“MODELO PREDICTIVO PARA ESTIMAR LA DESERCIÓN DE ESTUDIANTES EN UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR”

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CONTROL DE GESTIÓN

Estudiante: Jonathan Vásquez
Profesor Guía: Jaime Miranda

Santiago, Mayo 2016

Dedicado a papá, mamá y mama.

"…Después de un triunfo (…) termina el partido y hay una sensación de efervescencia, (…) de la adrenalina al tope que genera excitación y felicidad. Pero son apenas cinco minutos…después hay un vacío enorme, grandísimo. Una soledad indescriptible…"

Marcelo Bielsa

AGRADECIMIENTOS

Agradezco a mi familia, quienes forjaron en mí los pilares de la persona que actualmente soy. Gracias por las enseñanzas, paciencia, apoyo incondicional, alegrías y valores entregados; pero principalmente, gracias por ser mi familia. Sin ustedes, esto no sería lo mismo.

También agradezco a todo el círculo de apoyo que tuve durante este proceso. Primero que todo, a mi pareja Pilar, quien me entregó su apoyo, ánimo y fuerzas para terminar; a Ariel La Paz, quien me incentivó y direccionó en los primeros pasos de la profesión académica; a mis compañeros de programa, con quienes disfruté, reí y crecí como profesional; a mis compañeros de la oficina 1905, que me animaron en las últimas etapas de la tesis y hasta el día de hoy recuerdan mis horas "*conectado a la matrix*"; y a mi profesor guía, Jaime Miranda, quien confió en mis capacidades y me corrigió en el difícil camino de esta investigación.

Finalmente, agradezco al ex Director de Escuela de Sistemas de Información y Auditoría, Ariel La Paz; al Secretario de Estudios, César Ortega; a la Directora y Asistente Social de Bienestar Estudiantil, Marcia Oyarce; y al Jefe de Unidad de Sistemas de Infotecnología, Rodrigo López, todos funcionarios de la Facultad de Economía y Negocios que dedicaron su tiempo a la obtención, facilitación y comprensión de los datos usados en esta tesis.

Índice de Contenido

INTRODUCCIÓN 6
Objetivos de Investigación 10
CAPITULO 1 - MARCO TEÓRICO 11
1.1 DESERCIÓN 11
1.1.1 Definición de la Deserción 11
1.1.2 Modelos Teóricos de la Deserción 13
1.2 MINERÍA DE DATOS Y LA DESERCIÓN 21
1.2.1 Identificación de Factores 22
1.2.2 Identificación de Predictores 23
CAPÍTULO 2 – METODOLOGÍA 27
2.1 Selección 29
2.2 Preprocesamiento 30
2.3 Transformación 31
2.4 Minería de Datos 32
2.4.1 Máquinas de Aprendizaje 33
2.4.2 Clusterización 44
2.4.3 Clasificadores 46
2.4.4 Desbalance 47
2.5 Interpretación y Evaluación 48
CAPITULO 3 – CASO DE ESTUDIO 52
3.1 Descripción del Caso Real 52
3.2 Análisis de la deserción 54
4. EXPERIMENTO COMPUTACIONAL 58
4.1 Descripción de la Muestra de Estudio 58
4.1 Bases de Datos 58
4.2 Variables Escogidas 61

4.3 Preprocesamiento y Transformación 65
4.4 Modelos Predictivos 66
4.5 Implementación de Modelos Predictivos en Software 69
4.6 Desempeño de los Modelos 72
CAPÍTULO 5 - RESULTADOS 84
5.1 Predictores 84
5.2 Perfiles por Semestre 111
CAPÍTULO 6 – DISCUSIÓN Y CONCLUSIONES 132
ANEXOS 136
Anexo 1 – Modelos por semestre 136
Anexo 2 – Diccionario de datos de variables obtenidas de la Base de Datos SAD y Becas y Créditos 138
Anexo 3 – Diccionario de datos de variables obtenidas de la Base de DEMRE 140
Anexo 4 – Procesos en Software 143
Anexo 5 – Resumen Variables por Semestre 147
REFERENCIAS 148

Índice de Figuras

Figura 1. Niveles y tipos de deserción.. 12
Figura 2. Modelo planteado por Spady. 14
Figura 3 Modelo planteado por Tinto. 15
Figura 4. Relación entre las variables planteadas por Bean.. 19
Figura 5. Relación entre las variables planteadas por Bean en su segundo estudio.. 21
Figura 6. Etapas del KDD y sus respectivas entradas/salidas.. 29
Figura 7. Representación gráfica de la aplicación del algoritmo SVM. 35
Figura 8. Representación gráfica del resultado obtenido por un algoritmo basado en la máquina de aprendizaje Árbol de Decisiones. 36
Figura 9. Representación gráfica de la segmentación según algoritmo de Árbol de Decisión. 37
Figura 10. Descripción gráfica de un perceptrón. 40
Figura 11. Ejemplo de estructura de una red neuronal artificial. 42
Figura 12. Representación gráfica de la validación cruzada. 50
Figura 13. Ejemplo del semestre 1 para el proceso de transformación de los atributos aplicado en cada semestre. 143
Figura 14. Ejemplo del semestre 1 para el proceso de clusterización y los subprocesos de validación cruzada. 144
Figura 15. Ejemplo del encadenamiento de operadores dentro del operador X-Validations. 144
Figura 16. Encadenamiento de la validación cruzada para cada cluster (lado izquierdo) y sin cluster (lado derecho). 145
Figura 17. Encadenamiento de aplicación de umbral. 145
Figura 18. Ejemplo de encadenamiento para ROS con el subproceso antes del modelo.. 146
Figura 19. Ejemplo de encadenamiento al interior del subrpoceso para la aplicación de la técnica ROS. 146

Índice de Tablas

Tabla 1. Lista de variables planteadas por Bean. 16
Tabla 2. Ejemplo de matriz de confusión. 51
Tabla 3 Vacantes por tipo de ingreso. 53
Tabla 4. Número de tipo de deserciones por año de ingreso. 55
Tabla 5. Cantidad y Distribución de ingresos por año. 58
Tabla 6. Número de registros por clase y por semestre según bases de datos SAD, DEMRE y Bienestar, ... 60
Tabla 7. Identificador y descripción de la aplicación de técnicas de desblance. 66
Tabla 8. Identificador y descripción de la aplicación de clusterización. 67
Tabla 9. Grilla por Máquina de Aprendizaje 68
Tabla 10. Identificador y descripción de la aplicación de clasificador con umbral. 69
Tabla 11 Ranking de 10 modelos con la mejor precisión para el primer semestre. 72
Tabla 12. Desempeño promedio por técnica para el primer semestre. 73
Tabla 13 Ranking de 10 modelos con la mejor precisión para el segundo semestre. 74
Tabla 14 Desempeño promedio por técnica para el segundo semestre 75
Tabla 15 Ranking de 10 modelos con la mejor precisión para el tercer semestre. 76
Tabla 16 Desempeño promedio por técnica para el tercer semestre. 77
Tabla 17 Ranking de 10 modelos con la mejor precisión para el cuarto semestre. 78
Tabla 18 Desempeño promedio por técnica para el cuarto semestre. 79
Tabla 19 Ranking de 10 modelos con la mejor precisión para el quinto semestre. 80
Tabla 20 Desempeño promedio por técnica para el quinto semestre. 81
Tabla 21. Ranking de 10 modelos con la mejor precisión para el sexto semestre. 82
Tabla 22. Desempeño promedio por técnica para el sexto semestre. 82
Tabla 23. Desempeño promedio por técnica para el sexto semestre. 83
Tabla 24 Peso de las variables para el Cluster 1 del primer semestre. 86
Tabla 25 Peso de las variables para el Cluster2 del primer semestre. 87
Tabla 26. Peso de las variables para el Cluster 1 del segundo semestre. 89
Tabla 27. Peso de las variables para el Cluster 2 del segundo semestre. 91
Tabla 28. Peso de las variables para el Cluster 3 del segundo semestre 93
Tabla 29. Peso de las variables para el tercer semestre. 95
Tabla 30. Peso de las variables para el primer cluster del cuarto semestre. 97
Tabla 31. Peso de las variables para el tercer clúster del cuarto semestre. 99

Tabla 32. Peso de las variables para el primer cluster del quinto semestre**.** 101
Tabla 33. Peso de las variables para el segundo cluster del quinto semestre. 103
Tabla 34. Peso de las variables para el primer clúster del sexto semestre. 105
Tabla 35. Peso de las variables para el segundo cluster del sexto semestre. 107
Tabla 36. Peso de las variables para el tercer cluster del sexto semestre. 110
Tabla 37 Perfiles primer clúster para el primer semestre. 112
Tabla 38 Perfiles segundo clúster para el primer semestre. 113
Tabla 39 Perfiles primer clúster para el segundo semestre. 115
Tabla 40 Perfiles segundo clúster para el segundo semestre. 116
Tabla 41 Perfiles segundo clúster para el segundo semestre. 117
Tabla 42 Perfiles para el tercer semestre. 119
Tabla 43 Perfiles primer clúster del cuarto semestre. 120
Tabla 44 Perfiles segundo clúster del cuarto semestre. 122
Tabla 45. Perfiles primer clúster del quinto semestre. 124
Tabla 46 Perfiles segundo clúster del quinto semestre. 125
Tabla 47. Perfiles primer clúster del sexto semestre. 127
Tabla 48 Perfiles segundo clúster del sexto semestre. 128
Tabla 49 Perfiles tercer clúster del sexto semestre 130
Tabla 50. Modelos aplicados por semestre según combinación de técncias de mienría de datos. 137
Tabla 51. Descripción y definición de los datos obtenidos desde base de datos Becas y SAD. 139
Tabla 52. Descripción y definición de los datos obtenidos desde base de datos DEMRE. . 142
Tabla 53. Resumen Variables identificadas como predictores según mejores modelos por Semestre. 147

Índice de Gráficos

Gráfico 1. Comportamiento deserción por semestre. 57

INTRODUCCIÓN

La deserción es entendida como la salida temprana o la falla de un individuo en completar un programa de estudio en el que se inscribió, el cual puede ser de carácter voluntario o involuntario, o bien, de transferencia a otro programa o abandono institucional. La deserción involuntaria ocurre cuando por decisión institucional el estudiante no puede seguir sus estudios por razones académicas o disciplinaria, mientras que la voluntaria se manifiesta a través de la renuncia formal o del abandono no informado del estudiante (Tinto & Cullen,1975). Por otro lado, la deserción del tipo de transferencia entre programas se manifiesta cuando el estudiante se inscribe en otro programa distinto al que estaba matriculado, mientras que la del tipo abandono institucional es cuando el estudiante abandona la institución educacional en la cual estaba inscrito. En el caso de la transferencia, esta puede ser a otro programa de la misma institución o bien, a una institución totalmente diferente (Himmel, 2002).

Investigadores de distintas disciplinas han demostrado su interés en estudiar la deserción. Investigaciones realizadas en el área de la sicología, economía, sociología, y ahora último, minería de datos han aportado en el entendimiento de este fenómeno, siendo los costos asociados a la deserción como la motivaciones de estudio. Algunos ejemplos de estos costos son la frustración y deuda financiera que genera al individuo, el congelamiento del financiamiento a la institución educacional, la pérdida de una vacante que pudo ser utilizada por otro estudiante que sí finalizaría el programa, el estancamiento del desarrollo educacional de la sociedad[1] y la disminución de un profesional que aportaría al país, principalmente en aquellas profesiones mayormente demandadas (Tinto, 2007). Sin embargo, aun con toda la investigación desarrollada desde distintas disciplinas, el fenómeno de la deserción sigue ocurriendo y pocas son las herramientas que se han generado para mitigar sus efectos negativos. Esto genera una oportunidad para que nuevas disciplinas, principalmente aplicadas, respondan al desafío del mejoramiento de la gestión de la deserción.

