forma se compara la provincia de residencia del estudiante con la provincia del colegio. Estas variables son del tipo binomial o dicotómico, es decir toma valor positivo o negativo según sea el caso.
2) Caracterización del estudiante: Estas variables hacen referencia a algunos comportamientos particulares del estudiante que reflejan de algún modo los atributos intrínsecos del estudiante. Estos pueden ser si el estudiante trabaja o no trabaja y cuantas horas a la semana trabaja, el género, la edad del estudiante a la PSU calculada con su fecha de nacimiento y el mes de nacimiento, y la situación de egreso del colegio.
3) Contexto familiar: Estas variables describen cómo está compuesto el núcleo familiar del estudiante, si trabajan sus padres, el nivel de estudio de sus padres, el nivel de ingreso bruto familiar, cuantas personas trabajan en la familia, las opciones de financiamiento con las que cuenta la familia además de la cobertura de salud.

7.4 Selección de Variables mediante Técnicas de Minería de datos

Esta etapa consiste en aplicar técnicas de Minería de datos para listar variables importantes y seleccionar las principales para dejar fuera variables que sean ruido en los datos. La etapa de selección de variables mediante técnicas de Minería de datos se utiliza para tanto para el segundo enfoque de Modelos en Serie como también para el tercer enfoque de Modelos Combinados.

En RapidMiner existen variadas técnicas para realizar procesos iterativos de mejora de desempeño del modelo en cualquiera de sus etapas. Si nos centramos en la selección de variables es posible estudiar el Information Gain, Matriz de Covarianza y Matriz de Correlación de las variables. Pero además es posible generar un modelo que identifique variables importantes antes de predecir. La primera técnica utilizada antes de generar el modelo predictivo corresponde al Árbol de decisión. Realizar un árbol de decisión es muy útil puesto que es un proceso de inducción recursiva. Este proceso genera una asignación de nodos, raíz y hojas del árbol y a mayor cantidad de hojas y ramas más puro se vuelven las asignaciones de las instancias dentro de determinada hoja. Por lo tanto se generan grandes árboles para su estudio y obtención de variables importantes. La generación de estos subgrupos permite obtener reglas de clasificación que son de interés para luego seleccionar variables.

La segunda técnica utilizada corresponde al Sequential Backward Elimination. Esta técnica permite iterar modelos cambiando las variables con las que se genera para eliminar una a una las variables que no mejoran el modelo. Es decir elimina variables hasta que se cumple una regla de detención del algoritmo, en el caso particular de esta tesis se utiliza el criterio de detención de disminución de desempeño del modelo. Esto quiere decir que el proceso se ejecuta, eliminando variables hasta que el desempeño disminuye.

Las técnicas de selección revisadas en los párrafos anteriores permiten generar mejor desempeños en los modelos predictivos en base a eliminar el ruido de variables que no se detectan a simple análisis estadístico, como también el ajuste de las variables para cada modelo.

7.5 Pre procesamiento y transformación

Como vimos en el Capítulo 4, es necesario realizar tratamientos particulares para utilizar las técnicas de Minería de datos a los datos según el formato de origen de ellos. Estos datos ya se encuentran limpios de valores outliers así como también de valores faltantes ya que se prefiere

reducir la base de datos debido a que no eran relevantes. Esta etapa es transversal para todos los enfoques de solución de Enfoque Singular, Modelos en Serie y Modelos Combinados

En esta etapa de preparación de los datos se transforman los atributos nominales o polinomiales a binominales creando n columnas, una columna por cada categoría diferente en el atributo inicial. De las columnas generadas se remueve una de ellas de la base de entrenamiento para evitar problemas con la multicolinealidad. A su vez las variables que son numéricas son normalizadas y se transforman entre los rangos de 0 y 1. El primer tipo de dato a transformar son los numéricos puesto que tienen distintos rangos, tales como las notas, los puntajes de las pruebas y el ingreso bruto familiar. El segundo tipo de dato a transformar son las variables nominales y polinomiales, es decir las categorías de datos. Por ejemplo la cobertura de salud del estudiante que son de hasta cinco tipos, o la rama y régimen del colegio. Para poder operar con estos en forma apropiada en las diferentes técnicas de Minería de datos se genera una variable (columna) adicional para cada una de las categorías, es decir que existirán tantas nuevas variables como categorías tenga la variable inicial. Luego para operar con estas es necesario reducir la muestra en una de estas variables generadas para no tener problemas de multicolinealidad y se utilizan n-1 categorias. Además se remueven de la base de datos los registros incompletos.

7.6 Análisis del sobre ajuste

Como fue discutido en el Capítulo 4 es importante revisar los modelos en la búsqueda de sobre ajuste pues no permite la aplicación de los modelos predictivo en casos reales.

Una forma de evaluar el sobre ajuste es comparando los desempeños de predicción del modelo en el entrenamiento versus los desempeños predictivos en la validación de la predicción en la etapa de testeo. En los modelos revisados se experimentó que el utilizar Bagging en técnicas como RL y SVM generó un over-fitting tan grave que la predicción obtenía un desempeño muy pobre en la etapa de testeo en validación cruzada. Por lo que Bagging se resguarda solo para uso en DT. Esto se valida comparando los resultados con predicciones fuera del entrenamiento en la etapa de testeo de validación cruzada.

A continuación el Gráfico Nº 2, presenta el desempeño del modelo en la etapa de testeo es decir el Accuracy comparándolo con el porcentaje de acierto en el entrenamiento. Los ejes del gráfico corresponden al porcentaje de acierto y a la complejidad del modelo, es decir mayor uso de Bagging y Stacking. La complejidad del modelo viene dada por la cantidad de técnicas de Minería de datos y técnicas de combinación son aplicadas. Se observa que a medida que se complejiza el modelo por ejemplo utilizando dos o más bagging la predicción del entrenamiento

supera el 90% pero la predicción real en el testeo cae bruscamente, siendo un caso de sobre ajuste grave y por ello se evitaron o desecharon modelos demasiado complejos en el enfoque de Modelos Combinados Secuenciales. Así también se evitaron técnicas de booster como Adaboost en las técnicas de aprendizaje de RL y SVM, ya que favorecían fuertemente el sobre ajuste.

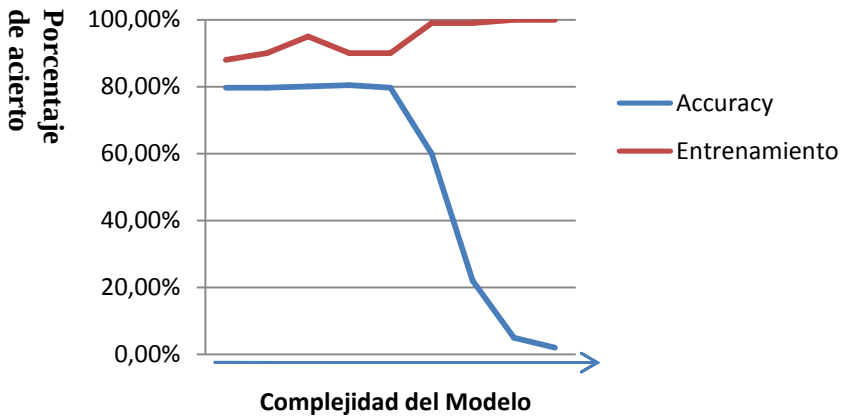


Gráfico Nº 2. Sobre ajuste de modelos demasiado complejos.

El desempeño de los modelos combinados utilizados tiene un porcentaje de diferencia de predicción con el entrenamiento dentro de un rango aceptable que es calculado por la validación cruzada y los porcentajes de predicción mencionados son siempre sobre datos de testeo no utilizados para el entrenamiento.

Capitulo 8 – Resultados experimentales

En este capítulo se discuten los resultados entregados por los modelos implementados en Software Rapidminer así como también los distintos hallazgos particulares por cada enfoque de solución; los enfoques de solución son Enfoque de Modelo Singular, Modelos Secuencial, Modelos Combinados y Modelos Secuencial Combinados. Se describen entonces los desempeños de los cuatro enfoques y las variables explicativas para las clases estudiadas de postulantes a FEN del enfoque de Modelo Secuencial Combinado.

8.1 Resultados de Enfoque de Modelo Singular

Los resultados del enfoque singular provienen de la aplicación de KDD utilizando un solo modelo, bajo una sola técnica de aprendizaje ya sea SVM, RL o DT. Como explica el Capítulo 5 son modelos con una sola técnica que no utilizan selección de variables mediante técnicas complejas ni tampoco combinan sus resultados entre sí. Se revisa a continuación el desempeño de los modelos predictivos, para Support Vector Machine (SVM), Regresión Logística (RL) y Árboles de Decisión (DT, por sus siglas en ingles). El desempeño de los modelos será revisado utilizando el accuracy o performance en general de cada modelo y además el Ratio de la Clase Positiva. La clase positiva corresponde al estudiante Postulante Destacado mientras que la clase negativa corresponde al estudiante Postulante No Destacado. La tabla Nº 23 presenta los resultados de los 10 mejores modelos de este enfoque de solución.

Para el caso de la técnica de aprendizaje SVM se ejecutan todas sus combinaciones entre el parámetro C y los costos de clasificación. En anexo 4 se encuentran los Modelos Singulares SVM con sus parámetros. Para SVM se obtiene un máximo de acierto de predicción de un Accuracy de 80% y por su parte el Ratio Verdadero Positivo alcanza un 34,9%. En este caso tener un 80% de Accuracy en general no resulta relevante debido al pobre desempeño del modelo en cuanto al Ratio Verdadero Positivo.

Para el caso de la técnica de aprendizaje RL se ejecutan todas sus combinaciones de C y los costos de clasificación de los clasificadores. En este caso, para RL se obtiene un máximo un acierto de predicción, es decir Accuracy de 81% y por su parte el Ratio Verdadero Positivo alcanza un 45,45%. En este caso tener un 81% de Accuracy tampoco resulta relevante debido al pobre desempeño del modelo en cuanto al Ratio Verdadero Positivo.

Para el caso de la técnica de aprendizaje DT se ejecutan combinaciones de sus parámetros de profundidad de árbol, tamaños mínimos para crear ramas, poda y pre poda. En este caso, para DT se obtiene un máximo un acierto de predicción, es decir Accuracy de 82% y por su parte el Ratio Verdadero Positivo alcanza un 50%. Cabe destacar que como se presenta en el Capítulo 4,

la técnica DT tiene gran variabilidad de predicción debido a su adaptación. La adaptación depende de que tan complejo sea el modelo en términos de ramas y árboles creados. Ahora bien esta adaptación puede subir el Ratio Verdadero Positivo pero lo hace en desmedro importante del Accuracy. Aún así tener un 82% de Accuracy tampoco resulta relevante debido al pobre desempeño del modelo en cuanto al Ratio Verdadero Positivo.

En general tienen bajo desempeño los modelos predictivos del enfoque singular. Esto es debido a la complejidad del problema, en particular debido a la baja separación entre las clases como fue revisado en Capítulo 6 sección 3. En donde se aprecian diferencias entre las clases en proporciones entre un 3% y 6%. Sumado a esto se observa en los datos claramente que estudiantes con similares variables explicativas de colegio y desempeño pre universitario tiene resultados muy diferentes. Se presentan en la tabla Nº 23 el ranking de los mejores resultados para este enfoque, se presentan resultados diferentes a modelos con diferentes parámetros utilizados.