[1] Es importante recalcar que el estancamiento del desarrollo educacional impacta de manera negativa al Indicador de Desarrollo Humano, o IDH, el cual se utiliza para establecer el nivel de desarrollo del país

Chile está experimentando cambios en el ámbito educacional, manifestándose algunos desde una perspectiva social a través de la participación de agentes relacionados con la educación tales como institutos, universidades, profesores y estudiantes; y otros cambios más del tipo político-legal, tales como la promulgación de nueva reforma educacional. Estos cambios generan impactos importantes en el sistema educativo chileno, puesto que para cualquier nación el sistema educacional es la máquina que empuja el crecimiento económico y no-económico gracias a la formación de profesionales que aporten al país. Por lo tanto, lograr una alta efectividad y eficiencia en el sistema educacional, el cual es medido por los índices de retención/deserción, toma importancia a nivel institucional y nacional.

Con el objetivo de establecer el estado actual del sistema educacional respecto a su eficiencia y efectividad en la formación de profesionales, el Centro de Estudios del Ministerio de educación publicó el 2012 un informe que cuantifica el desgaste del cuerpo estudiantil en el sistema educacional superior chileno. Los resultados mostraron que para el 2010, los Centros de Formación Técnica e Institutos Profesionales alcanzaron una retención del 64,7% y 64,3% respectivamente para el primer año de los programas, mientras que en las universidades fue de 78%. Al agrupar por género, las mujeres presentaron una tasa mayor de retención y aquellos estudiantes que recibieron algún beneficio económico estudiantil, presentaron una tasa del 82,5% de retención, ampliamente mayor al 62,7% presentado por los estudiantes que no recibieron beneficio. Si bien el Centro de Estudios del Ministerio no ha realizado un estudio respecto de la deserción en todos los años, estima que de 100 estudiantes que ingresan a alguna institución educacional superior, solamente 50 de ellos logran terminar los estudios (Centros de Estudios MINEDUC, 2012)

La ineficiencia en la retención de estudiantes se ha mantiene en los últimos años, tal como lo plantea una noticia publicada el 20 de Abril del 2016 en el diario Las Ultimas Noticias (LUN), en la que se discute un ranking de las carreras con mayor deserción en Chile. Según la nota, la carrera con mayor deserción alcanza apenas un 19.2% de retención el primer año, muy distante del 94.7% obtenido por las carreras de Medicina. Adicionalmente enfatizan en que se debe tener importancia con este indicador, puesto que según el investigador Mathias Gómez de la organización Educacion 2020, este se encuentra fuertemente ligado al cumplimiento de estándares educacionales y son un reflejo de la gestión académica de las instituciones. Indica también que una institución de educación superior de calidad debiese desarrollar mecanismos

de retención de sus alumnos, o bien, la detección temprana de una posible renuncia al programa para así poder apoyarlo y evitar la deserción. En general, la eficiencia y efectividad del sistema educacional en formar profesionales está actualmente poco trabajada por las instituciones educacionales y organizaciones gubernamentales competentes, lo que genera efectos negativos importantes para el país, sobre todo si se considera que el gobierno proyecta una fuerte inyección de inversión por la reforma educacional y la intención de utilizar este indicador como parte de las acreditaciones de las instituciones educacionales.

Adicional a la cuantificación del problema de la deserción es importante entender por qué se genera este fenómeno, que desde la perspectiva de las ciencias sociales significa identificar los factores y predictores de la deserción. La identificación de estos elementos puede ser realizado a través de la aplicación de distintas herramientas y metodologías, siendo las más comunes las provenientes de la econometría hace treinta años y los modelos de minería de datos en los últimos siete años. Entre las primeras investigaciones se encuentra el trabajo realizado por Pyke y Sheridan (1993), quienes desde una perspectiva econométrica usaron regresiones logísticas, obteniendo como resultado que los factores financieros y de tiempo de permanencia en el programa influyen positivamente a la retención. Diez años más tarde, Sadler aplicó herramientas econométricas para explicar la deserción el primer año en un programa de enfermería, identificando que aquellas alumnas que manifestaban una mayor relación interna con la profesión (*being nurse* o ser enfermera en español) tenían una menor deserción respecto de aquellas que sentían una relación más externa (*doing nurse profession* o aplicando la profesión de enfermera en español) (Sadler, 2003). Gracias al desarrollo de técnicas y modelos de Minerías de Datos (DM, por su nombre en inglés *Data Mining*), permitió su aplicación en estudios relacionados con la educación, siendo la identificación de predictores de la deserción como una de las más importantes (Peña-Ayala, 2014). Por ejemplo, Denle (2010) a través del uso de técnicas de minería de datos identificaron como predictores aquellos relacionadas con el éxito educacional preuniversitario y universitario, como también los que indican si el estudiante recibió algún tipo de ayuda financiera, ya sea beca o préstamo

Mientras que a finales de los ’90 a nivel internacional el estudio de la deserción se consolidaba teóricamente, en chile los investigadores comenzaban a reaccionar tardíamente, realizando principalmente revisiones bibliográficas de estas investigaciones. Sin embargo, esta corriente cambió a inicios del 2000, donde comenzaron a generarse estudios relacionados con la

cobertura de la educación superior y el efecto de la deserción sobre ella. Por ejemplo, Gonzalez y Uribe (2002) cuantificaron que la cobertura de la educación superior alcanzaba el 25% aproximadamente de los estudiantes que completaron su educación secundaria el año 2000 y proyectaron un aumento de 20 puntos para los siguientes diez años. Adicionalmente midieron económicamente los costos directos que acarreaba la deserción en la educación superior, estimando una pérdida de $47.000 MM de pesos anuales para el país. Si bien, estudios similares de la época generaban alertas respecto a los costos de la deserción, la cantidad de investigaciones posteriores sigue siendo baja.

El desafío de realizar investigaciones que ayuden a la gestión de la deserción toma más importancia actualmente, ya que la educación superior Chilena está en un proceso de cambios estructurales, siendo el más importante el aumento del acceso gratuito a las instituciones educacionales. Por lo tanto, la generación de nuevas herramientas que mejoren la gestión de los recursos gracias a la disminución de la deserción, permitirá que la inversión realizada por el Estado sea reflejado en beneficios para la sociedad, el cual permitirá el desarrollo nacional. Adicionalmente, las instituciones educacionales podrían mejorarán su gestión educacional, cumpliendo estándares de calidad y reflejarla en la obtención de acreditación nacional.

Objetivos de Investigación

En esta tesis se realizará un estudio del fenómeno de la deserción en el contexto universitario chileno. Se analizará el fenómeno específicamente en un programa de la Universidad de Chile, gestionada por la escuela de Sistemas de Información y Auditoría de la Facultad de Economía y Negocios. Los objetivos generales y específicos de esta investigación son:

1. **Objetivo General**: Construir modelos predictivos para la detección temprana de las deserciones en el programa de Ingeniería en Información y Control de Gestión de la Facultad de Economía y Negocios de la Universidad de Chile.
2. **Objetivos Específicos**:
    a. Identificar los predictores que describen la deserción.
    b. Aplicar técnicas de minería de datos para generar modelos que permitan predecir la deserción y la no deserción semestral.

El documento se estructura de la siguiente manera. En la sección siguiente, **Capitulo 1**, se hará una revisión literaria relacionada con la deserción, iniciando con la descripción de las definiciones establecidas durante los años 70 y terminando con la discusión de investigaciones de la disciplina de minería de datos en la deserción estudiantil. Posteriormente, en el **Capítulo 2**, se describirá la metodología de investigación, describiendo en detalle cada uno de las etapas y las técnicas de minería de datos utilizadas en cada una de ellas. Luego, en el **Capítulo 3** se detalla el caso de estudio, describiendo la institución universitaria y el comportamiento de deserción en esta. Posteriormente, en el **Capitulo 4** se detalla el experimento realizado y se presenta el desempeño de los modelos desarrollados, para que en el **Capítulo 5** se analicen los resultados obtenidos de los mejores modelos los cuales son principalmente dos: predictores y perfiles de los estudiantes que desertan y no desertan. Finalmente, en el **Capítulo 6**, la última sección de este documento, se realizan las conclusiones y discusiones.

CAPITULO 1 - MARCO TEÓRICO

En este capítulo se hará una revisión bibliográfica de las investigaciones relacionadas con la deserción. En la primera sección del capítulo se analizarán los modelos teóricos que describen el fenómeno de deserción realizados por autores internacionales y nacionales. Posteriormente se hará una revisión de los estudios relacionados con la disciplina minería de datos respecto a la deserción de los estudiantes.

1.1 DESERCIÓN

1.1.1 Definición de la Deserción

Los trabajos realizados por Spady y por Tinto y Cullen en los años 70 entregan modelos teóricos que describen la deserción, los cuales son utilizados como base teórica por varios investigadores actuales, aun cuando ambos trabajos no son los primeros en describir y definir el fenómeno de la deserción. En el caso del trabajo de Spady, el autor se basó en la teoría del suicidio de Durkheim (1951) y definió la deserción como un resultado a la falta de integración del individuo a su entorno educacional, siendo el primero en aplicar los principios del suicidio en el contexto de la deserción (Spady, 1970ª). Por otro lado, Tinto y Cullen (1975) realizaron una revisión y síntesis de los estudios relacionados con la deserción hasta ese entonces, lo que derivó en una definición ampliamente consensuada en la literatura. De esta manera, desde la visión más simple, los autores entienden este fenómeno como el acto permanente en que el individuo deja la institución en la cual se encontraba registrado, pudiéndose identificar distintos tipos y niveles de deserción.

Posterior a los modelos de deserción generados por Spady, Tinto y Cullen, las investigaciones tomaron como objetivos principales el complementar, explicar y/o validar estos modelos. Los investigadores que formaron parte de esta expansión del conocimiento provenían de distintas disciplinas, por lo que (Braxton, Shaw Sullivan, & Johnson, 1997) se basaron en las categorías implícitamente nombradas por Tinto en sus publicaciones iniciales para clasificar estas investigaciones. Los autores identificaron 5 categorías: (1) Teorías psicológicas, (2) Teorías sociológicas, (3) Teorías económicas, (4) Teorías organizacionales y (5) Teorías de interacción, las cuales junto al análisis la importancia de las disciplinas participantes, permitieron identificar el estado del arte de la investigación sobre la deserción. Como

conclusión de los resultados obtenidos, generaron una lista de 6 recomendaciones para investigaciones futuras con un fuerte llamado a reducir la escasez de estudios empíricos como también la integración de distintas disciplinas para soportar los factores que planteaba la teoría desarrollada durante las décadas de los ’60 al ‘90.

A finales de los años 90, los investigadores internacionales ya habían consolidado la teoría de la deserción y comenzaban a cambiar el foco a estudios empíricos que permitan validar los modelos teóricos. Por otro lado, Himmel en el 2002 publica un artículo que cual aborda la deserción desde la perspectiva conceptual, de modo que las futuras investigaciones puedan considerar las distintas dimensiones desde las cuales puedan aportar al estudio de la deserción a nivel nacional. El artículo permitió en primera instancia definir la deserción, basándose en los distintos estudios realizados a nivel internacional. De este modo, la autora definió la deserción como el abandono prematuro de un programa de estudios antes de alcanzar el título o grado, y se considera un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore (Himmel, 2002). Adicionalmente, declara que es necesario precisar la existencia de dos tipos de deserción: (1) Voluntaria y (2) No Voluntaria. La deserción del tipo voluntaria se puede generar cuando el estudiante renuncia a la carrera o bien, abandona el programa sin informar a la institución educacional. Mientras que la no voluntaria se produce cuando el organismo educacional decide desvincular al estudiante del programa, según el reglamento institucional vigente y causas académicas o disciplinarías.

De acuerdo a los dos tipos y según el nivel en que puede la deserción, es posible graficar el fenómeno según como lo muestra el mapa conceptual en la **Figura 1**.

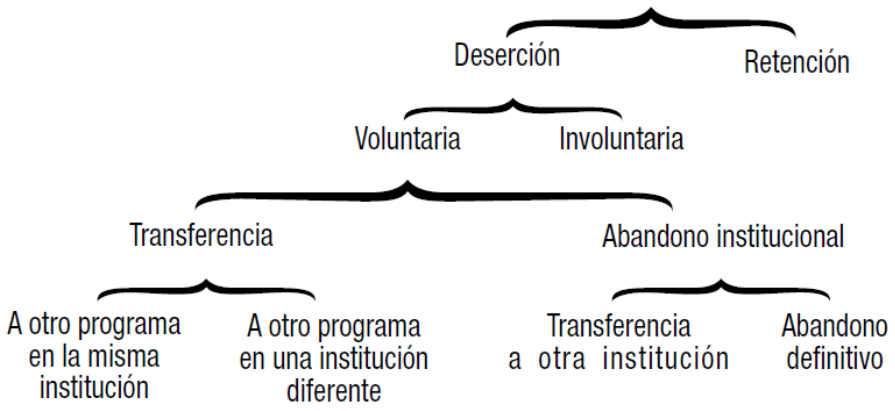

Figura 1. Niveles y tipos de deserción. Fuente: (Himmel, 2002).

Es importante destacar que el artículo de Himmel es usado como base teórica por muchos de los estudios nacional sobre la deserción actualmente (Barrios, 2013; De MagalhaesL-Calvet, 2013; Díaz, 2008; Ministerio de Educación, 2012). Por ejemplo, el artículo publicado por el Centro de Estudios del Ministerio de Educación el 2012 utiliza las definiciones y modelos discutidos por Himmel para estudiar la deserción en el sistema educacional superior para los años 2007, 2008 y 2009.

El trabajo realizado por Himmel permite obtener una base teórica respecto a los distintos tipos, niveles y visiones desde los que se puede estudiar la deserción en Chile. Adicionalmente, entrega una revisión de los modelos más importantes, lo cual permite identificar inicialmente los factores o variables potenciales que explicarían la deserción.

1.1.2 Modelos Teóricos de la Deserción

Con el objetivo de entender la evolución de la teoría respecto de la deserción, se mostrarán de manera cronológica tres modelos teóricos que son usados por la mayoría de los investigadores que estudian la deserción: **(1)** modelo basado en la teoría del suicidio, **(2)** modelo basado en la teoría del intercambio y **(3)** el modelo basado en el modelo de productividad del ambiente laboral.

1970: Spady y su modelo basado en la teoría del suicidio

Uno de los primeros estudios relacionados con la deserción es el desarrollado por Spady. El autor utiliza los principios del suicidio de Durkheim, el cual establece que la decisión de suicidarse no puede ser explicado solamente por factores individuales, puesto que es un hecho social, ya que es generado por la ruptura del individuo con su sistema social debido a su imposibilidad de integrarse a la sociedad (Durkheim, 1951). Siguiendo esta lógica, Spady establece que la deserción sería un resultado de la no integración del individuo con su entorno educacional y alude a que el entorno familiar y sus características afectan fuertemente al estudiante, ya que estos lo exponen a influencias, expectativas y demandas que podrían afectar tanto en su integración social con sus pares en el ambiente universitario como en el rendimiento académico.

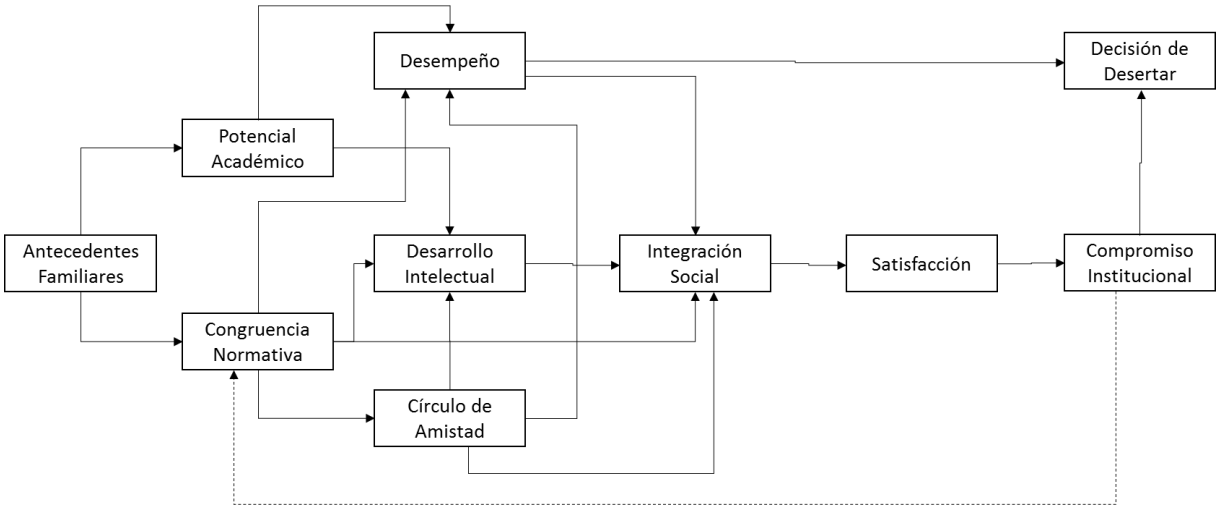

Figura 2. Modelo planteado por Spady. Fuente: (Spady, 1970a)

Tal como se ve en la **Figura 2**, Spady plantea que los **antecedentes familiares** impactan directamente en el potencial académico y en la congruencia normativa del estudiante, el cual se refiere a la compatibilidad de las actitudes, intereses y disposición personal del individuo con las características del medio. Tanto el potencial académico y la conferencia normativa afectan el desempeño académico, el desarrollo intelectual y su integración con los pares en su ambiente educacional. Cada atributo que defina estos factores impacta directamente a la integración social del individuo, que en consecuencia definirá el nivel de satisfacción y consecuentemente el compromiso con la institución educacional. Todos estos factores tendrán incidencia en la decisión final del estudiante para desertar, ya sea voluntaria o involuntariamente.

1975: Tinto y su modelo basado en la teoría del intercambio

En el año 1975 Tinto realiza una revisión de los modelos desarrollados hasta entonces respecto de la deserción. Dentro de esta revisión, resalta el trabajo realizado por Spady y siete años más tarde complementa su modelo incorporando la teoría del intercambio desarrollado por Nye. La teoría del intercambio plantea que los seres humanos evitan aquellas conductas que les generan costos de algún tipo y buscan beneficios en las relaciones, interacciones y estados emocionales que generan con sus pares y la institución educacional (Nye, 1976). Bajo esta perspectiva, para Tinto los estudiantes se mantendrían en el programa que se inscribieron siempre y cuando los beneficios percibidos superen el esfuerzo, dedicación y otros costos

personales; y si existe alguna otra actividad que le genere mayores beneficios la decisión final del estudiante podría desencadenar en una deserción (Tinto, 1982).

Según el modelo mostrado en la **Figura 3**, se desprende que cuando un estudiante ingresa a un programa de educación superior, este plantea inicialmente sus compromisos con la institución y con sus objetivos personales de obtener un grado en la institución. Tales compromisos serán afectados por sus antecedentes familiares, como por ejemplo nivel sociocultural; por sus atributos personales, como edad y género; y por su experiencia académica preuniversitaria. Luego de un tiempo razonable estando en el programa, el estudiante reevaluará sus compromisos iniciales de acuerdo a su integración social y su desempeño académico en la institución, cuyos efectos podrían desencadenar la deserción si el estudiante percibe que los costos sean mayores que los beneficios.

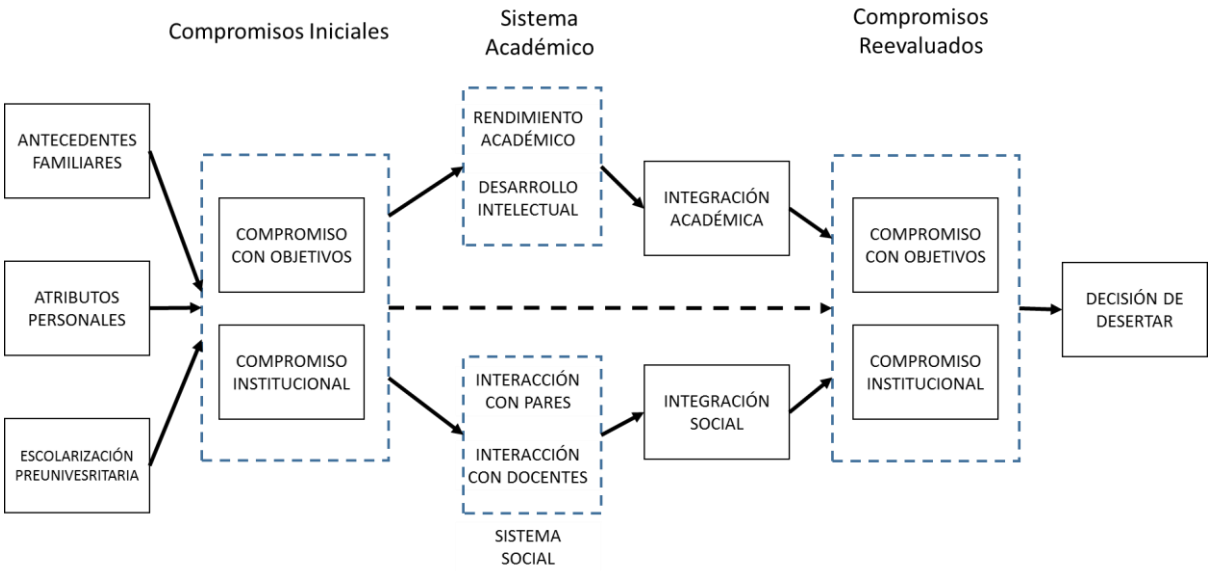

Figura 3 Modelo planteado por Tinto. Fuente:(Tinto & Cullen, 1975).

1985: Bean y su modelo basado en la productividad del ambiente laboral

En los siguientes años, Bean complementa los modelos desarrollados por Tinto y Spady a través de dos trabajos publicados en el 1980 y 1985. En el primero trabajo Bean toma los factores planteados por Tinto y Spady y define variables para testear si estos modelos tienen evidencia empírica. Cinco años más tarde, en el segundo trabajo amplía su estudio incluyendo

en el análisis a estudiantes no tradicionales como respuesta a un cambio en el acceso a la educación superior de la época que hasta entonces era restringida solamente a un grupo de elite. Este cambio generó un aumento en la heterogeneidad del cuerpo estudiantil en la educación superior, generando un nuevo desafío de evidencia empírica de los modelos de Tinto y Spady (Bean & Metzner, 1985; Bean, 1980).

En artículo publicado en 1980, Bean planteó un conjunto de variables que podrían definir la deserción. Adicionalmente, con el objetivo de identificar las causalidades entre estas variables, define que las variables relacionadas con el antecedente del estudiante, tales como estado socioeconómico, desempeño académico previo y residencia actual (supuestamente diferente previo ingreso al instituto educacional), impactarían en los determinantes organizacionales o bien, de la institución educacional y estos a su vez, en la decisión de desertar (Bean, 1980). El conjunto de variables y sus relaciones se encuentran en la **Tabla 1** y **Figura 4**.