Ranking	Técnica	¿Utiliza Clasificador?	Accuracy	Ratio Verdadero Positivo
1	DT	Sin Clasificador	81,5%	37,1%
2	DT	Sin Clasificador	81,4%	50,0%
3	SVM	Sin Clasificador	81,2%	25,0%
4	RL	Si	80,6%	45,4%
5	RL	Si	80,6%	45,5%
6	SVM	Sin Clasificador	79,7%	34,9%
7	SVM	Sin Clasificador	79,6%	26,3%
8	DT	Sin Clasificador	79,4%	35,3%
9	SVM	Si	78,9%	26,8%
10	SVM	Sin Clasificador	78,9%	24,8%

Tabla Nº 23. Resultados Modelos Singulares

8.2 Resultados de enfoque de Modelos Secuenciales

Se revisa a continuación el desempeño de los modelos predictivos, para las dos etapas del modelo en serie. En primera etapa se describen los principales resultados de la selección de variables y luego se describen los resultados de la segunda etapa sobre técnicas de aprendizaje. Las técnicas de aprendizaje utilizadas fueron Support Vector Machine (SVM), Regresión Logística (RL) y Árboles de Decisión (DT, por sus siglas en ingles).

En cuanto a los principales resultados del modelo en la primera etapa del enfoque de Modelos Secuenciales, es conveniente iniciar la revisión presentando el análisis de Information Gain para las variables explicativas estudiadas. Este análisis presenta cuanta información entrega cada variable independiente a la clase o variable dependiente. El uso de Information Gain y forma de

cálculo se presentó en el Capítulo 4. En el Gráfico Nº 3 se presenta el Information Gain para las variables más seleccionadas por el análisis de la primera etapa. Muchas categorías de variables quedan fuera, por lo tanto la tabla completa se presenta en Anexos. Por su parte en la primera etapa se busca encontrar variables relevantes mediante técnicas de Minería de datos, de ellas se utiliza Árbol de decisión (DT) y Sequential Backward Elimination (SBE). Ambas entregan un listado de variables. En el Gráficos Nº 4 y Gráficos Nº 5 se describe la cantidad de veces que estas variables están presentes en los 48 modelos generados. Como se puede ver en el gráfico Nº 3 existen muchas variables con poco peso y muy pocas de ellas destacan con un Information Gain mayor, esto tiene que ver con lo similar que resultan las clases en sus datos de entrenamiento. Por su parte las técnicas de selección de variables se relacionan fuertemente con el Information Gain. Las técnicas de selección de variables también se relacionan entre sí en la mayoría de las variables.

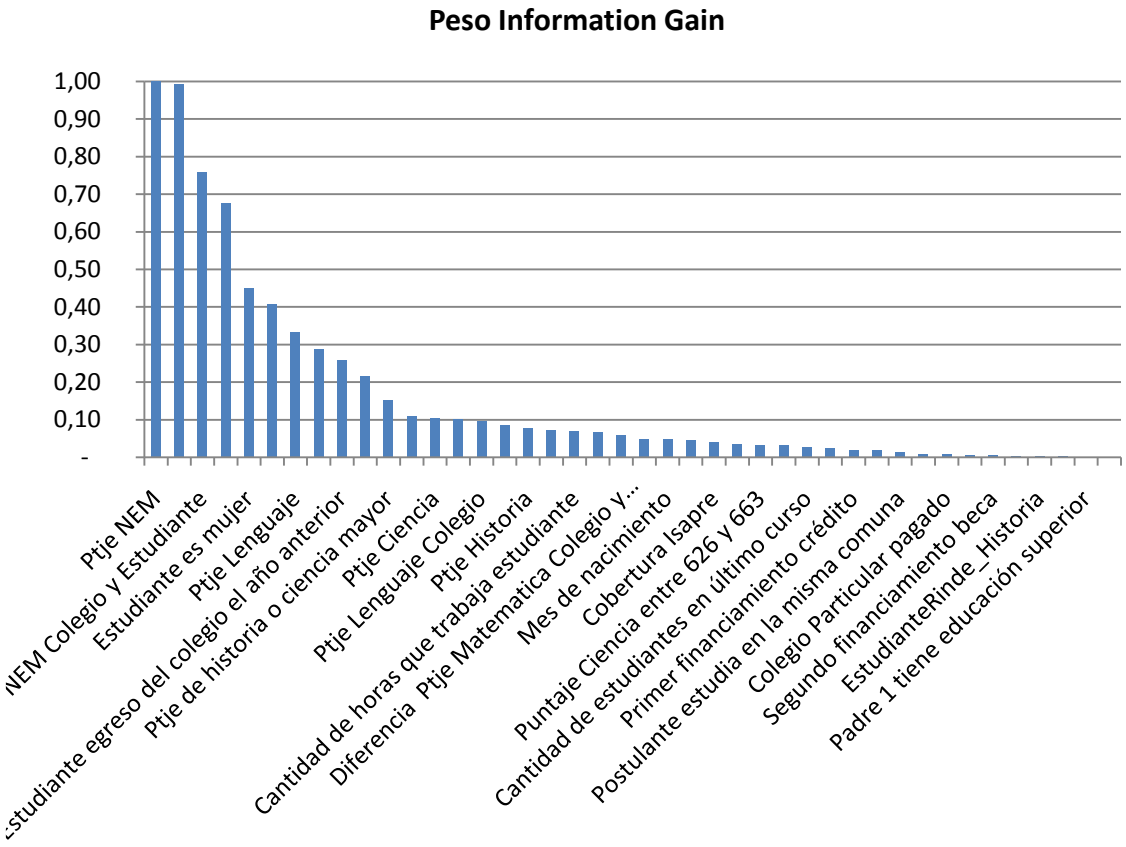


Gráfico Nº 3. Information Gain.

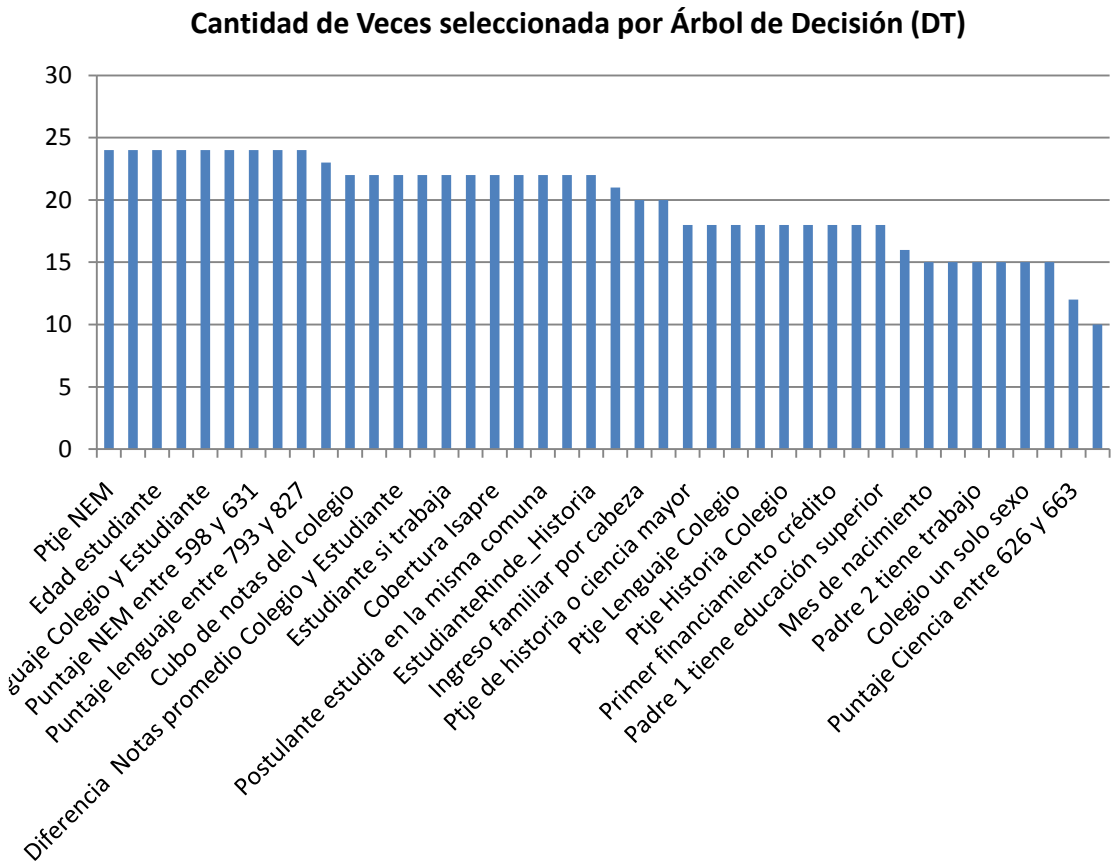


Gráfico Nº 4. Cantidad de veces que es seleccionada una variable por árbol de decisión.

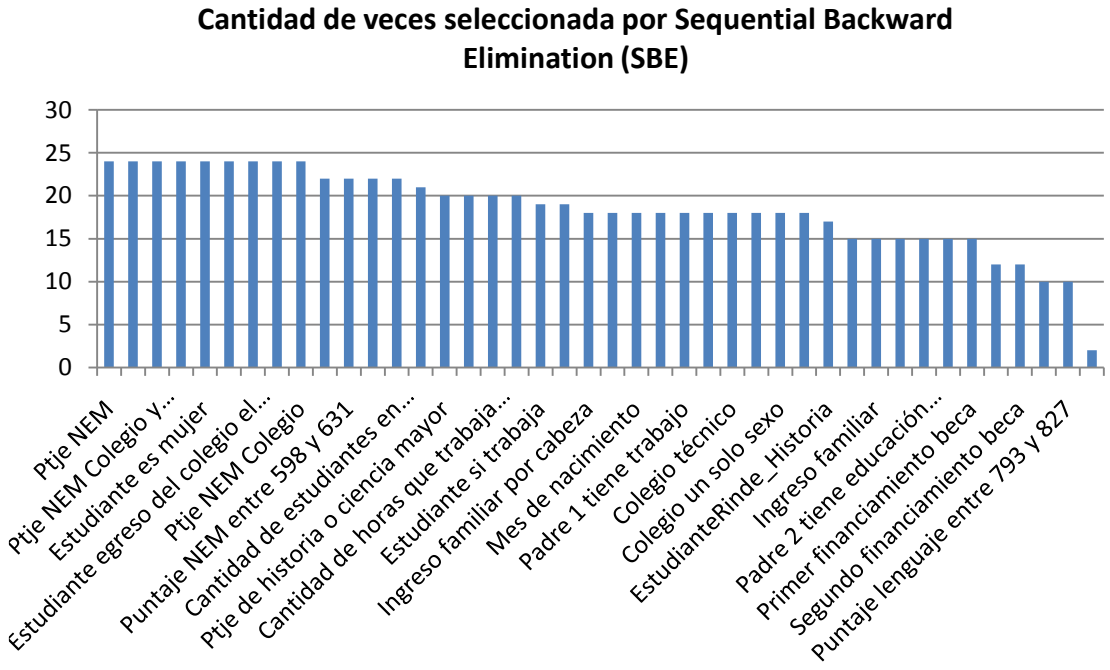


Gráfico Nº 5. Cantidad de veces que es seleccionada una variable por Backward Elimination

Utilizando entonces las variables descritas en el gráfico Nº 3 y Nº 4, se generan modelos predictivos para cada caso. Para el caso de la técnica de aprendizaje SVM se ejecutan todas sus combinaciones entre el parámetro C y los costos de clasificación. Para SVM se obtiene un máximo un acierto de predicción, es decir Accuracy de 81,43% y por su parte el Ratio Verdadero Positivo alcanza un 47,83%. Nuevamente el Accuracy en general no resulta relevante debido al pobre desempeño del modelo en cuanto al Ratio Verdadero Positivo.

Para el caso de la técnica de aprendizaje RL se ejecutan todas sus combinaciones de C y los costos de clasificación de los clasificadores. En este caso, para RL se obtiene un máximo un acierto de predicción, es decir Accuracy de 82,49% y por tu parte el Ratio Verdadero Positivo alcanza un 62,89%. Este ratio tiene mucho mejor desempeño que todos los otros encontrados, pero aún así no es suficiente para tener una predicción confiable.