Variable	Definición
Variables de Antecedentes	
Desempeño Previo	Grado en que el estudiante ha demostrado sus logros académicos previos.
Estatus Socioeconómico	Grado en que los padres del estudiante han logrado como estatus a través de la ocupación familiar.
Residente en el Estado	Si el estudiante es un residente del estado en donde la institución educacional está
Distancia a Casa	Distancia de su residencia actual a la casa de sus padres.
Tamaño de la ciudad	Tamaño de la comunidad donde el estudiante pasó la mayor parte de su tiempo en su crecimiento.
Determinantes Organizacionales (Basado en (Price, 1977)*)*	
Rutina	Grado en que el rol de ser un estudiante es visto como una rutina.
Desarrollo	Grado en que un estudiante cree que él/ella se está desarrollando como resultado de ir a una institución de educación superior (IES).
Valor Práctico	Grado en que un estudiante percibe que su educación será utilizada para emplearse.

Tabla 1. Lista de variables planteadas por Bean. Fuente: (Bean, 1980).

Variable	Definición
Calidad Institucional	Grado en que la IES es percibida como una proveedora de buena educación.
Integración	Grado en que el estudiante participa en relaciones primarias o cuasiprimarias (tiene amigos cercanos)
Promedio de Notas Universitario	Grado en que un estudiante demuestra su capacidad para desempeñarse en una IES.
Compromiso de Metas	Grado en que obtener un grado universitario es percibido como importante.
Comunicación (Requerimientos/Reglas)	Grado en que la información sobre ser un estudiante es vista o entregada.
Justicia Distributiva	Grado en que un estudiante cree que es tratado justamente por la institución. Por ejemplo: recibe premios y castigos proporcionalmente a su esfuerzo realizado en su rol como estudiante.
Centralización	Grado en que un estudiante cree que participa en los procesos de tomas de decisión. Por ejemplo: centros de estudiantes, consejeros, etcétera.
Advisor	Grado en que un estudiante cree que su *advisor* es útil.
Relación con Funcionarios	Nivel de contactos informales con los miembros de la facultad.
Trabajo en el Campus	Necesidad de tener un trabajo en el campus universitario para permanecer en la escuela
Major (área)	El área de uno de los campos de estudio
Major(certeza)	Grado en que un estudiante es poco indeciso en que se está especializando
Alojamiento	Cuando una persona vive *On Campus*
Organización del Campus	El número de miembros en la organización del campus
Oportunismo (Transferencia/Trabajo/Hogar)	Grado en que un rol alternativo (como estudiante, empleado o dependiente en casa de los padres) existe en el ambiente externo (otra universidad, en una empresa o volver a casa de los padres).
Variables de Intervención	
Satisfacción	Grado en que siendo un estudiante es visto positivamente
Compromiso Institucional	Grado de lealtad hacia la pertenencia del estudiante en la organización

Tabla 1 (Continuación). Lista de variables planteadas por Bean. Fuente: (Bean, 1980).

Se destacan del listado las variables de calidad de la institución y la satisfacción del estudiante con la institución. Estos se pueden relacionar directamente con la satisfacción y su compromiso institucional, mientras que de manera transitiva con la decisión de desertar. Guiado por la literatura, Bean relacionó las variables y realizó regresiones lineales para el caso de los estudiantes mujeres y hombres. La relación entre estas variables se presenta a continuación en la **Figura 4.**

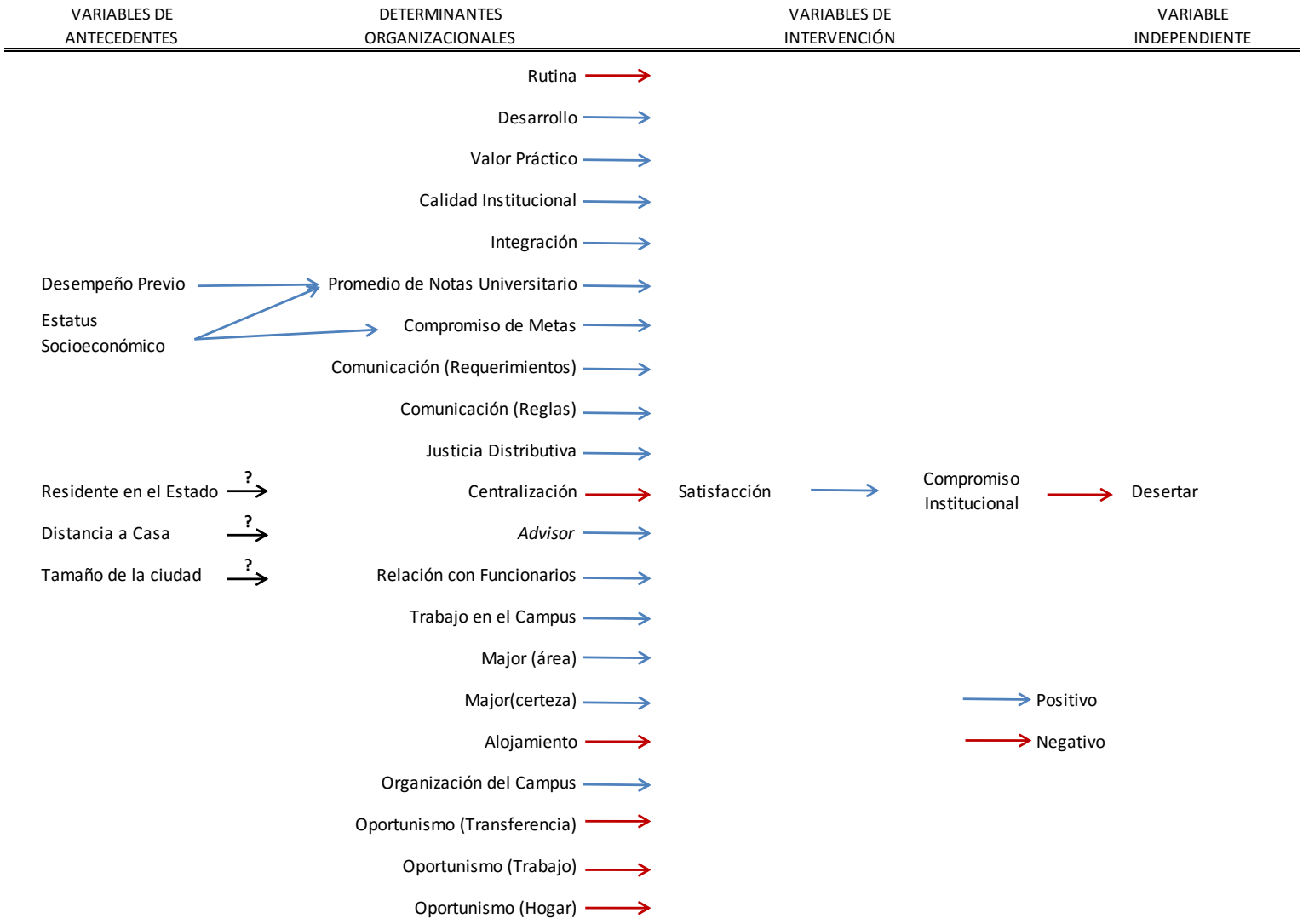

Figura 4. Relación entre las variables planteadas por Bean. Fuente: (Bean, 1980).

Tal como se planteó en los párrafos anteriores, las variables de desempeño académico previo y estatus socioeconómico impactan en la relación del estudiante con su entorno educacional. De esta manera, aquellos estudiantes que tienen antecedentes de excelencia académica en el colegio, tenderían a obtener mejores desempeños en la universidad, lo que aumentaría el grado de satisfacción y compromiso institucional. En consecuencia, el estudiante tendería a no desertar.

Para el caso de las variables relacionadas con el lugar de donde proviene, el autor no logró obtener resultados significativos estadísticamente, pero sí fueron interesantes los encontrados respecto a los valores positivos y negativos que reflejaban una relación directa o inversamente proporcional con los factores de su entorno social educacional. A modo de ejemplo, la distancia hacia la ciudad natal del estudiante impacta positivamente en el grado que este percibía la calidad institucional y la organización que existía en el campus; probablemente, mientras más lejos de su ciudad, menor conexión con esta tenía, por lo que tendía a considerar y valorar mejor la nueva institución en que la que se encontraba inmerso.

Cinco años después Bean se asocia con Metzner y amplía su estudio de la deserción a estudiantes no tradicionales (Bean & Metzner, 1985). En este nuevo trabajo plantea un modelo en el que se destaca la utilización de variables sociodemográficas como el género, edad y etnia del estudiante. Las variables planteadas y las relaciones causales entre estas son mostradas en la **Figura 5**. El estudio planteó la heterogeneidad que existe en el cuerpo estudiantil, lo que se traduce en que para el fenómeno de deserción es importante considerar características propias del estudiante. Por lo tanto, para el estudio que se realizará en esta tesis deberán ser consideradas variables tales como el género, edad, satisfacción institucional, estatus socioeconómico, desempeño académico escolar y universitario para su análisis en el impacto de la deserción.

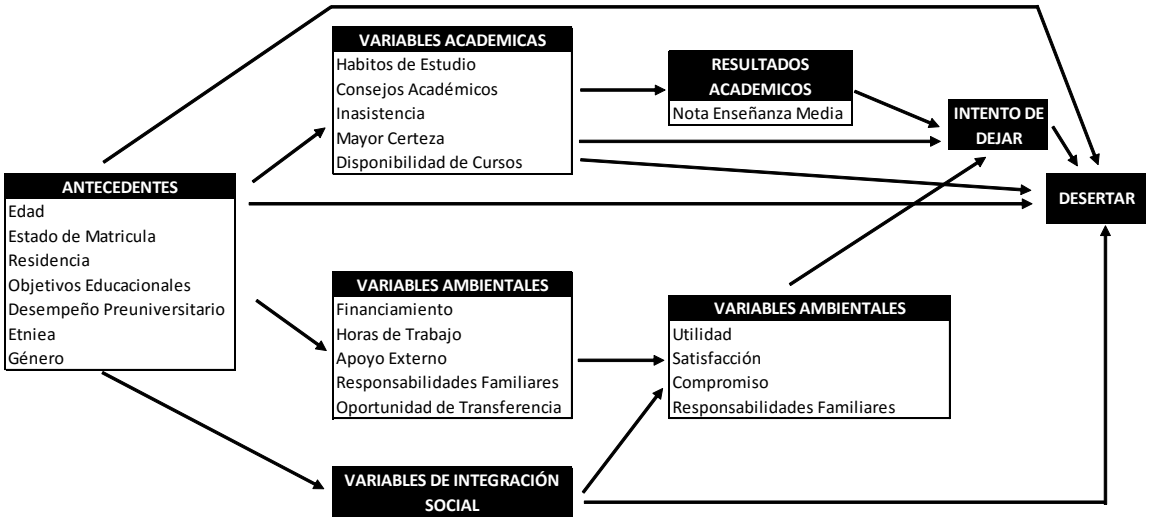

Figura 5. Relación entre las variables planteadas por Bean en su segundo estudio. Fuente:(Bean & Metzner, 1985).

1.2 MINERÍA DE DATOS Y LA DESERCIÓN

Parte del estudio de la deserción estudiantil es detectar los factores que expliquen la deserción. Estos factores pueden ser utilizados como predictores para detectar tempranamente si un alumno finalmente dejará sus estudios. Este objetivo de estudio ha sido cubierto por investigaciones econométricas, cuyos investigadores detectan con cierta significancia estadística los factores que podrían explicar la deserción. Sin embargo, el objetivo de los estudios econométricos es identificar variables explicativas de la deserción, abriendo una puerta para que otras disciplinas, tales como la minería de datos, aporten al estudio la generación de modelos predictivos.

En las siguientes dos secciones se discutirá los aportes realizados por la econometría respecto a la significancia estadística de factores y sobre el uso de estos en la aplicación de técnicas de minería de datos.