Para el caso de la técnica de aprendizaje DT se ejecutan combinaciones de sus parámetros de profundidad de árbol, tamaños mínimos para crear ramas, poda y pre poda. En este caso, para DT se obtiene un máximo un acierto de predicción, es decir Accuracy de 82,21% y por su parte el Ratio Verdadero Positivo alcanza un 59,09%. La Tabla Nº 24 resume los resultados de modelos destacados en este enfoque.

Técnica	¿Aplica Cluster?	¿Aplica Árbol de Decisión?	¿Aplica SBE?	¿Aplica Clasificador?	Accuracy	Ratio Verdadero Positivo
RL	Sin Cluster	Si utiliza	Si utiliza	Clasificador Sin	82,49%	62.89%
DT	Sin Cluster	Si utiliza	Si utiliza	Clasificador Sin	82,21%	59.09%
DT	Sin Cluster	Si utiliza	No	Clasificador	81,54%	37.10%
SVM	Sin Cluster	Si utiliza	Si utiliza	Clasificador Sin	81,43%	47.83%
DT	Sin Cluster	No	No	Clasificador	81,40%	42.42%
RL	Sin Cluster	Si utiliza	Si utiliza	Clasificador Sin	81,32%	49.03%
DT	Sin Cluster	Si utiliza	No	Clasificador	81,27%	46.51%
SVM	Sin Cluster	Si utiliza	No	Clasificador	81,15%	46.97%
SVM	Sin Cluster	No	No	Clasificador	80,82%	43.22%
RL	Cluster	Si utiliza	No	Clasificador	80,80%	47.33%

Tabla Nº 24. Desempeño de los modelos del Enfoque de Modelos Secuenciales

En general tienen bajo desempeño los modelos predictivos del enfoque en serie. Los mejores resultados se dan al utilizar ambas técnicas de selección de variables, en particular Sequential Backward Elimination tiende a aumentar la precisión del modelo. Además de la complejidad en el problema de predicción se detecta gran variabilidad entre los postulantes que desafían a los modelos a encontrar el ajuste necesario. Por lo anterior pese a que el Enfoque de Modelos en Serie es superior al Enfoque de Modelo Singular, se mantiene un desempeño predictivo pobre el que no responde aún a los problemas a resolver por esta tesis.

8.3 Resultados de Enfoque de Modelos Combinados

El enfoque de solución de Modelos Combinados permite mezclar los resultados de diferentes modelos mientras sean del mismo tipo de técnica de aprendizaje. En esta tesis se combinan resultados para SVM, RL y DT.

Se obtienen mejores resultados para la combinación de modelos de DT en general. Alcanza hasta un accuracy de 80,9% con un Ratio Verdadero Positivo de un 65%, siendo este resultado uno de los mejores modelos hasta ahora. Ahora bien, el resto de los modelos se encuentran cercanos al 50% de Ratio Verdadero Positivo. Para el caso de SVM se obtiene hasta un 80,8% de accuracy con un Ratio Verdadero Positivo de un 30%, siendo un resultado de poco interés debido al bajo ratio positivo. Mientras que para el caso de RL se obtienen resultados similares, con un 80,1% de accuracy pero un Ratio Verdadero positivo de un 54,4%. La Tabla Nº 25 muestra los diez mejores modelos de combinación.

Cabe destacar que muchos modelos combinados sufren de un sobre ajuste del modelo muy alto. Esto quiero decir que se obtiene un resultado de predicción muy alto en el entrenamiento mientras que luego de aplicar la predicción a datos fuera de la muestra de entrenamiento el resultado de predicción es muy bajo. Estos casos tienen una predicción en entrenamiento sobre 70% de accuracy y bajo 30% en los datos de prueba. Esto es particularmente acentuado en SVM.

Ranking	Técnica Combinada en Bagging	¿Utiliza Clasificador?	Accuracy	Ratio Verdadero Positivo
1	DT	Sin Clasificador	81,1%	55,7%
2	DT	Si	80,9%	65,0%
3	SVM	Sin Clasificador	80,8%	30,0%
4	RL	Si	80,1%	54,4%
5	SVM	Si	79,5%	54,5%
6	RL	Sin Clasificador	79,2%	41,9%
7	SVM	Sin Clasificador	79,1%	39,4%
8	DT	Sin Clasificador	78,9%	52,9%
9	SVM	Si	78,4%	29,8%
10	SVM	Sin Clasificador	75,2%	37,2%

Tabla Nº 25. Desempeño de los modelos del Enfoque de Modelos Combinados

8.4 Resultados de Enfoque de Modelos Combinados Secuenciales

8.4.1 Desempeño de Modelos Combinados Secuenciales

Los resultados de los enfoque de Modelos Combinados Secuenciales resultan de mucho mejor desempeño que los dos enfoques de modelos anteriores. Los resultados serán revisados según los modelos combinados y las técnicas de combinación, es decir si se utiliza SVM, RL o DT y si se combinan utilizando Bagging o Stacking.

Se obtienen un mejor desempeño por un lado, gracias a la capacidad de la técnica de Bagging de mejorar los resultados de un tipo de modelo a través de la combinación de diversos resultados. Esta técnica en árboles de decisión tiene un gran alcance y mejora mucho su desempeño. Por otro lado la técnica de Stacking permite combinar las diversas predicciones de los tres modelos para entregar un nuevo resultado, que de acuerdo al experimento realizado, resulta en un mejor desempeño. Ahora bien cabe destacar que a mayor complejidad de los modelos mayor es el riesgo de sobre ajustar el modelo o riesgo de over-fitting. Esto quiere decir que el modelo no es útil en la predicción de nuevas instancias no utilizadas en el entrenamiento. Se revisa en detalle este tema en la sección 8.4 para comparar modelos. Por lo que Bagging se considera sólo para uso en DT.

Dado lo anterior, el mejor desempeño utiliza un modelo complejo pero sin caer en un over-fitting. Este modelo corresponde a la combinación de Stacking con SVM, RL y sólo utilizar Bagging en DT. Si bien es cierto que es un modelo complejo, su over-fitting es de menor relevancia y permite realizar buenas predicciones. En particular este modelo combinado obtiene

resultados de predicción con un Accuracy entre 80,4% y 76,74%. La precisión de este modelo en base al Ratio Verdadero Positivo varía entre un 100% y un 80% en la mayoría de los modelos, lo anterior según los parámetros utilizados en las técnicas. Por su parte se destaca también el desempeño de 3 modelos que combinan sólo dos técnicas, una de ellas no utiliza Bagging. En un modelo combinado de RL y DT se obtiene un Accuracy de 80,04% y una precisión de un 78,13% situándolo como uno de los mejores modelos entre los modelos combinados. Se presenta a continuación la Tabla Nº 26 con un ranking de los 20 mejores modelos que presenta estos resultados.

Rank	ID Modelo	Selección Decision Tree	Sequential Backward Elimination	Cantidad Modelos Stacking	RL en Stacking	SVM en Stacking	Aprendizaje que utiliza Bagging	Accuracy	Ratio Verdadero Positivo
1	28	Aplica DT	Aplica SBE	3	RL	SVM	DT	80,40%	89,47%
2	35	Aplica DT	Aplica SBE	2	RL	No	DT	80,04%	78,13%
3	29	Aplica DT	Aplica SBE	3	RL	SVM	DT	79,67%	86,36%
4	30	Aplica DT	Aplica SBE	3	RL	SVM	DT	79,67%	86,36%
5	31	DT	Aplica SBE	3	RL	SVM	DT	79,67%	86,36%
6	40	No	Aplica SBE	3	RL	SVM	DT	79,49%	89,47%
7	41	No	No Aplica	3	RL	SVM	DT	79,49%	89,47%
8	42	No	Aplica SBE	3	RL	SVM	DT	79,49%	89,47%
9	32	No	Aplica SBE	3	RL	SVM	DT	79,49%	89,47%
10	34	Aplica DT	Aplica SBE	2	No	SVM	DT	79,49%	72,73%
11	46	Aplica DT	Aplica SBE	2	RL	No	DT	79,49%	89,47%
12	24	No	No	3	RL	SVM	DT	79,30%	81,82%
13	44	Aplica DT	Aplica SBE	2	No	SVM	DT	79,30%	85,00%
14	25	No	Aplica SBE	3	RL	SVM	DT	79,12%	80,95%
15	38	Aplica DT	Aplica SBE	3	RL	SVM	DT	79,12%	80,95%
16	36	Aplica	Aplica SBE	2	RL	SVM	No	78,94%	75,00%

		DT Aplica							
17	48	DT	Aplica SBE	2	RL	SVM	No	78,94%	75,00%
18	37	No	Aplica SBE	3	RL	SVM	DT	78,75%	92,31%
19	18	No	Aplica SBE	3	RL	SVM	DT	77,84%	87,50%
20	17	No	Aplica SBE	3	RL	SVM	DT	76,74%	100,00%

Tabla Nº 26. Desempeño 20 mejores Modelos de Enfoque Combinado.

En la tabla Nº 26 se presenta que el mejor Ratio Verdadero Positivo se alcanza con un Accuracy de un 76,74% por lo tanto el uso de este modelo o el primero del ranking que combina una mejor Accuracy y un muy buen Ratio Verdadero Positivo depende de los recursos disponibles en el proceso de llamados a los Postulantes Destacados. Es decir, que si se cuenta con muchos recursos se recomienda llamar a la mayor cantidad de estudiantes, si se requiere ser más preciso se prefiere entonces el modelo con mayor precisión. Se discute esto en el Capítulo 9.

Como se observa entre los modelos la única técnica de aprendizaje que utiliza Bagging es el DT. Por dos factores, el primero es el over-fitting que genera utilizarlo en SVM y RL. El segundo factor tiene que ver con que un solo modelo DT no entrega resultados de predicciones potentes, pero al combinar variados de ellos como lo hace Bagging permite considerar al mismo tiempo la variabilidad de predicción de DT y transformarlo en un modelo útil.

Luego se destaca que la combinación de sólo dos modelos entrega menor desempeño y precisión que combinar tres modelos. Por ejemplo en el modelo ubicado en el segundo lugar del ranking se obtiene una precisión de solo 78,13% muy por debajo del primer modelo que obtiene un 89,47% de precisión.

Cabe destacar que la cantidad de combinaciones posibles en este caso es muy grande, se resumen en la tabla Nº 26 los resultados más interesantes. Como se comentó en esta sección el problema de over – fitting afectó a variados modelos muy complejos descartando del análisis todos los modelos que tienen Bagging en RL y SVM, así como también las combinaciones de modelos utilizando DT sin Bagging que tienen un pobre desempeño. Es claro que en los modelos con buena predicción el Accuracy del Enfoque de Modelos Combinados ronda el 80%, con 13 modelos entre 79,67% y 79,12%, mientras que la moda del Ratio Verdadero Positivo es de un 89,47% con un máximo de hasta 100% de precisión

.

8.4.2 Variables explicativas de los Modelos Combinados Secuenciales

La combinación de modelos mediante Stacking utiliza una técnica de aprendizaje adicional para considerar las predicciones de los otros modelos. En este caso la técnica utilizada fue Naive Bayes. Como fue revisado en el Capítulo 4, esta técnica se basa en el análisis y extensión del Teorema de Bayes y por lo tanto expresa las variables en términos de media y varianza para crear una distribución normal para cada clase.

Revisemos a continuación las diferentes distribuciones entre clases que indican cuán diferente y explicativa es una variable con respecto a la clase. De los modelos combinados generados secuenciales se realiza un ranking de las variables con respecto a las veces que se utiliza estas variables. El gráfico Nº 6 a continuación resume estos resultados.