1.2.1 Identificación de Factores

Los investigadores econométricos se han enfocado en demostrar de manera empírica los factores que autores durante los setenta y ochenta plantearon como explicativos del fenómeno de la deserción (Spady, 1970b; Tinto, 1982; Tinto & Cullen, 1975). Por ejemplo, Pyke & Sheridan (1993) utilizaron regresiones logísticas en una base datos de 601 estudiantes de programas postgrados, en donde la información académica, demográfica y financiera fueron utilizadas como potenciales variables explicativas de la deserción. Los resultados mostraron que para ambos tipos de estudiantes, cuanto mayor es el tiempo en el programa y el financiamiento otorgado, mayores son las probabilidades de que el estudiante se gradúe.

En otro estudio relacionado a programas de pregrado de enfermería, la autora comparó dos herramientas evaluativas que se utilizaban para el ingreso al programa: *Grade Point Averages* (GPA – Promedio de notas colegio) y Ensayo. Este último se basaba en que el estudiante plasmaba en un ensayo los argumentos de su interés de ingresar al programa, además de explicar la profesión y sus trabajos relacionados con la salud. Cada ensayo era evaluado por una comisión en cinco dimensiones enfocadas fuertemente en criterios de escritura: organización, enfoque, desarrollo de ideas, calidad de inglés, congruencia de ideas con los valores y congruencia de ideas con las normas y comportamiento de la profesión de enfermería. A través del uso de estadística descriptiva y *test-t* de diferencia de medias, la autora encontró diferencia significativa en los ensayos entre estudiantes que desertaban y completaban el programa, siendo los puntajes más altos en estos últimos. Adicionalmente, de manera cualitativa, encontró diferencias respecto de la relación que tenían los postulantes con la profesión. Aquellos estudiantes que completaban el programa, tendían a escribir su relación con la enfermería de manera más interna (*being nurse*) que externa (*doing nurse profession*) (Sadler, 2003). Respecto del desempeño académico anterior evaluado a través de sus notas (*GPA*), no encontró diferencia estadísticamente significativa, a diferencia de otros estudios previamente publicados que sí lo hicieron como el de (Byrd, Garza, & Nieswiadomy, 1999).

En conclusión, podemos afirmar que, tal como fue planteado teóricamente en los años 70 y 80, los factores relacionados con el actual entorno del estudiante, el desempeño académico previo y su compromiso con sus objetivos tienen directa relación con el fenómeno de deserción.

1.2.2 Identificación de Predictores

Gracias a que las capacidades de procesamiento de los computadores aumentara y el costo de almacenamiento disminuyera, la implementación y desarrollo de técnicas de minería de datos a través de la programación de algoritmos experimentó un crecimiento. Investigadores de esta área de conocimientos identificaron el desafío de utilizar estas herramientas en estudios relacionados con la educación, creando una nueva disciplina denominada *Educational Data Mining* (EDM) o Minería de Datos Educacional. Esta disciplina es considerada como emergente y se preocupa del desarrollo de técnicas, métodos y modelos para el tratamiento único de datos que provienen de la gestión educacional. Una de las ramas del EDM es el desarrollo de modelos que apoyen a la gestión de la deserción, los cuales utilizan como entrada los factores de la deserción identificados como estadísticamente significativos por la econometría y genera modelos predictivos del fenómeno (Peña-Ayala, 2014)

Al analizar los estudios de la deserción relacionados con EDM, se destacan tres problemas a los que se han enfrentado los investigadores. Una de ellas es la generalización de los predictores, ya que los factores identificados como más importantes por la econometría no son aplicables en todo contexto educacional. El segundo problema es la definición de la temporalidad de la variable dependiente, que en minería de datos serían las que reflejan la deserción y la no deserción del estudiante. Finalmente, un tercer problema es el relacionado con el desbalance en las bases de datos, ya que por la naturaleza del fenómeno de la deserción, la mayoría de las bases tiene más eventos registrados para la no deserción que la deserción.

La generalización de predictores

En Yu, DiGangi, Jannasch-Pennell, & Kaprolet (2010) aplican tres técnicas de minería de datos para estudiar la deserción aludiendo a que si bien muchos investigadores

habían utilizados técnicas paramétricas, como análisis de regresiones lineales y logísticas, esta nueva perspectiva era nueva y permitía detectar relaciones no lineales y no convencionales entre los factores y la variable dependiente que refleja la deserción. En su estudio, las variables que captaban las horas de traslado a la universidad, la residencia (dentro o fuera del estado) y la etnia eran consideradas como cruciales para predecir la salida temprana de un estudiante del programa. Estos factores diferían respecto a los encontrados en investigaciones econométricas, en donde se indicaba que el desempeño académico en el colegio era el predictor más importante para la deserción. De esta manera, los autores concluían que los factores identificados por la econometría no eran generalizables y llamaban a la aplicación de técnicas de minería datos en otras instituciones para identificar sus propios factores y predictores.

Siguiendo la misma línea de investigación, Delen (2010) identificó que los factores considerados como mejores predictores para la universidad de su investigación eran aquellas relacionadas con el éxito educacional preunivesritario y universitario, como también si el estudiante recibió algún tipo de ayuda financiera (beca o préstamo). A diferencia de Yu, DiGangi, Jannasch-Pennell, & Kaprolet (2010), Denle sí obtuvo resultados que no contradecían lo identificado hasta entonces por la econometría.

Si bien los estudios econométricos permiten seleccionar inicialmente un conjunto de factores de la deserción, para luego identificar los predictores, las técnicas de minería de datos ayudan a evaluar si tales factores son aplicables en diferentes contextos educacionales, ya que como vimos anteriormente, para dos estudios en distintas instituciones, se obtuvieron distintos predictores.

Temporalidad y Desbalance

Dada la naturaleza de ocurrencia de la deserción, existen dos características que diferencia a los estudios de EDM con la deserción: temporalidad y balance de las bases de datos.

La ocurrencia de la deserción no es siempre la misma para todos los programas de estudio. Esto se debe a la duración en semestres y las particularidades de cada programa no es la misma entre distintas instituciones educacionales. De hecho, es posible identificar variaciones a nivel de ciudad, región, país y áreas de formación. Esto se traduce en que el la ocurrencia de la deserción no siempre se concentra en el mismo semestre para todos los centros educacionales, lo que obligaba a los investigadores identificar cuando los alumnos abandonan el programa de estudio. Por ejemplo, Alkhasawneh y Hargraves (2014) formularon un modelo híbrido con el objetivo de predecir la retención para el primer año del programa; mientras que Yu et al (2010) estudiaron también la deserción pero para el segundo y tercer año. El efecto de la temporalidad también genera una diferencia entre naciones, puesto que la extensión de los programas varía incluso entre países, siendo particularmente más largo en Latinoamérica que en los países europeos o norteamericanos.

En relación con las características de la base de datos que se utilizan en las investigaciones donde se aplica minería de datos, los investigadores han identificado un desbalance en los registros que caracterizan a los alumnos que desertan y los que no, siendo en general mucho menos el número de estudiantes que deserta. Esto genera un problema en los modelos de minería, puesto que los algoritmos reciben como entrada menos volumen de datos para aprender de los registros identificados como deserta, entregando predicciones con baja precisión para la deserción y alta precisión para la retención. En definitiva, es importante la evaluación y aplicación de técnicas de balanceo en cada estudio relacionado con la deserción (Delen, 2010; Thammasiri, Delen, Meesad, & Kasap, 2014).

Desde los años setenta los estudios a nivel internacional relacionados con deserción son variados, de los cuales el investigador chileno Diaz realizó una revisión y análisis (Díaz, 2008). Sin embargo, la revisión entregó como resultado que el volumen de investigaciones nacionales en donde se aplican técnicas de minería de datos es relativamente bajo. De la misma manera lo plantea Himmel, donde reclama por una escasez de investigaciones sobre la deserción a nivel nacional cuando realizó el análisis y síntesis de los modelos teóricos (Himmel, 2002). Desde ese entonces, el fenómeno ha

tomado mayor interés por distintas instituciones e investigadores, buscando como objetivo inicial el entendimiento de la deserción y posteriormente la aplicación de distintas técnicas que apoyen la generación de medidas para mitigar los costos de este fenómeno (Díaz, 2008; Morales, Fuentes, Riquielme, & Kraemer, 2011; Morales, Riquelme, Bascuñan, & Navarrete, 2014).

En resumen, los factores identificados por la econometría no son generalizables para todas las instituciones en donde ocurre la deserción, por lo que se incentiva la realización de más estudios con el uso de minería de datos con el objetivo de identificar aquellos predictores específicos para la institución. Adicionalmente, dado que los programas varían en su duración, genera que es el investigador deba definir la temporalidad de este fenómeno. Esto también responde a que las decisiones de retirarse del programa tomadas en los primeros semestres no sean comparables con las tomadas en los últimos semestre, debido a que a medida que avanza el tiempo, más factores entran en juego en la decisión del alumno, tales como la relación con sus pares y el desempeño académico. Finalmente, en la mayoría de las investigaciones en donde se aplican técnicas de minería de datos se debe considerar el efecto en el desempeño de los modelos que puede generar el desbalance de las bases de datos.

CAPÍTULO 2 – METODOLOGÍA

El **Capítulo 2** describe la metodología que se utilizará en este estudio. Inicialmente se hará una introducción de la metodología planteada por Fayyad, Piatetsky-Shapiro y Smyth nombrada como *Knowledge Discovery in Databases* (**KDD**), cuyas etapas claramente especificadas permiten la aplicación apropiada de técnicas de minería de datos en cualquier proyecto o investigación de esta disciplina. Posteriormente se describirá en detalle cada etapa y se mostrará la relación con el proyecto actual.

Los autores Fayyad, Piatetsky-Shapiro y Smyth en el año 1996 propusieron una metodología que planteaba un conjunto de actividades agrupadas y ordenadas con el objetivo de aplicar correctamente las técnicas de minería de datos y de esta manera, aumentar la probabilidad de éxito de los proyectos de minería de datos, como también el desempeño y confiabilidad de los modelos generados. Los autores indicaron que el primer paso es el trato correcto de los datos y que posteriormente estos deben ser procesados para permitir la detección y extracción de patrones (o bien conocimientos). Para esta última función, la disciplina Minería de Datos (que en ese entonces era principalmente desarrollada desde el área de la ciencia de la computación) parecía la con mayor potencial y más indicada para cumplirla (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Los autores plantearon 5 etapas que todo proyecto, investigación u organización relacionados con la minería de datos deberían aplicar. El nombre que otorgaron a la metodología fue *Knowledge Discovery in Databases* (KDD)*,* que en español significa Descubrimiento de Conocimiento en base de datos. Fayyad y sus co-investigadores indicaron que el problema básico abordado por la metodología KDD es uno de mapeo de datos de bajo nivel para identificar conocimientos y patrones en ellos. Tal metodología permite la aplicación en distintas áreas tales como marketing (Linoff & Berry, 2011; Shaw, Subramaniam, Tan, & Welge, 2001), finanzas (Diaz, Theodoulidis, & Sampaio, 2011), detección de fraudes (Brause, Langsdorf, & Hepp, 1999), fabricación (Harding, Shahbaz, & Kusiak, 2006), telecomunicaciones (Hung, Yen, & Wang, 2006; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012), medicina (Lavrač, 1999), recursos humanos

(Chien & Chen, 2008) y ahora último, educación (Miranda & Vásquez, 2015; Romero & Ventura, 2010; Vásquez, Ortega, Lee, & Silva, 2015; Wilson, Eva, & Lobb, 2013)

Según la **Figura 6**, el KDD consta principalmente de cinco etapas en la que cada una tiene entradas, actividades y resultados definidos. En la etapa inicial, la base de datos es recuperada para que, bajo ciertos criterios, se seleccionen los datos (registros y variables) que se utilizarán en la obtención del conocimiento. El resultado de esta etapa genera un conjunto de datos con los cuales trabajar. Posteriormente la base de datos es preprocesada y limpiada de todo tipo de ruido que pueda generar algún problema en el procesamiento.