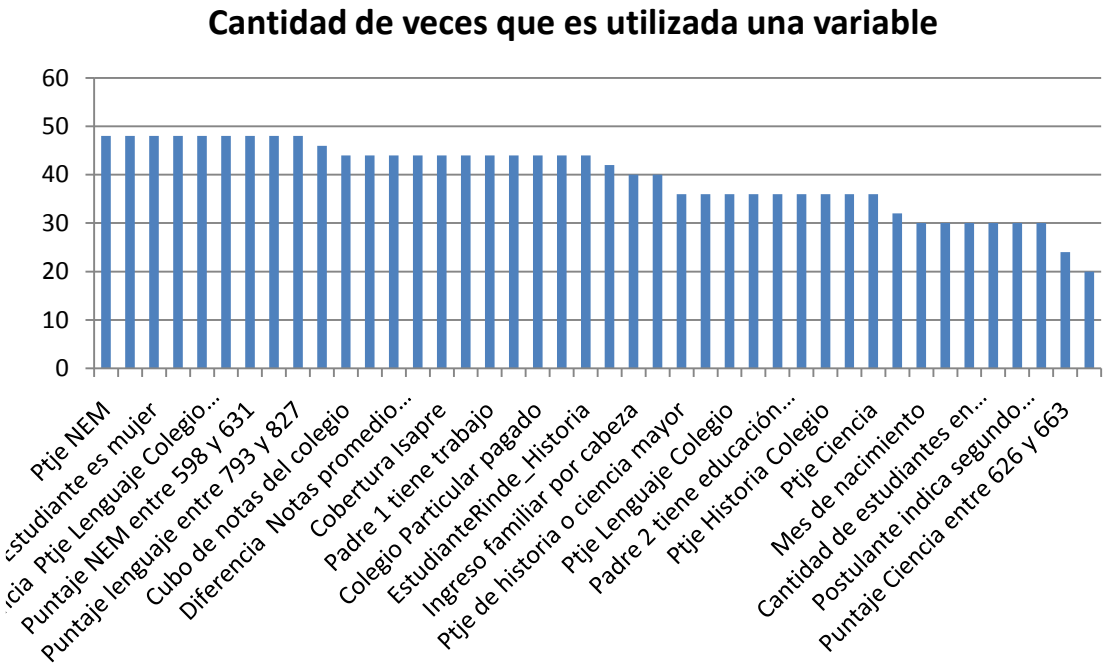
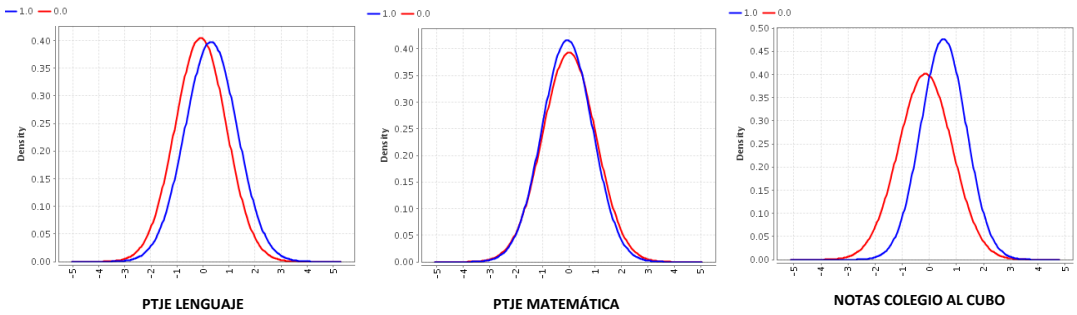


Gráfico Nº 6. Cantidad de veces que se utilizan las variables

En los gráficos a continuación la clase 1 en azul corresponde a Postulante Destacado mientras que la clase 0 en rojo corresponde a postulante no destacado. Se presenta a continuación los gráficos de distribución por variable y clase del mejor modelo presentado en la sección 8.4.1 anterior. Cabe destacar que estas variables se repiten para los otros modelos.

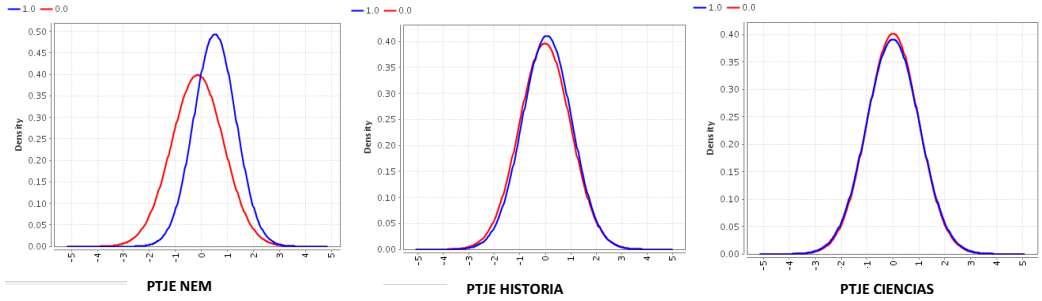
Ahora bien sobre los puntajes de sus pruebas PSU, en cuanto al puntaje de Lenguaje existe una media mayor para los Postulantes Destacados que desplaza la distribución a la derecha. Mientras que para el puntaje de matemática las distribuciones son muy similares, sólo se aprecia una leve densidad mayor en los Postulantes Destacados y una leve mayor media para los

Postulantes No destacados. Por su parte la diferencia en la distribución respecto al cubo de las notas es marcada la media superior para los Postulantes Destacados, así como también una mayor densidad y menor varianza. Estas distribuciones se aprecian en los Gráficos Nº 7.



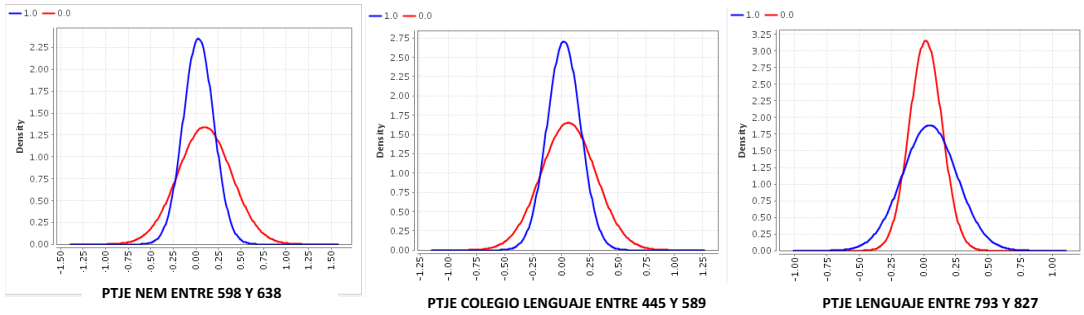
Gráficos Nº 7. Distribución para Notas, puntajes de Lenguaje y Matemática.

Por su parte el Puntaje NEM también muestra una media mucho mayor para los Postulantes Destacados, mientras que las pruebas de Ciencias e Historia y Ciencias Sociales no presentan diferencias en la distribución entre las clases.



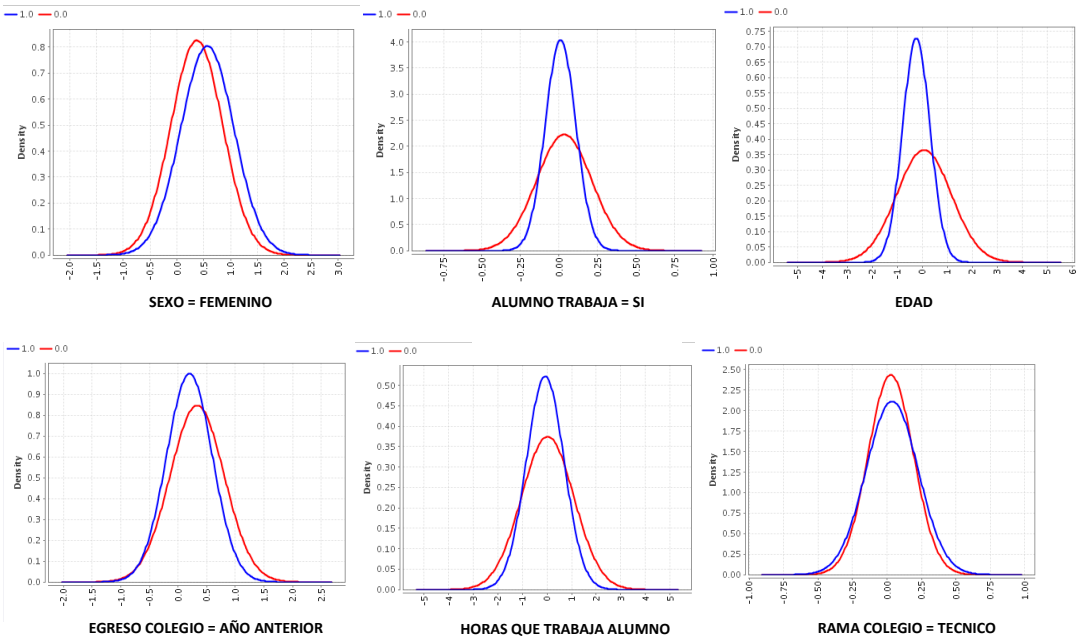
Gráficos Nº 8. Distribución para puntajes NEM, Ciencia e Historia.

En los gráficos Nº 9 siguientes, podemos ver que existe una clara diferencia entre la varianza de los Postulantes destacados versus los No destacados, es decir que la mayor cantidad de Postulantes Destacados se encuentran en un grupo NEM entre 598 y 638, como también los colegios de los Postulantes Destacados están más concentrados en los puntaje en lenguaje de hasta 589.



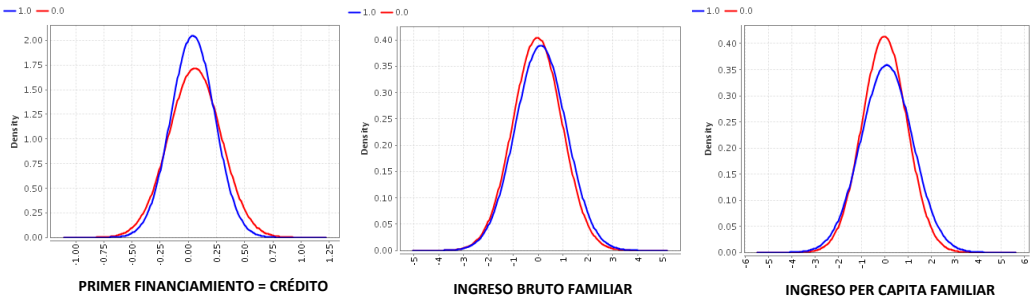
Gráficos Nº 9. Grupo Puntaje NEM estudiante y Grupo Colegio Puntaje Lenguaje.

En cuanto a las variables sobre las características personales de un estudiante se aprecian distribuciones diferentes para los estudiantes que trabajan y además la cantidad de horas que trabajan y la edad, así como también el sexo femenino. Las diferentes distribuciones se encuentran en los gráficos Nº 10. Así como también es clara la diferencia en la media para los Postulantes No Destacados que egresaron del colegio un año anterior. Los colegios técnicos tienen una menor varianza en Postulantes No Destacados.



Gráficos Nº 10. Distribución para características de los estudiantes.