Una vez que se obtiene un conjunto de registros y atributos totalmente limpios, estos son transformados de acuerdo a los requerimientos que la técnica o modelo de minería de datos tenga para ser aplicada. Al cumplir tales requisitos, se obtiene una base apta para la aplicación de técnicas de minería de datos. Finalmente, el producto obtenido, que es descrito como *patrones* por los autores, es interpretado y evaluado, el cual se transforma en el conocimiento buscado al inicio de la metodología. Cabe destacar que durante el desarrollo de la metodología existe la posibilidad de volver a alguna etapa si la persona o equipo ejecutor así lo considera necesario.

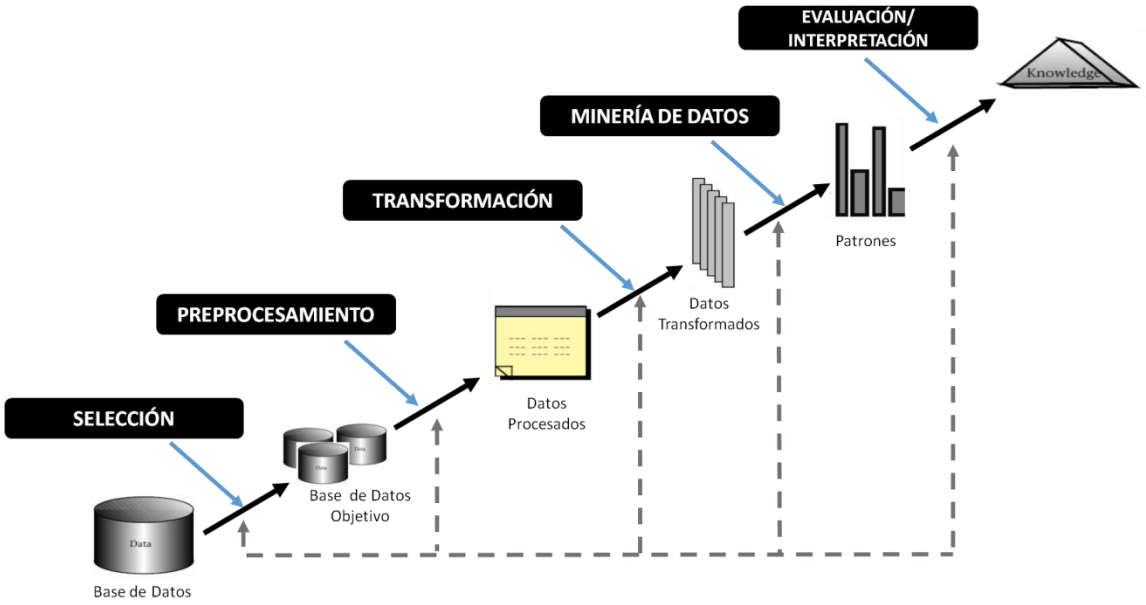

Figura 6. Etapas del KDD y sus respectivas entradas/salidas. Fuente:(Fayyad et al., 1996).

2.1 Selección

En esta etapa se seleccionan las variables y registros con los que finalmente se trabajará. De acuerdo a la literatura, existen técnicas estadísticas y matemáticas que se recomiendan para realizar esta selección con el objetivo de identificar las variables que cumplan los siguientes requisitos: potencialmente explicativas del fenómeno en estudio, cuentan con nulo o poco error de registro, estarán disponibles en el futuro si se vuelve a realizar el análisis y son medibles antes que ocurra el evento en estudio.

En este estudio se realizará inicialmente una exploración de los datos, de esta manera se podrá detectar inicialmente las posibles variables que sean predictores de la deserción Adicionalmente se obtendrá información respecto a las escalas y tipos de datos, detección de valores extremos, sesgos en los datos y qué tan dispersos están.

2.2 Preprocesamiento

El conjunto de variables y registros seleccionados deben estar libres de ruidos para no generar sesgo durante su procesamiento en la etapa de Minería de Datos. Por lo tanto, se hace necesario un preprocesamiento de la base de datos para cumplir con un nivel mínimo de calidad de los datos. Si bien no existe un estándar único de calidad de datos, en este estudio consideraremos los siguientes criterios:

- **Completitud:** No existencia de datos faltantes (*missing values*) en los atributos.
- **Consistencia:** El formato y codificación en un atributo en particular deben ser idénticos para todos los registros.
- **Coherencia:** Los datos deben responder a reglas lógicas básicas según el contexto de la base. En este punto se deben evaluar los datos o registros identificados como *outlier*. En otras palabras, aquellas observaciones que distan mucho de las otras.
- **Validez:** Los registros deben ser coherentes con la organización y/o actividad que se documenta.

Con el objetivo de que los datos cumplan con los 4 criterios, se evaluará los siguientes tratamientos a los datos vacíos:

a. **Reemplazo Missing Values por Moda y Media**: Dependiendo del tipo de dato, los valores vacíos pueden ser reemplazados por la moda (variables categóricas) o media (variables numéricas).

b. **Reemplazar Missing Values por Modelo Predictivo:** Los datos vacíos o *outliers* pueden ser reemplazado a través del uso de modelos predictivos tales como máquinas de aprendizajes que permitan la predicción del valor más posible para el dato, considerando los otros datos que sí tienen valor.

c. **Reducir Registros**: En el caso que la cantidad de registros con valores vacíos es muy pequeña en comparación con la base, dichos registros podrían ser no

considerados para la base de entrenamiento. De esta manera se asegura que los datos usados para el entrenamiento y validación reflejan el comportamiento real del fenómeno en estudio.

Para esta tesis decide utilizará la reducción de la base de datos, seleccionando los registros en donde ningún atributo es nulo o *missing value*.

2.3 Transformación

En la etapa *Minería de Datos* se aplican muchos modelos cuyos requisitos es que todas las variables sean numéricas. Sin embargo, en la realidad existen datos que por naturaleza no cumplen esta condición, tales como la variable de texto categórica que almacena la ciudad de residencia del estudiante.

Se realiza el siguiente tratamiento a cada variable:

1. **Variables Binominales**: Para este tipo de datos los valores posibles son solamente dos. De esta manera, la solución de transformación es asignar un valor numérico a cada categoría n de las variables evaluadas.

2. **Variables Polinominales:** En este caso los valores posibles son generalmente más de dos categorías. La forma en que se deben transformar a números es generando $n - 1$ atributos binomiales, donde n es la cantidad de categorías.

Además de que los datos deben ser numéricos, algunos algoritmos de minería requieren que estos estén normalizados para obtener resultados eficientes. Al mismo tiempo, la normalización soluciona el problema de las diferencias generadas en los rangos y medidas que ocupan las variables, como por ejemplo la edad y el ingreso bruto familiar.

Las bases de datos utilizadas en este estudio contienen datos del tipo binomial y polinominal, por lo que las transformaciones anteriormente explicadas serán aplicadas

según corresponda. Así mismo, estos se normalizarán para que los algoritmos sean implementados eficientemente.

2.4 Minería de Datos

En esta fase se realiza el modelamiento propiamente tal a través de la aplicación distintas técnicas. El objetivo es extraer patrones y conocimientos previamente desconocidos, los cuales se obtendrían a través del procesamiento de los datos. Las técnicas aplicadas pueden ser de caracterización, asociación, *clusterización*, clasificación y regresión (Han, Kamber, & Pei, 2011).

Las tareas relacionadas con la clasificación hacen referencia a la construcción de modelos donde se categorizan los registros en grupos predefinidos o clases ya conocidas. En general, la variable dependiente en los modelos de categorización es del tipo categórico o discreto. Por otro lado, cuando se habla de actividades de regresión, se hace referencia al uso de modelos para predecir variables del tipo real. La decisión de aplicar tareas de categorización o regresión dependen crucialmente del objetivo que se busca.

En la actualidad existen variadas técnicas de minería de datos que permiten la ejecución de las tareas anteriormente descritas. Respecto de la categorización y regresión es posible identificar máquinas de aprendizaje que gracias a los avances tecnológicos, la implementación de estas se hace cada vez más accesible para los usuarios. Máquinas de aprendizaje tales como *Support Vector Machine, Decision Tree y Artifical Neural Net* han sido programadas y utilizadas como librerías en *softwares* de minería de datos. Para esta tesis se ha decidido aplicar las siguientes máquinas: **(1)** *Support Vector Machine* (SVM), **(2)** *Decision Tree* (DT), **(3)** *Artifical Neural Net* (ANN), **(4)** *Logistic Regression* (LR).

2.4.1 Máquinas de Aprendizaje

Support Vector Machine

El Support Vector Machine (SVM) fue introducido por (Vapnik & Chervonenkis, 1964). Hoy en día, desde la disciplina de máquinas de aprendizaje el *SVM* es un modelo del tipo de aprendizaje supervisado el cual utiliza algoritmos de clasificación y análisis de regresión. En términos prácticos, se puede explicar el funcionamiento del SVM como la clasificación de un conjunto de registros a través de la separación con un hiperplano, el cual minimiza el costo de error de clasificación de cada registro a una de las dos clases en estudio.

Formalmente se define el Support Vector Machine como la función de separación representada por un hiperplano en el espacio $\mathbb{R}^n$, el cual maximiza el margen de separación. En una separación de puntos por dos clases, el problema se denomina del tipo linealmente separable. Bajo este concepto, en un espacio $\mathbb{R}^n$ con finitos puntos que representan observaciones, estos pueden ser separados por infinitos hiperplanos. Sin embargo, los algoritmos que aplican Support Vector Machine identifican la separación lineal óptima, el cual está definido como la máxima capacidad de generalización y el mínimo error empírico de clasificación. El margen de separación es definido como la distancia entre el par de paralelos canónicos, por lo que el margen es igual a dos veces el mínimo entre los puntos de entrenamiento y el hiperplano de separación. Estos puntos los cuales se minimiza la distancia de separación del hiperplano, son llamados vectores de soporte (*support vectors*). Sobre estos se obtienen las reglas de clasificación (Vercellis, 2009).

Matemáticamente, si se define $\boldsymbol{w}$ como el vector de coeficientes del hiperplano y $\boldsymbol{b}$ como el intercepto, se define el hiperplano dado como:

$$\mathbf{w}'\mathbf{x} = b$$

Mientras que los dos hiperplanos paralelos canónicos son:

$$\mathbf{w}'\mathbf{x} - b - 1 = 0. \qquad \mathbf{w}'\mathbf{x} - b + 1 = 0$$

Entonces el margen de separación δ es definido como:

$$\delta = \frac{2}{\|w\|}$$

Donde $\|w\| = \sum_{j \in N} w_j^2$. En orden de determinar los coeficientes $\mathbf{w}$ y el intercepto b, el hiperplano que optimiza la separación se define por la solución del siguiente problema cuadrático con restricciones lineales:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$s.a. \quad y_i(\mathbf{w'x_i} - b) \geq 1, \quad i \in \mathcal{M}$$

Sin embargo, en la mayoría de los casos los m puntos no son linealmente separables. Entonces la ecuación anterior se formula de la siguiente manera:

$$\min_{w,b,e} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} e_i$$
$$s.a. \quad y_i(\mathbf{w'x_i} - b) \geq 1 - e_i, \quad i \in \mathcal{M}$$
$$e_i \geq 0 \qquad i \in \mathcal{M}$$

Donde C refleja el costo de error de clasificación.

A modo de ejemplo, en la **Figura 7** se muestra la aplicación de SVM en donde divide un mapa de registros en dos categorías. Aquellos puntos ubicados al lado derecho del hiperplano son cateogrizados como no fuga (cuadrado azul) y aquellos ubicados al lado izquierdo como fuga (círculo rojo). El hiperplano generado estaría minimiza el costo de error de clasificación y adicionalmente, maximizaría la distancia entre los dos grupos dentro de un margen establecido.

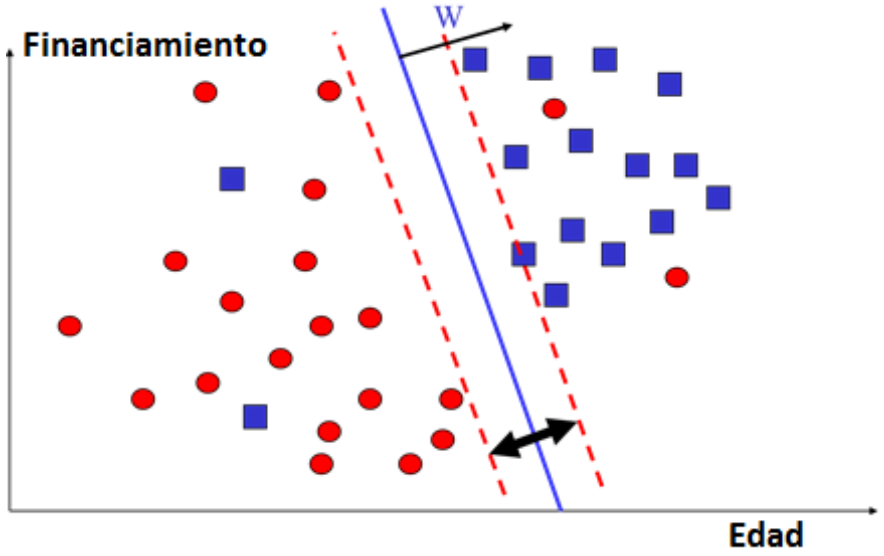

Figura 7. Representación gráfica de la aplicación del algoritmo SVM.

Parámetros

Los parámetros del Support Vector Machine son los costos asociados al error de clasificación. Por lo tanto, en esta tesis se optimizará el **C** que otorgue el mejor desempeño de los modelos.

Árboles de Decisiones

Los árboles de decisión, o *Decision Tree* (DT) en inglés, en la disciplina de minería de datos es considerado un modelo de clasificación planteado por (Quinlan, 1986), el cual se basa en las teorías de decisiones para realizar clasificaciones a los base de datos en donde se apliquen algoritmos minería de datos. Quinlan ha realizado grandes aportes a los algoritmos de árboles de decisiones, siendo los más reconocidos el C4.5 e ID3.

Los árboles de decisión segmentan los registros utilizando técnicas matemáticas y estadísticas, introduciendo el concepto de entropía (índice de incertidumbre o desorden), el cual sirve para identificar el siguiente atributo de segmentación.

Gráficamente los árboles se componen de nodos, ramas y hojas. Los nodos son puntos de unión, en donde se refleja una toma de decisión. Las ramas representan los arcos de

conexión entre nodos, y las hojas son nodos terminales en donde se refleja la decisión final, es decir, la clasificación.

En el libro Data Science for Business(Provost & Fawcett, 2013) se muestran dos claros ejemplos de la segmentación y representación gráfica de un árbol de decisión. Respecto de la **Figura 8**, se puede ver la secuencia de nodos que reflejan la decisión de segmentación de un conjunto de datos. Los nodos finales reflejan la clasificación final con la probabilidad correspondiente.

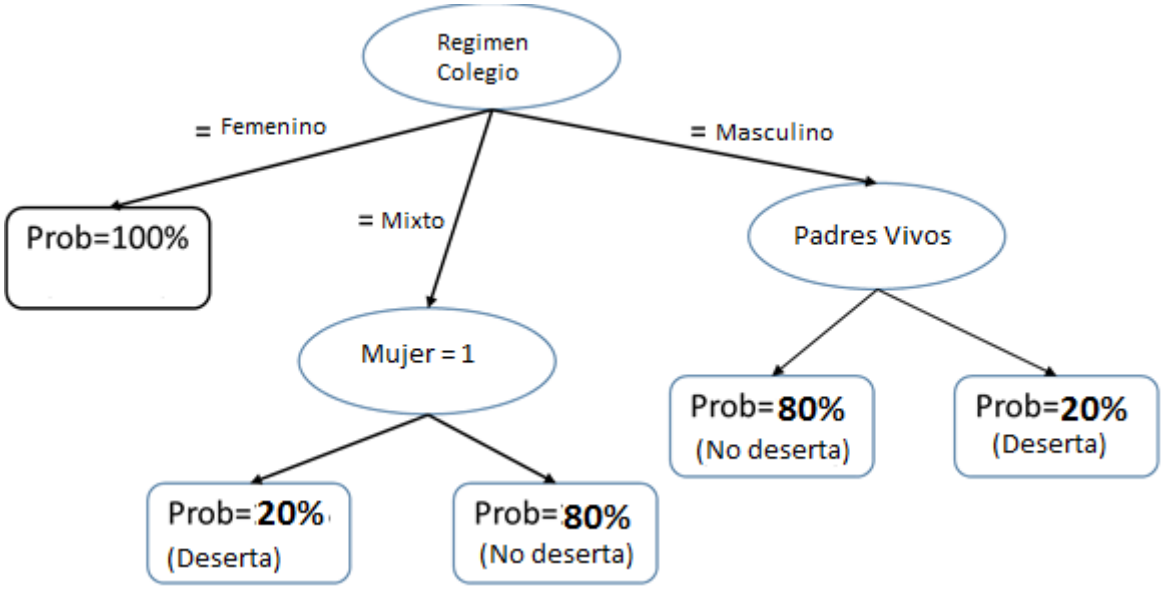

Figura 8. Representación gráfica del resultado obtenido por un algoritmo basado en la máquina de aprendizaje Árbol de Decisiones.

En la **Figura 9** se muestra de manera gráfica la segmentación realizada a través del árbol de decisión. Cada división refleja un nodo y los puntos dentro de un cuadro reflejan las hojas o nodos finales del árbol.

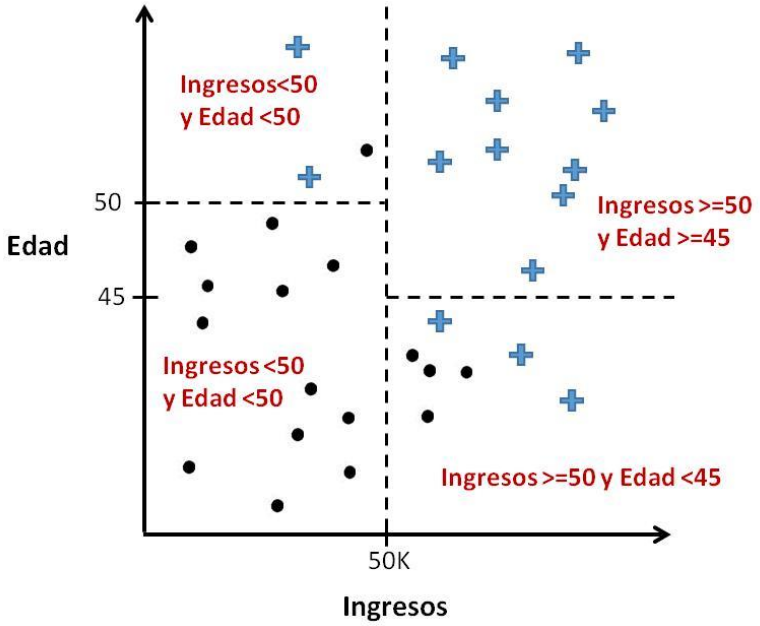

Figura 9. Representación gráfica de la segmentación según algoritmo de Árbol de Decisión. Fuente: (Provost & Fawcett, 2013).

La segmentación previamente descrita, requiere de algunos pasos específicos antes de implementar el algoritmo de clasificación. Los elementos necesarios a describir son:

Reglas de Separación

Para cada nodo del árbol, se debe especificar el criterio por el cual se separarán los conjuntos de datos. Estas reglas varían por el número de nodos descendientes, el número de atributos y la evaluación de métricas. Dentro de estas últimas se encuentran dos mayormente usados: (1) Índice de Clasificación Errónea, (2) Entropía y (2) Gini.

El índice de Clasificación Errónea para un nodo q es calculado como:

$$Miscl(q) = 1 - \max_h p_h$$

Donde p_h es la proporción de observaciones en el nodo clasificados a la clase h. Este índice mide la proporción de ejemplos mal clasificados cuando todas las instancias del nodo q son asignados a la mayoría de la que ellos pertenecen, de acuerdo al principio de voto por mayoría (Vercellis, 2009).

Por otro lado, el índice de entropía para un nodo q es calculado como:

$$Entropy(q) = -\sum_{h=1}^{H} p_h log_2 p_h$$

El índice en sí mismo es un índice de diversidad, ya que mide las diferencias de distribución de los grupos de clasificación.

Finalmente, el índice Gini para un nodo q es calculado como:

$$Gini(q) = 1 - \sum_{h=1}^{H} {p_h}^2$$

El índice Gini mide la uniformidad en la distribución de la clasificación de las instancias y sirve para distinguir la diferencia entre dos grupos categorizados de manera dicotómica.

Criterios de Detención y Poda

En cada nodo del árbol se deben definir criterios de detención para el algoritmo. Se entiende como criterios de detención aquellos que indican si la construcción de una rama del árbol debería continuar recursivamente o bien, el nodo se debe considerar como hoja. Adicionalmente, con el objetivo de evitar el excesivo crecimiento de un árbol durante el desarrollo del algoritmo recursivo se generan criterios de pre-poda, como también reducir el número de nodos después de que el árbol haya sido generado (poda).

En la literatura existen muchos criterios utilizados para la detención del algoritmo y poda, siendo los más comunes el tamaño del nodo, la pureza y el mejoramiento del rendimiento del árbol (Vercellis, 2009).

1. **Tamaño del Nodo:** Este consistente en el número de observaciones bajo el nodo. El algoritmo recursivo terminará si el tamaño llega a ser menor de un umbral.

2. **Pureza:** Se entiende como pureza la proporción de observaciones en un nodo que pertenecen a la misma clase. Mientras más alta es esta proporción, mayor pureza tiene el nodo. El algoritmo recursivo se detendrá una vez conseguido un mínimo nivel de pureza.

3. **Mejoramiento**: Un algoritmo continuará la segmentación recursiva en una rama, si la división genera un mejoramiento en el desempeño del modelo. Para esta tesis la evaluación será a través de la precisión.

Parámetros

Los parámetros para un árbol varían según el algoritmo utilizado. En el caso de esta tesis los parámetros utilizados para optimizar el rendimiento de los árboles serán: (1) Máximo Profundidad (Tamaño Poda), (2) Máximo tamaño separación, (3) Aplicación Poda y (4) Aplicación Prepoda.

Redes Neuronales Artificiales

Las redes neuronales artificiales o *artificial neural net* (ANN) fueron introducidas inicialmente como concepto de red neuronal por los neurólogos (McCulloch & Pitts, 1943). Quince años más tarde, (Rosenblatt, 1958) generó el primer perceptrón simple basado en los conceptos de red neuronal, proponiendo así los fundamentos de una red neuronal artificial.

Una red neuronal se compone de redes de nodos llamados neuronas. Cada nodo recibe un conjunto de entradas provenientes de otros nodos y entregan una salida. Esta salida se compone por tres funciones:

1. **Función de propagación**: Es la función que se compone por la sumatoria de las entradas multiplicados por un peso de interconexión.