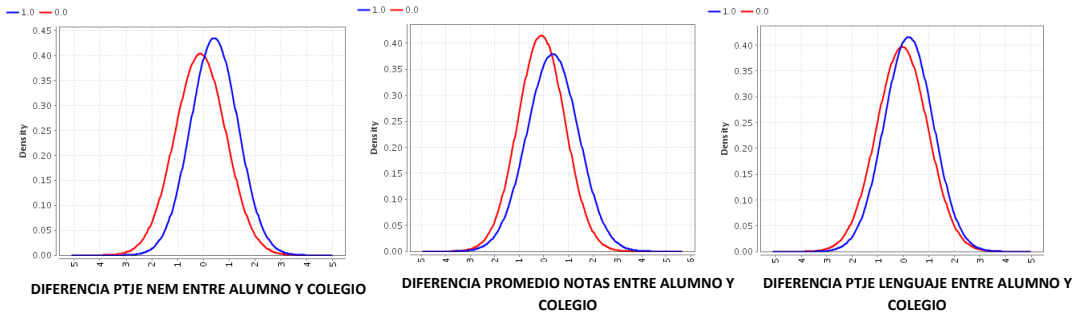
En cuanto a variables referentes a situación financiera de los estudiantes se aprecia que los estudiantes que planean utilizar crédito tienen una mayor concentración y menor varianza para los Postulantes Destacados. Mientras que el ingreso familiar no distribuye las clases en forma notoriamente diferente, si lo hace el ingreso por persona en una familia, en donde se aprecia una mayor media para los Postulantes Destacados pero también menor varianza.



Gráficos Nº 11. Distribución sobre variables de información financiera.

Además cabe destacar que se presentan distribución de interés para las diferencias obtenidas en puntaje por el estudiante y el promedio de su colegio, por ejemplo los Postulantes Destacados

tienen mayor media y menor varianza en las diferencia en el promedio NEM con su colegio. Así también ocurre con la nota del colegio aunque aumenta la varianza. En este caso nuevamente las notas de lenguaje tienen mayor media y menor varianza para los postulantes destacados.



Gráficos Nº 12. Distribución de diferencias de puntajes entre estudiante y promedio de su colegio.

Finalmente muchas variables no tienen distribuciones diferentes entre las clases. Las variables de diferencia del estudiante con el promedio de matemática de su colegio, los puntajes de historia y ciencia, si el estudiante rinde historia no tienen distribuciones diferentes. De la misma forma las características del colegio sobre particular pagado, y cantidad de estudiantes en el último o penúltimo curso. Se destaca que la educación de los padres 1 y 2 no tienen distribuciones diferentes, así como tampoco si tienen trabajo o no.

En resumen, se observa en muchos de los arboles de decisión los patrones siguientes, en que el puntaje NEM para que un estudiante sea Postulante Destacado parte desde 702, el puntaje de Lenguaje y Comunicación debe estar por sobre los 675, mientras que la edad debe ser menor a 20 años, y el promedio de notas debe ser mayor a 64. En árboles de decisión la rama del colegio no permite predecir claramente si un estudiante es un Postulante Destacado o un Postulante No Destacado. Las variables explicativas encontradas en estos experimentos son el puntaje de lenguaje y comunicación y el puntaje NEM, tanto del estudiante como del colegio. Además la edad menor a 20 años y el género femenino son variables explicativas muy fuertes en variados modelos. Las variables sobre situación económica tales como Ingreso por persona del grupo familiar, si un estudiante trabaja y cuantas personas trabajan se encuentran en variados modelos pero con un peso menor. El puntaje de matemática para el estudiante se repite en todos los modelos, pero las diferencias de distribución entre las clases es mucho menor que los otros puntajes.

8.5 Comparación de los enfoques de solución

A continuación se comparan los resultados de los cuatro enfoques, singular, secuencial, combinado y combinado secuencial utilizando el mejor modelo de cada enfoque y sus indicadores de desempeño como accuracy y el ratio verdadero positivo. Los modelos singulares en base a los indicadores de desempeño de los 3 modelos distan de obtener un buen desempeño predictivo debido a la complejidad de predicción que no alcanzan a abarcar. Por ello tienen un Ratio Verdadero Positivo muy bajo con respecto a los otros tres enfoques de solución. En el Gráfico Nº 13 se presenta el mejor modelo de cada enfoque.

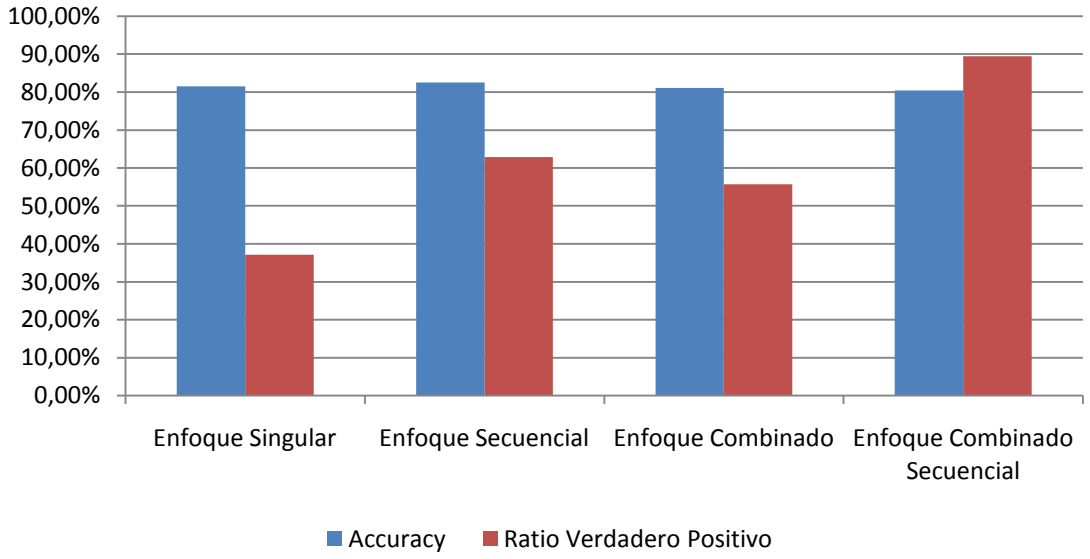


Gráfico Nº 13. Desempeño y Ratio Verdadero Positivo por Enfoque

También se observa que todos los enfoques tienen un accuracy similar por sobre un 80%, pero la diferencia radico en el Ratio Verdadero Positivo, pues es este indicador el que permite detectar cuanto sirve un modelo para detectar la clase positiva, y en esta tesis detectar a los Postulante Destacado. Por su parte el enfoque secuencial y el enfoque combinado tienen similar accuracy y Ratio Verdadero Positivo, aunque el enfoque secuencial es superior en el Ratio Verdadero Positivo por 5 puntos. En este mismo sentido el enfoque secuencial resulta más eficaz puesto que no tiene tantos modelos con un sobre ajuste alto como es el caso del enfoque combinado. Finalmente, es el Enfoque Combinado Secuencial el que se destaca por un Ratio Verdadero Positivo cercano al 90%. Además en algunos de sus modelos este indicador de desempeño alcanza hasta un 100% de acierto. Si bien es cierto que algunos modelos del enfoque

combinado secuencial sufren de sobre ajuste, son fáciles de detectar y desechar, para conservar los modelos útiles para predecir los resultados de futuros postulantes a FEN.

Los Modelos Combinados Secuenciales permiten entonces alcanzar un desempeño de predicción alto, hasta un 80,4% pero también la flexibilidad necesaria para escoger si enfocarse en el Ratio Verdadero Positivo. Esto por ejemplo se ve reflejado en el modelo con un 76,74% de Accuracy y un Ratio Verdadero Positivo de 100%. Es entonces parte del tomador de decisión apuntar los esfuerzos de captación mediante un modelo con alto Accuracy o un modelo con alto Ratio Verdadero Positivo. Un modelo con alto Ratio Verdadero Positivo se enfoca solo en llamar a la mayor cantidad de Postulantes Destacados a menor error, siendo más eficiente en el uso de los recursos gracias a su precisión en la predicción. Por su parte el modelo con mayor Accuracy es más eficaz a la hora de contactar a los estudiantes, pues sugiere una predicción acertada en forma más amplia a menor error.

Lo anterior guarda relación con los costos del error tipo I y los costos del error tipo II. El costo del error tipo I, es no contactar a un estudiante que en realidad era Postulante Destacado cuando se predice Postulante No Destacado. Es decir, dejar pasar a un Postulante Destacado, el que podría ser atraído por otra universidad y eventualmente postular a otra universidad. El costo asociado al error tipo I es entonces, el costo de oportunidad de no contactar con un postulante destacado. Por otro lado el costo del error tipo II, guarda relación con contactar a estudiantes que son en realidad Postulantes No Destacados cuando se predice Postulantes Destacados. El costo asociado radica en el costo de contactar al estudiante, más el costo asociado de ofrecer un incentivo (si existe) a postular a la universidad, más el costo asociado a tener un Postulante No Destacado cursando la carrera, tales como costos de ramos reprobados y atrasos en el egreso universitario, y finalmente se agrega el costo de oportunidad de no contactar a un Postulante Destacado debido a que el tiempo y recursos del proceso de captación de estudiantes son escasos y no permite llamar a todos los estudiantes. Se desprende intuitivamente que el costo del error tipo II es mucho mayor que el costo del error tipo I, de hecho el costo del error tipo I, en cuanto a costo de oportunidad está incluido en parte en el costo del error tipo II. Por lo tanto se sugiere reducir el error tipo II todo lo posible, sujeto a los recursos disponible para captación. Es decir que se sugiere el uso del modelo con mayor Accuracy pues el Ratio Verdadero Positivo alcanza casi un 90%, siendo uno de los modelos que más reduce el error tipo II. En términos prácticos para resumir estos resultados, actualmente el porcentaje de alumnos destacados en FEN es de un 22%. Con la aplicación del modelo predictivo, es posible contactar para postular a FEN en forma certera a un 89,4% de Postulantes Destacados sobre el total de estudiantes. Ahora bien, si los estudiantes son aceptados o no, depende del proceso de selección en base a ranking de postulación.

8.6 Variables explicativas de los enfoques de solución

Las variables explicativas no presentan mayor variación entre un enfoque y otro, de hecho los resultados de variables del Enfoque Secuencial resultan claves para los resultados de predicción en el Enfoque Secuencial Combinado. Por lo tanto a continuación se presentan los principales hallazgos sobre las variables explicativas que son transversales a los enfoques de solución. Para presentar estas variables podemos clasificarlos en tres ámbitos. El primer ámbito es la distribución diferente entre las clases, el segundo es la cantidad de usos de las variables, mientras que el tercero guarda relación con cotas encontradas para variables de puntaje.

El primer ámbito sobre distribución diferente, se desprende de la comparación de normales por clase resultante del Enfoque Secuencial Combinado. Estas variables muestran normales diferentes entre los Postulantes Destacados y los Postulantes No Destacados. Las principales variables de este ámbito son las notas, el sexo femenino, la edad, el puntaje de lenguaje y el puntaje NEM, si el alumno trabaja y el tipo de egreso del colegio. Aparecen también variables de financiamiento tales como si el alumno tiene como principal financiamiento el crédito y el ingreso bruto familiar. Toma importancia también las diferencias de puntaje del alumno y el promedio de su colegio en cuanto al puntaje NEM y al puntaje de lenguaje. Cabe destacar que matemáticamente, las diferencias del puntaje NEM del alumno con el promedio del puntaje NEM del colegio son también una expresión de un ranking. Si bien es cierto que no es la misma métrica utilizada por la métrica actual de puntaje NEM en base a ranking, es en términos prácticos una comparación del alumno con su propio colegio y en forma intuitiva estos resultados muestran que el ranking es una variable explicativa de desempeño pero de menor forma que las otras variables justo antes mencionadas.

En cuanto al segundo ámbito de análisis de variables utilizadas, podemos estudiar la cantidad de usos que tienen los modelos de ellas. En particular se destaca que en todos los enfoques de solución están presentes las notas, el puntaje NEM, el sexo mujer y la diferencia de puntaje de lenguaje entre el alumno y el promedio de su colegio. Luego se observan variables presentes en los resultados de predicción más altos que se evidencian en a lo largo del Capítulo 8, las que son la diferencia del puntaje NEM entre el alumno y el colegio, así como también la diferencia de notas, si al menos uno de los padres del alumno trabaja, la edad del alumno, el ingreso bruto familiar por persona y si el colegio es particular.

El tercer ámbito de estudios sobre las variables explicativas, hace referencia al uso de cotas por los modelos de predicción. Se observa que en los modelos de DT a lo largo de los enfoques de solución, se generan cotas o límites que concentran Postulantes Destacados o bien Postulantes No Destacados. Se observa que los Postulantes Destacados se concentran en las cotas de

puntaje de Lenguaje entre 793 y 827 puntos, así también por sobre el límite del puntaje de lenguaje de 675 puntos. En cuanto al puntaje NEM se observa una cota inferior de 702 puntos para los Postulantes Destacados, mientras que en cuanto al puntaje de matemática se observan cotas inferiores que varían alrededor de 700 puntos. Los Postulantes No Destacados por su parte se concentran entre las cotas de puntaje de lenguaje de 445 y 589 puntos, a su vez que se concentran entre las cotas de puntaje NEM entre 598 y 638 puntos.

Resulta clave entonces considerar el puntaje de lenguaje, pues guarda relación con la capacidad de expresión y comprensión de la comunicación, herramienta clave para profesionales relacionados a los negocios según evidencia Nohria y Eccles (1992) en diversas estructuras organizacionales. De este análisis de variables se desprenden las características de los postulantes más influyentes en la predicción de desempeño académico. Cabe destacar que este análisis se construye como resultado de la comparación de los enfoques, y se observa entonces que las mismas variables están presentes en diferentes enfoques. La diferencia en el resultado de predicción por cada enfoque radica en la ponderación y uso que da cada modelo a estas variables.

Capitulo 9 –Conclusiones

La aplicación de esta tesis busca ser un apoyo para la selección de postulantes universitarios en dos puntos clave, en el corto plazo se busca apoyar en el momento de contactar a los estudiantes para la selección. Luego, en el mediano plazo apoyar la dirección de planes de difusión universitaria. En el corto plazo ocurre la etapa de captación de estudiantes. Esto ocurre después de la entrega de resultados de las pruebas de selección a la casa de estudios, y con ello es posible identificar en forma temprana y rápida los mejores estudiantes que denominamos Postulantes Destacados. Para esta tesis un Postulante Destacado cumple con el perfil de interés indicado por la Dirección de Escuela.

El perfil de selección de interés guarda relación con un buen desempeño académico futuro. Para esta tesis, un buen desempeño quiere decir tener una nota promedio ponderada superior a 5,5 en los últimos tres años de estudio, según lineamientos de la Dirección de Escuela. Estos equivalen al 22,2% del total de los estudiantes. Se quita el ruido entonces de los primeros años en que los estudiantes dependen de su formación académica secundaria como base. En particular este perfil de estudiante mantiene un alto desempeño académico y presenta mejores notas también en el primer y segundo año.

Los resultados del estudio muestran que es posible aumentar significativamente la cantidad de candidatos de alto desempeño invitados a postular en primera prioridad. Esto permite mejorar el proceso clave de captación, dirigiendo los recursos y aprovechando mejor la ventana de tan sólo días para invitarlos a ser parte de la casa de estudios. Esta tesis sustenta sus raíces en la teoría del Knowledge Discovery Database, apoyándose en todas sus etapas para obtener un resultado útil para la casa de estudios de esta caso, en donde es posible priorizar en forma efectiva la limitada cantidad de invitaciones y promoción de la casa de estudios hacia sus candidatos según el perfil definido como objetivo. Es posible identificar entre un 100% y 80,95% de los Postulantes Destacados en forma correcta utilizando el enfoque de Modelos Combinados Secuenciales. Los modelos Combinados Secuenciales alcanzan un Accuracy de hasta un 80,4% y rondan la moda de un Accuracy de 79,48%. De hecho el mejor modelo, que se sugiere utilizar alcanza un 80,4% de Accuracy en el proceso de selección de postulantes y permite dar un paso sustantivo en la cantidad de Postulantes Destacados contactados. Según se aprecia estadísticamente el porcentaje de alumnos destacados en FEN es de un 22% actualmente y con la aplicación del modelo predictivo, es posible contactar para postular a FEN en forma certera a un 89,4% de Postulantes Destacados sobre el total de estudiantes. Sin embargo, si estos Postulantes Destacados se convierten en estudiantes universitarios de FEN o no, depende del proceso de selección universitaria en base a ranking de prioridades de carreras.

En el desarrollo de esta tesis se generaron modelos predictivos basados en cuatro enfoques. El primer enfoque de Modelo Singular no es suficiente para abarcar el problema de predicción detrás de los datos. Por su parte el segundo enfoque de Modelo en Serie y tercer enfoque de Modelos combinados arroja guías y pistas sobre las variables más importantes pero aún así el Ratio Verdadero Positivo de este enfoque no alcanza un nivel suficiente. Es entonces, el cuarto enfoque el que genera los mejores resultados de esta tesis. El enfoque de Modelos Combinados Secuenciales utiliza las tres técnicas de aprendizaje, Support Vector Machine, Árbol de Decisión y Regresión Logística, para generar un mejor desempeño de predicción combinando modelos de menor desempeño.

Cabe destacar que las técnicas de Bagging y Stacking resultan ser técnicas clave en el proceso. Por su parte Bagging mejora el desempeño de Árbol de Decisión gracias a su combinación de múltiples modelos de este mismo tipo. Bagging logra extraer resultados útiles de Árboles de Decisión convirtiendo la debilidad que muestra en la variabilidad de la predicción en una fortaleza que mejora el desempeño mucho gracias a predicciones más integras abarcando datos más diversos entre sí. Stacking resulta clave en la generación de modelos combinados integrando los resultados de modelos de diferente tipo en una sola predicción bajo la técnica de aprendizaje de Naive Bayes. Esta complejidad alta en los modelos generados tiene un mayor riesgo de sobre ajuste a los datos. Se controla el sobre ajuste a los datos comparando la predicción con datos no entrenados contra la predicción de datos entrenados a través del proceso de validación cruzada. Se desechan modelos demasiado complejos con más de una técnica de Bagging, o la aplicación de Bagging en RL y SVM. Así como también el uso de más de una técnica de Stacking. Los boosters como Adaboost generan un sobre ajuste al combinarlos con las otras técnicas por lo que se limita su uso.

Los resultados de predicción de esta tesis aumentan la cantidad de estudiantes contactados que se caracterizan por una rápida adaptación al ambiente universitario y su probable mejor performance académica, reduciendo los costos que implican estudiantes de bajo desempeño y difícil adaptación. Las variables utilizadas por este estudio se respaldan con literatura especializada al respecto. Las variables utilizadas por esta tesis corresponden tanto a características del colegio como del estudiante más el desempeño pre universitario del estudiante. En las características del estudiante se busca ahondar en definiciones sobre su núcleo familiar y su actividad, así como también variables que rescaten el carácter del estudiante, por ejemplo si trabaja o no, y a su vez si considera vivir en forma independiente o no, entre otras. De ellas las variables explicativas del perfil de captación para postulantes universitarios destacados o no que más destacan según su desempeño en el largo plazo son los puntajes de Lenguaje y puntaje NEM tanto del estudiante como el promedio del colegio. Lo anterior es revelador con respecto a los estudios actuales que centraban su atención en el puntaje de

matemática como una única variable importante de desempeño. El puntaje de matemática no representa diferencias significativas en las distribuciones de las clases, es así como toma un rol secundario en los modelos de predicción. En términos prácticos, los modelos plantean que si el estudiante presenta un puntaje de matemática suficiente para ingresar a FEN, es decir sobre los 700 puntos, es el puntaje de lenguaje, NEM del estudiante y diferencia del puntaje NEM del estudiante con el promedio del colegio los que marcan mayor diferencia en el desempeño académico futuro. Se observa que los datos geográficos tanto de los estudiantes como del colegio tienen relevancia media pero solo por algunos modelos. Mientras que variables relacionadas con las diferencias entre el puntaje de NEM, Lenguaje y Matemática del estudiante y el promedio de su colegio se mantienen presentes en variados modelos. La distribución de clases marca diferencias con mayores medias a estas diferencias para los Postulantes Destacados.

Desde las características particulares y familiares del estudiante, se destaca que es marcada la distribución de clases sobre el género femenino del estudiante. De hecho todos los modelos consideran un peso importante a que el sexo sea femenino. Además la edad menor a 20 años, si el estudiante trabaja y si el estudiante rinde la PSU el mismo año de egreso son variables explicativas muy influyentes en variados modelos que permiten redirigir los esfuerzos de captación universitaria.

Además de esto, en la literatura a la fecha se han utilizado modelos que incluyen análisis particulares para el caso de cada país, universidad y carrera que les entrega información mucho más rica que la de esta tesis, por ejemplo pruebas de evaluación deportiva para una carrera deportiva o bien exámenes de placement de gobierno para orientar a los estudiantes según sus capacidades. En nuestro país no se realizan pruebas de placement y orientación ni tampoco existe contacto importante con la facultad previa a la PSU. Los registros de contactos con estudiantes se orientan a la transmisión de información de la Facultad hacia el estudiante y no viceversa para obtener información de ellos. Resulta interesante aumentar los datos sobre características para generar predicciones a futuro sobre desempeño.

Los resultados de esta tesis entregan un modelo de Minería de Datos capaz de predecir a los Postulantes Destacados de acuerdo al perfil definido para ellos. Es interesante destacar el perfil de captación puede ser cambiado para apuntar a los estudiantes que la escuela considere de mayor interés para recibir la atención del área de difusión universitaria. Los estudios posteriores apuntan por un lado profundizar el modelo predictivo con mayores variables de ingreso, y por modelar diferentes perfiles de estudiantes diferentes al alto desempeño académico. Sin embargo, en nuestro país resulta necesario realizar pruebas de apoyo para la selección de estudiantes, tales como pruebas de placement y de aptitudes a la carrera. Más aún cuando los beneficios de

gratuidad estatal se extienden a mayor cantidad de estudiantes. La literatura correlaciona fuertemente esas variables con el desempeño futuro, pero en nuestro país ni siquiera se realizan en forma interna por universidad.

Se discute entonces la selección de estudiantes actual con foco fuerte en matemática cuando en realidad, el puntaje de matemática es una señal necesaria pero no suficiente para alcanzar un buen desempeño académico. Resultan claves los puntajes de Lenguaje y NEM tanto del estudiante, como del colegio e incluso se destaca la diferencia entre el promedio del colegio y el estudiante. Esta última variable resulta un buen indicador de que tan bueno es un estudiante con respecto al ambiente formativo de donde proviene.

Las limitaciones de esta tesis guardan relación con las variables utilizadas y su procedencia. Es decir que las predicciones están limitadas a las datos existentes, que como se explica son de menor cantidad de variables a los utilizados en estudios de otros países. A su vez, estos datos provienen de encuestas DEMRE que son completadas por la familia del estudiante y su veracidad es responsabilidad de quienes completaron la encuesta. La investigación se limita a las variables consideradas en esta encuesta DEMRE por su carácter oficial como parte del Proceso de Selección Universitaria, y se dejan fuera otras variables de otras fuentes gubernamentales o universitarias. De la misma forma se limita el perfil de selección de acuerdo a los lineamientos de la Dirección de Escuela y los datos disponibles. Cabe destacar que esta investigación se limita a considerar un buen desempeño académico en base a las notas universitarias dejando de lado la discusión de desempeño académico suficiente. Siendo entonces tarea de estudios posteriores la predicción de otros perfiles de interés al momento de postular a la universidad tales como alumnos desertores. Los datos utilizados en esta tesis se encuentran entre el año 2004 y 2010 para generar un modelo predictivo, y la aplicación de este modelo está limitada a los cambios del contexto en un futuro.

Al predecir si un estudiante es un Postulante Destacado se permite mejorar el proceso de atracción de estos estudiantes en base a llamados e invitaciones telefónicas dirigidas, siendo tan particulares como ofrecer becas o beneficios dirigidos como por ejemplo alojamiento según sea el caso de estudiante de una región extremo del país. Siendo esta una ventaja competitiva con respecto a las otras universidades que realizan procesos similares. La aplicación de estos modelos puede ser integrada al proceso de captación de estudiantes gracias a que es de rápida ejecución una vez que el modelo ya se encuentra entrenado y diseñado. Siendo interesante además seleccionar otro tipo de perfil de estudiante para dirigir los esfuerzos de promoción universitaria, tales como alumnas con el perfil de captación adecuado para aumentar la inclusión de ellas, estudiantes con foco en el outreach universitario o bien estudiantes futuros candidatos de post grado.

ANEXOS

Anexo 1 – Postulantes según dependencia colegio

La siguiente tabla muestra la distribución de postulantes según dependencia de colegio y año.

Año Ingreso	GRUPO DEPENDENCIA	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado / Total Dependencia
2004	1	31	85	116	27%
	2	11	48	59	19%
	3	16	61	77	21%
2005	1	25	98	123	20%
	2	15	53	68	22%
	3	11	53	64	17%
2006	1	24	112	136	18%
	2	12	78	90	13%
	3	8	41	49	16%
2007	1	27	114	141	19%
	2	20	57	77	26%
	3	10	35	45	22%
2008	1	37	93	130	28%
	2	19	66	85	22%
	3	14	31	45	31%
2009	1	35	89	124	28%
	2	21	55	76	28%
	3	9	42	51	18%
2010	1	40	138	178	22%
	2	14	48	62	23%
	3	12	41	53	23%
Total general		**411**	**1438**	**1849**	

Anexo 2 – Postulantes según dependencia colegio

La siguiente tabla muestra la edad por año de ingreso

Año Ingreso / Edad	Postulante Destacado	Postulante No Destacado	Total general	% Postulante Destacado	% Postulante No Destacado
2004	58	194	252	23%	77%
18	1		1	100%	0%
19	43	106	149	29%	71%
20	10	39	49	20%	80%
21	3	21	24	13%	88%
22	1	15	16	6%	94%
Entre 23 y 25		9	9	0%	100%
Mayor a 26		4	4	0%	100%
2005	51	204	255	20%	80%
18	3	2	5	60%	40%
19	32	120	152	21%	79%
20	14	46	60	23%	77%
21	1	16	17	6%	94%
22		12	12	0%	100%
Entre 23 y 25	1	4	5	20%	80%
Mayor a 26		4	4	0%	100%
2006	44	231	275	16%	84%
18	1	5	6	17%	83%
19	29	139	168	17%	83%
20	9	49	58	16%	84%
21	1	14	15	7%	93%
22	2	9	11	18%	82%
Entre 23 y 25	2	13	15	13%	87%
Mayor a 26		2	2	0%	100%
2007	57	206	263	22%	78%
18		4	4	0%	100%
19	41	121	162	25%	75%
20	14	53	67	21%	79%

21	1	13	14	7%	93%
22		6	6	0%	100%
Entre 23 y 25	1	8	9	11%	89%
Mayor a 26		1	1	0%	100%
2008	70	190	260	27%	73%
18	1		1	100%	0%
19	49	107	156	31%	69%
20	20	50	70	29%	71%
21		13	13	0%	100%
22		6	6	0%	100%
Entre 23 y 25		12	12	0%	100%
Mayor a 26		2	2	0%	100%
2009	65	186	251	26%	74%
18		2	2	0%	100%
19	47	107	154	31%	69%
20	13	46	59	22%	78%
21	2	11	13	15%	85%
22		5	5	0%	100%
Entre 23 y 25	3	13	16	19%	81%
Mayor a 26		2	2	0%	100%
2010	66	227	293	23%	77%
18	2	3	5	40%	60%
19	37	118	155	24%	76%
20	21	58	79	27%	73%
21	2	24	26	8%	92%
22	2	10	12	17%	83%
Entre 23 y 25	2	9	11	18%	82%
Mayor a 26		5	5	0%	100%
Total general	**411**	**1438**	**1849**	**22%**	**78%**

Anexo 3 – Listado de variables

A continuación se presenta el listado de variables utilizadas y su descripción.

Variable	Descripción
Mes de nacimiento	Mes de nacimiento
Cubo de notas del colegio	Promedio de notas al cuadrado
Ptje NEM	Ptje NEM
Diferencia Ptje NEM Colegio y Estudiante	Diferencia Ptje NEM Colegio y Estudiante
Diferencia Notas promedio Colegio y Estudiante	Diferencia Notas promedio Colegio y Estudiante
Ptje Lenguaje	Ptje Lenguaje
Estudiante es mujer	Estudiante es mujer
Edad estudiante	Edad estudiante
Diferencia Ptje Lenguaje Colegio y Estudiante	Diferencia Ptje Lenguaje Colegio y Estudiante
Estudiante egreso del colegio el año anterior	Estudiante egreso del colegio el año anterior
Ptje de historia o ciencia mayor	Ptje de historia o ciencia mayor
Puntaje NEM entre 598 y 631	Puntaje NEM entre 598 y 631
Ingreso familiar	Ingreso familiar
Cobertura Isapre	Cobertura Isapre
Ingreso familiar por cabeza	Ingreso familiar por cabeza
Cantidad de horas que trabaja estudiante	Cantidad de horas que trabaja estudiante
Ptje Historia	Ptje Historia
Puntaje Ciencia entre 626 y 663	Puntaje Ciencia entre 626 y 663
Ptje Matematica	Ptje Matematica
Puntaje Colegio lenguaje entre 485 y 519	Puntaje Colegio lenguaje entre 485 y 519
Padre 2 tiene trabajo	Padre 2 tiene trabajo
Padre 1 tiene trabajo	Padre 1 tiene trabajo
Diferencia Ptje Matematica Colegio y Estudiante	Diferencia Ptje Matematica Colegio y Estudiante
Postulante estudia en la misma comuna	Postulante estudia en la misma comuna
Ptje Lenguaje Colegio	Ptje Lenguaje Colegio
Cantidad de estudiantes en penúltimo curso	Cantidad de estudiantes en penúltimo curso
Puntaje lenguaje entre 793 y 827	Puntaje lenguaje entre 793 y 827
Colegio Particular pagado	Colegio Particular pagado
Estudiante si trabaja	Estudiante si trabaja
Colegio un solo sexo	Colegio un solo sexo
Postulante indica primer financiamiento de crédito	Postulante indica primer financiamiento de crédito
Cantidad de estudiantes en último curso	Cantidad de estudiantes en último curso
EstudianteRinde_Historia	EstudianteRinde_Historia
Postulante indica segundo financiamiento de beca	Postulante indica segundo financiamiento de beca
Padre 2 tiene educación secundaria	Padre 2 tiene educación secundaria
Colegio técnico	Colegio técnico
Estudiante_Vivirá_Solo	Estudiante_Vivirá_Solo

Ptje NEM Colegio	Ptje NEM Colegio
Ptje Historia Colegio	Ptje Historia Colegio
Padre 1 tiene educación superior	Padre 1 tiene educación superior
Ptje Ciencia	Ptje Ciencia
Postulante indica primer financiamiento de beca	Postulante indica primer financiamiento de beca

Anexo 4 – Modelos Enfoque Singular

A continuación se encuentra la descripción del enfoque singular incluyendo la grilla de los parámetros.

Técnica	Tipo de Aplicación	Grilla de Parametros	Descripción
Clasificador	Clasificador	First Cost y Second Cost	El proceso se ejecuta asignandoles costos diferentes al error tipo I y error tipo II, estos costos son los First y Second Cost.
	Sin Clasificador	No aplica	No aplica
Aprendizaje	Support Vector Machine	C: 0,05,1,10,20,25,50,70,100,1000	Técnica basada en vectores, donde C corresponde al error de clasificación.
	Árboles de Decisión	Profundidad: 5,10,15,20	Técnica de árboles de decisión, la profundidad indicada cuantos nodos o hojas puede tener una rama.
		Minimo Split: 2,5,10,15,20	Split significa cual es el tamaño minimo se separación
		Poda: si / no	Significa recortar ramas y hojas del arbol luego de construido
		Prepoda: si / no	Significa recortar ramas y hojas del arbol por evaluación
	Regresión logística	C: 0,05,1,10,20,25,50,70,100,1000	Técnica basada en regresiones, donde C corresponde al error de clasificación.

Anexo 5 – Modelos de Enfoque de Modelos Secuenciales

A continuación se encuentra la descripción de técnicas para los modelos secuenciales.

ID Modelo en Serie	Clusterización	Selección de Variables por Árbol	Selección de Variables por SBE	Uso de Clasificador	Aprendizaje
1	Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Support Vector Machine
2	Sin Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Support Vector Machine
3	Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Árboles de Decisión
4	Sin Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Árboles de Decisión
5	Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Regresión logística
6	Sin Cluster	Decision Tree	Sequential Backward Elimination	Clasificador	Regresión logística
7	Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Support Vector Machine
8	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Support Vector Machine
9	Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Árboles de Decisión
10	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Árboles de Decisión

11	Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Regresión logística
12	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Clasificador	Regresión logística
13	Cluster	Decision Tree	Sin SBE	Clasificador	Support Vector Machine
14	Sin Cluster	Decision Tree	Sin SBE	Clasificador	Support Vector Machine
15	Cluster	Decision Tree	Sin SBE	Clasificador	Árboles de Decisión
16	Sin Cluster	Decision Tree	Sin SBE	Clasificador	Árboles de Decisión
17	Cluster	Decision Tree	Sin SBE	Clasificador	Regresión logística
18	Sin Cluster	Decision Tree	Sin SBE	Clasificador	Regresión logística
19	Cluster	Sin Decision Tree	Sin SBE	Clasificador	Support Vector Machine
20	Sin Cluster	Sin Decision Tree	Sin SBE	Clasificador	Support Vector Machine
21	Cluster	Sin Decision Tree	Sin SBE	Clasificador	Árboles de Decisión
22	Sin Cluster	Sin Decision Tree	Sin SBE	Clasificador	Árboles de Decisión
23	Cluster	Sin Decision Tree	Sin SBE	Clasificador	Regresión logística
24	Sin Cluster	Sin Decision Tree	Sin SBE	Clasificador	Regresión logística
25	Cluster	Decision Tree	Sin SBE	Sin Clasificador	Support Vector Machine
26	Sin Cluster	Decision Tree	Sin SBE	Sin Clasificador	Support Vector Machine
27	Cluster	Decision Tree	Sin SBE	Sin Clasificador	Árboles de Decisión
28	Sin Cluster	Decision Tree	Sin SBE	Sin Clasificador	Árboles de Decisión
29	Cluster	Decision Tree	Sin SBE	Sin Clasificador	Regresión logística
30	Sin Cluster	Decision Tree	Sin SBE	Sin Clasificador	Regresión logística
31	Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Support Vector Machine

32	Sin Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Support Vector Machine
33	Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Árboles de Decisión
34	Sin Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Árboles de Decisión
35	Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Regresión logística
36	Sin Cluster	Sin Decision Tree	Sin SBE	Sin Clasificador	Regresión logística
37	Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Support Vector Machine
38	Sin Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Support Vector Machine
39	Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Árboles de Decisión
40	Sin Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Árboles de Decisión
41	Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Regresión logística
42	Sin Cluster	Decision Tree	Sequential Backward Elimination	Sin Clasificador	Regresión logística
43	Cluster	Sin Decision Tree	Sequential Backward Elimination	Sin Clasificador	Support Vector Machine
44	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Sin Clasificador	Support Vector Machine
45	Cluster	Sin Decision Tree	Sequential Backward	Sin Clasificador	Árboles de Decisión

46	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Sin Clasificador	Árboles de Decisión
47	Cluster	Sin Decision Tree	Sequential Backward Elimination	Sin Clasificador	Regresión logística
48	Sin Cluster	Sin Decision Tree	Sequential Backward Elimination	Sin Clasificador	Regresión logística

REFERENCIAS

Acikkar, M., & Akay, M. F. (2009). Support vector machines for predicting the admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. *Expert Systems with Applications*, *36*(3), 7228-7233.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for Data Mining applications (Vol. 27, No. 2, pp. 94-105). ACM.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. The quarterly journal of economics, 488-500.

Al-Alwan, A. F. (2014). Modeling the Relations among Parental Involvement, School Engagement and Academic Performance of High School Students. International Education Studies, 7(4), 47-56.

Améstica Rivas, L., Gaete Feres, H., & Llinas-Audet, X. (2014). Segmentación y clasificación de las universidades en Chile: desventajas de inicio y efectos de las políticas públicas de financiamiento. Ingeniare. Revista chilena de ingeniería, 22(3), 384-397.

Arruabarrena, R. López-Cuadrado, J., Pérez, T. A., Vadillo, J. Á., (2002). Integrating adaptive testing in an educational system. In First International Conference on Educational Technology in Cultural Context.

Ayers, E., Nugent, R., & Dean, N. (2009). A Comparison of Student Skill Knowledge Estimates. International Working Group on Educational Data Mining.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society, 54(6), 627-635.

Beranuy, M., Oberst, U., Carbonell, X., &Chamarro, A. (2009). Problematic Internet and mobile phone use and clinical symptoms in college students: The role of emotional intelligence. Computers in human behavior, 25(5), 1182-1187.

Breiman, L., Friedman, J. H., 01shen, R. A., & Stone, C. J. (1984). Classification and regression trees. New York: Chapman and Hall.

Brugler, J., & Linton, O. (2014). Single stock circuit breakers on the London Stock Exchange: do they improve subsequent market quality? (No. CWP07/14). cemmap working paper, Centre for Microdata Methods and Practice.

Bumacov, V., & Ashta, A. (2011, June). The conceptual framework of credit scoring from its origins to microfinance. In Second European Research Conference on Microfinance.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in higher education, 12(2), 155–187.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. Review of educational Research, 55(4), 485–540.

Centros de Estudios MINEDUC. (2012, septiembre 30). Serie Evidencias: Deserción en la educación superior en Chile.

Chickering, A. W. (1991). Applying the seven principles for good practice in under graduate education. New directions for teaching and learning, 47.

Christenson, S. L., & Huebner, E. S. (2010). A study of the factorial invariance of the Student Engagement Instrument (SEI): Results from middle and high school students. School Psychology Quarterly, 25(2), 84.

Coase, R. H. (1937). The nature of the firm. economica, 4(16), 386-405.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. Expert Systems with Applications, 40(11), 4715-4729.

Diaz, D., Theodoulidis, B., & Sampaio, P. (2011). Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. Expert Systems with Applications, 38(10), 12757–12771.

Durkheim, E. (1951). Suicide: A study in sociology (JA Spaulding & G. Simpson, trans.). Glencoe, IL: Free Press.(Original work published 1897).

Duckworth, A. L., & Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. Psychological science, 16(12), 939-944.

Pelleg, D and Andrew Moore "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" by, Proceedings of the Seventeenth International Conference on Machine Learning, 2000

Eskew, R. K., & Faley, R. H. (1988). Some determinants of student performance in the first college-level financial accounting course. Accounting Review, 137-147.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Freeman, P. R. (1970). A multivariate study of students' performance in university examinations. Journal of the Royal Statistical Society. Series A (General), 38-55.

Goldman, R. D., & Slaughter, R. E. (1976). Why college grade point average is difficult to predict. Journal of Educational Psychology, 68(1), 9.

Goleman, D. (1998). Working with emotional intelligence. New York: Bantam Books.

González, L. E., & Uribe, D. (2002). Estimaciones sobre la "repitencia" y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones. Revista Calidad en la Educación Consejo Superior de Educación Diciembre del, 2002, 77.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.

Han, J., Kamber, M., & Pei, J. (2011). Minería de datos: concepts and techniques: concepts and techniques. Elsevier.

Haselgrove, S. (1994). Why the student experience matters. The student experience, 3-8.

Heiner, C. Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., (2005). An educational data mining tool to browse tutor-student interactions: Time will tell. In Proceedings of the Workshop on Educational Data Mining, National Conference on Artificial Intelligence (pp. 15-22). AAAI Press.

House, J. D. (2000). The effect of student involvement on the development of academic self-concept. The journal of social psychology, 140(2), 261-263.

Huang, Z., & Zheng, S. Yang, Q., Wang, X., (2007, November). Research of student model based on bayesian network. In Information Technologies and Applications in Education, 2007. ISITAE'07. First IEEE International Symposium on (pp. 514-519). IEEE.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6, 429–449.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13(1), 61-72.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Laudon, K. C., & Laudon, J. P. (2010). Management Information Systems: Managing the Digital Firm. 2011.

Levitz, R., & Noel, L. (2000). The earth-shaking, but quiet revolution, in retention management. Retrieved on August, 6, 2004.

Li, J., & Zaïane, O. (2004). Combining usage, content, and structure data to improve web site recommendation. E-Commerce and Web Technologies, 313-315.

Linoff, G. S., & Berry, M. J. (2011). Minería de datos techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Lowis, M., & Castley, A. (2008). Factors affecting student progression and achievement: prediction and intervention. A two- year study. Innovations in education and teaching international, 45(4), 333-343.

Loveman, J. (2003). "Diamonds in the data mine". Harvard Business Review. The Magazine, May, 2003.

Luan, J. (2002). Data mining and its applications in higher education. New directions for institutional research, 2002(113), 17-36.

Mass, C. F., Ovens, D., Westrick, K., & Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts?. Bulletin of the American Meteorological Society, 83(3), 407-430.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115–133.

Ministerio de Educación. (2012). Serie Evidencias: Deserción en la educación superior en Chile.

Miranda, J., & Vásquez, J. (2015). Student Attrition - Identifying Key Factors and Building a Predictive model in Universidad de Chile context (Vol. 2). Presentado en BAFI, Universidad de Los Andes, Santiago.

Mass, C. F., Ovens, D., Westrick, K., & Colle, B. A. (2002). Doesincreasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. Bulletin of the American Meteorological Society, 83(3), 407.

Mayer, J. D., DiPaolo, M., &Salovey, P. (1990). Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence. Journal of personalityassessment, 54(3-4), 772-781.

Mayer, J. D., &Geher, G. (1996). Emotional intelligence and the identification of emotion. Intelligence, 22(2), 89-113.

Mayer, J. D. (2002). Mayer-Salovey-Caruso emotional intelligence test. Multi-Health Systems, Toronto.

Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2005). Using semantic web mining technologies for personalized e-learning experiences. Proceedings of the web-based education, 461-826.

Mestre, J. M., Guil, R., Lopes, P. N., Salovey, P., & Gil-Olarte, P. (2006). Emotional intelligence and social and academic adaptation to school. Psicothema, 18.

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. Expert Systems with Applications, 38(12), 14984-14996.

Nohria, N., & Eccles, R. G. (1992). Networks and organizations: Structure, form, and action.

Pahl, C. (2003). Managing evolution and change in web-based teaching and learning environments. Computers & Education, 40(2), 99-114.

Payne, W. L. (1985). A study of emotion: developing emotional intelligence; self-integration; relating to fear, pain and desire.

Parker, J. D., Summerfeldt, L. J., Hogan, M. J., &Majeski, S. A. (2004). Emotional intelligence and academic success: Examining the transition from high school to university. Personality and individual differences, 36(1), 163-172.

Padmapriya, A. (2012). Prediction of higher education admissibility using classification algorithms. International Journal of Advanced Research in Computer Science and Software Engineering, 2(11), 330-336.

Pittman, K. (2008). Comparison of data mining techniques used to predict student retention. Nova Southeastern University.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. En ICML (Vol. 1).

Peña-Ayala, A. (2014). Educational Data Mining: A survey and a Data Mining-based analysis of recent works. Expert Systems with Applications, 41(4), 1432–1462.

Pole, Andrew. (2010). "How Target gest most out its guest data to improve marketing ROI". Predictive Analytivs World Washington, DC, Conference. October 18, 2010.

Price, J. L. (1977). The study of turnover. Iowa State Press.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. Canadian Journal of Higher Education, 23(2), 44–64.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81–106.

Rousseeuw P. J, "A graphical aid to the interpretation and validation ofcluster analysis," Journal of Computational Appl Math, vol 20, pp. 53–65, 1987.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational Minería de datos: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618.

Rubin, D. B., & Stroud, T. W. F. (1977). Comparing high schools with respect to student performance in university. ETS Research Report Series, 1977(2), 139-147.

Spady, W. G. (1970). Dropouts from Higher Education: An Interdisciplinary Review and Synthesis. Interchange, 1(1), 64–85.

Sadler, J. (2003). Effectiveness of student admission essays in identifying attrition. Nurse Education Today, 23(8), 620–627.

Support vector machines for predicting the admission decision of a candidate to the school of physical education and sports at Cukurova University. Acikkar (2009)

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. Imagination, cognition and personality, 9(3), 185-211.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., &Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. Personality and individual differences, 25(2), 167-177.

Shurkin, J. N. (1992). Terman's kids: The groundbreaking study of how the gifted grow up. Little, Brown and Co.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and Data Mining for marketing. Decision support systems, 31(1), 127–137.

The Measure of Inteligence, (1916, p.91-92), Genetic Studies of Genius (1925, 1947, 1959)

Terman, L. M., & Oden, M. H. (1947). The gifted child grows up: Twenty-five years' follow-up of a superior group (Vol. 4). Stanford University Press.

Terman's Kids: The Groundbreaking Study of How the Gifted Grow Up by Joel N. hurkin, Little Brown & Co, 1992, ISBN 0-316-78890-2

T. Miskelly, “Interactive student modeling,” pp. 88–94, ACM, 1998

Tekin, A. (2014). Early prediction of students’ grade point averages at graduation: AData Mining approach. Eurasian Journal of Educational Research 54, 207-226.

Tinto, V., & Cullen, J. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. Review of Educational Research, 45(1), 89–125.

Thomas, Liz (2002). Student retention in higher education: the role of institutional habitus. Journal of Education Policy. Volume 17. 423-442

Vapnik, V., & Chervonenkis, A. (1964). A note on one class of perceptrons. Automation and remote control, 25(1).

Vercellis, C. (2009). Business intelligence: Data Mining and optimization for decision making. Editorial John Wiley and Sons.

Vizcaíno, J. A. Olivas, and M. Prieto, "Modelos del estudiante en entornos de aprendizaje colaborativo," tech. rep., Escuela Superior deInformática, Universidad de Castilla-La Mancha, 2000

Williamson, O. E. (1985). The economic intstitutions of capitalism. Simon and Schuster.

Widyantoro, T. R. Ioerger, and J. Yen, "An adaptive algorithmfor learning changes in user Interests," (Kansas City, Missouri, UnitedStates), pp. 405–412, ACM, 1999

Yorke, M. (1998). The Times'"league table" of universities, 1997: a statistical appraisal. Quality Assurance in Education, 6(1), 58-60.

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning, 17(1), 118-133.

Zaiane, O. R., Xin, M., & Han, J. (1998, April). Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on (pp. 19-29). IEEE.

Zorrilla, M., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005). Web usage mining project for improving web-based learning sites. Computer Aided Systems Theory–EUROCAST 2005, 205-210.