2. **Función de activación**: Es la función que modifica la función anterior (aprendizaje). Puede que la configuración de la red no tenga esta función, por lo que la salida es la misma función de propagación**.**

3. **Función de transferencia:** Es la función que se aplica al valor que entrega la función de activación. Su principal es para acotar el rango de salida del nodo. Las más comunes son la función sigmodea (intervalos entre 0 y 1) y la tangente hiperbólica (intervalos entre -1 y 1).

Perceptrón

Gráficamente un perceptrón puede ser reflejado como la siguiente figura:

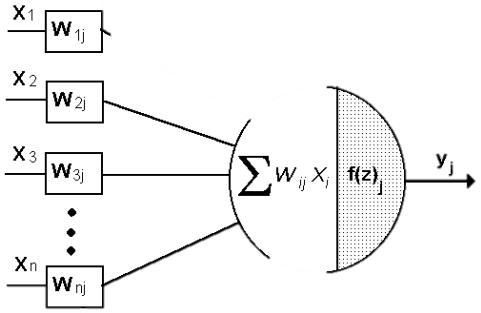

Figura 10. Descripción gráfica de un perceptrón.

La **Figura 10** es la forma más sencilla de una red neuronal y corresponde a una sola neurona de salida. Las entradas $x_1, x_2, \dots, x_n$ son valores de entradas que se combinan con distintos pesos $w_1, w_2, \dots w_n$ y entregan un resultado de salida $f(\mathbf{x})$.

Supongamos que los valores de los pesos ya han sido determinados durante el proceso de entrenamiento. Bajo este contexto, se puede determinar la predicción de una nueva observación bajo los siguientes pasos. Primero, la combinación linear de los pesos con los valores de entrada (variables explicativas) para una nueva observación pueden ser calculadas como:

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n - \mathcal{E} = \mathbf{w'x} - \boldsymbol{\mathcal{E}}$$

Entonces, la predicción de $f(\boldsymbol{x})$puede ser obtenida como:

$$f(\boldsymbol{x}) = g(w_1x_1 + w_2x_2 + \cdots + w_nx_n - \mathcal{E}) = g(\mathbf{w'x} - \boldsymbol{\mathcal{E}})$$

Donde $g(\cdot)$ es la función de activación y su propósito es mapear la combinación lineal al conjunto de posibles valores de la variable dependiente. En un problema de clasificación binaria estos valores son definidos como $[-1,1]$, entonces, una función de activación $g(\cdot)$ para este caso sería la función signo ($sgn(\cdot)$). Una vez identificada la función de activación, se implementa un algoritmo iterativo que determina los valores de los pesos w_i, examinando en secuencia, uno a uno las observaciones del vector **x.**

Redes Multinivel de Prealimentación

Una estructura más compleja de las redes neuronales son las del tipo multinivel de prealimentación. Se compone de tres elementos principales:

1. **Capa de Entrada:** Conjunto de neuronas que representan las variables/factores independientes.

2. **Capa Oculta:** Conjunto de neuronas que reciben la información de las neuronas de entradas, estudia sus patrones y determina el peso de cada una de forma iterativa.

3. **Capa de Salida**: Conjunto de neuronas que proporcionan la respuesta de la red neuronal. En un problema de clasificación, el número de neuronas en esta capa es igual al número de clases de clasificación.

A modo de ejemplo, la **Figura 10** muestra la estructura básica de una red neuronal.

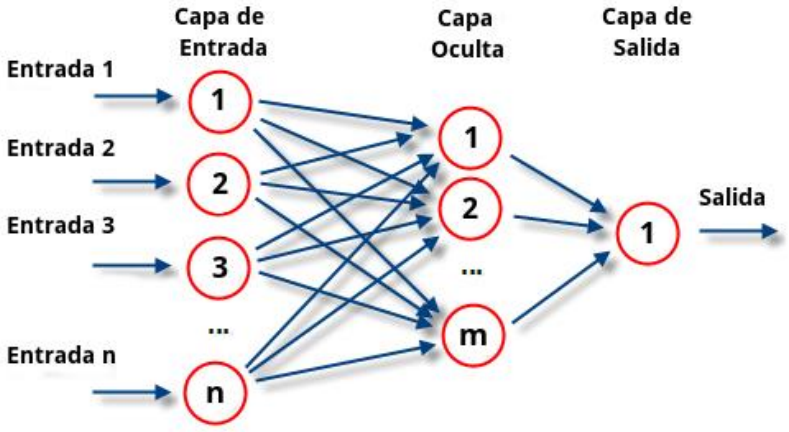

Figura 11. Ejemplo de estructura de una red neuronal artificial.

Cada nodo opera básicamente como un perceptrón. En otras palabras, los pesos estarán asociados a cada arco conectado entre nodos, y cada nodo es asociado a un coeficiente de distorsión ($\mathcal{E}$) y una función de activación ($g(\cdot)$). El método utilizado para determinar los pesos y los coeficientes de distorsión es denominado algoritmo de *backpropagation*, o en retropropagación en español.

Un algoritmo de retropropagación inicia con valores de pesos w_i aleatorios. Posteriormente, cada instancia de la base de entrenamiento es examinada secuencialmente, generando una predicción para cada una y obteniendo desempeños de correctas y malas clasificaciones. Estos resultados son utilizados como retroalimentación para ajustar los pesos y realizar nuevamente la examinación con pesos y coeficiente de distorsión ajustados.

Las redes neuronales tienen un alto desempeño predictivo, puesto que son capaces de capturar las relaciones complejas lineales y no lineales entre las variables

independientes y dependientes. Sin embargo, necesita de grandes volúmenes de información para obtener un buen desempeño.

Parámetros

Los algoritmos de redes neuronales varían en la configuración de los parámetros para su correcto funcionamiento. En el caso de esta tesis, se optimizarán dos parámetros principalmente: (1) Tamaño Capa Oculta y (2) Ciclos de Entrenamientos.

El Tamaño de la Capa Oculta indica la cantidad de nodos utilizados en esta capa de la red neuronal. Actualmente existen algoritmos que calculan de manera automática el tamaño ideal de la capa, siendo la mayoría de las veces $n-1$ nodos, donde n es la cantidad de variables dependientes. Sin embargo, no siempre este número entrega los mejores desempeños de las redes, por lo que en esta tesis la asignación automática será comparada con asignaciones manuales del número de nodos.

Los ciclos de entrenamientos hacen referencia al número de iteraciones que el algoritmo realizará para ajustar los valores de los pesos w_i y coeficiente de distorsión. Mientras mayor número de ciclos asignados, mayor probabilidad de obtener un mejor desempeño de las redes.

Regresión Logística

La Regresión Logística o *Logistic Regression* (LR) en inglés, es un caso especial de regresiones cuyo uso es para predecir el resultado de una variable dependiente categórica. Tiene bastante uso en los cálculos de probabilidades, donde se predice la ocurrencia de un evento en función de otros factores.

A modo de ejemplo, supongamos que la variable de respuesta y toma valores 0 y 1. De acurdo a lo que postula la regresión logística, la probabilidad posterior $P(y|\mathbf{x})$ de respuesta a la variable condicionada del vector $\mathbf{x}$ sigue una función logística definida como:

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

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

En las ecuaciones previamente mostradas, el algoritmo identifica los coeficientes **w** de forma iterativa, usualmente a través del método de máxima verosimilitud.

En general los modelos y algoritmos de regresión logística presentan la misma dificultad que las regresiones lineales, en otras palabras, pueden adolecer de problemas de multicolinealidad y sesgo.

Parámetros

En la etapa de entrenamiento, los algoritmos de regresiones logísticas obtienen las predicciones para un conjunto de observaciones $\mathbf{X}$ y al igual que los *support vector machines*, es posible calcular el costo de error de las clasificaciones erróneas. Adicionalmente, los algoritmos iteran con el objetivo de identificar el mejor conjunto de los valores $\boldsymbol{w_i}$, en donde el costo total de clasificaciones erróneas busca ser minimizado. En esta tesis, el parámetro de las regresiones logísticas será el costo **C** para cada error de clasificación**.**

2.4.2 Clusterización

Todos los modelos anteriormente explicados hacen referencia a la mayoría de los aprendizajes del modelo supervisado, esto quiere decir que la variable dependiente es conocida. Existen también los no supervisados, en los que la variable final no es conocida previamente. Un ejemplo de este es la *clusterización*, cuyo objetivo es agrupar los registros cercanos pero sin saber de ante mano cuáles son los segmentos propiamente tal y qué características tienen estos.

Técnicamente, un algoritmo de *clusterización* identifica segmenta un conjunto de observaciones $\boldsymbol{X}$ en $\boldsymbol{n}$ grupos, donde cada uno de estos se caracteriza por una observación tipo llamada centroide.

La segmentación e identificación de los centroides se puede obtener través de la siguiente solución del problema de programación lineal:

$$\begin{aligned} & \min_{x} \sum_{i,j\ \epsilon\ I} y_{ij} d_{ij} \\ s.a. \quad & \sum\nolimits_{j \epsilon I} y_{ij} = 1, \quad i\ \in I \\ & \sum\nolimits_{i \in I} x_i = N \\ & y_{ij} \leq x_i \quad i\ \in I \\ & x_i, y_{ij}\ \in [0,1] \end{aligned}$$

Donde y_{ij} representa la asignación de una observación a un centroide, x_i la asignación de una observación como centroide y d_{ij} refleja la distancia entre la $i-ésima$ y $j-ésima$ observación. En el problema de programación lineal se puede observar que dado un número N de segmentaciones a generar, se deben asignar todas las observaciones a un centroide y se buscará minimizar la distancia de estas asignaciones. La distancia d puede ser calculada a través de distintos índices, siendo el más común la distancia euclidiana, la cual se mide como:

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

En este estudio se aplicarán tanto las técnicas de aprendizaje supervisado y no supervisado con el objetivo de identificar las diferencias que se generan en el desempeño de la predicción y seleccionar la mejor combinación de aplicación de técnicas.

2.4.3 Clasificadores

Las cuatro técnicas de minería de datos explicados en la sección anterior pueden generar distintos desempeños si previamente se aplica un *umbral* de clasificación según un indicador que indica la probabilidad de que un registro pertenezca a una clase en particular. Tal indicador es conocido como *confidence*.

A modo de ejemplo, imaginemos que se cuenta con una base de 500 registros y es aplicada una máquina de aprendizaje, del cual se obtendrá un *confidence* de la clase positiva para cada registro. De manera predefinida, al ordenar de mayor a menor se podrá observar que todos los registros con igual o mayor *confidence* a 0.5 (o 50%) son clasificados como positivos. Eventualmente este estándar puede ser más o menos restrictivo, aumentando el *umbral de clasificador* en el caso que se requiera mayor restricción para catalogar un registro a la clase positiva y por tanto menor restricción para la clase negativa, o de lo contrario, disminuyendo el umbral si se requiere ser menos restrictivo para la clase positiva y más restrictivo para la clase negativa.

La selección del mejor umbral para cada modelo puede ser de manera manual según la experticia del investigador, o bien, utilizar el análisis conocido como *Receiver Operating Characetristic* (**ROC**), o simplemente curva ROC. Esta curva es un análisis gráfico que muestra los puntos en un cuadro cartesiano, cuyos ejes se componen por *Ratio del Verdadero Positivo* (Verdaderos Positivos o TP divididos por el total de la clase real positiva) y *Ratio Falos Positivo* (Falsos Positivos o FP dividido por la clase real negativa) para un sistema de clasificación binario mientras su umbral va variando. El ROC también se puede representar por el trazado que representa la fracción de los TP respecto del total de los reales positivos frente (TP/P) a la fracción de los FP de los reales negativos (FP/N).

El ROC se obtiene graficando puntos en un gráfico donde FP/N y TP/P son los ejes X e Y respectivamente, lo que representa relativos compensaciones entre los verdaderos positivos (beneficios) y falsos positivos (costos). Cada resultado de predicción o una

instancia de una matriz de confusión representa un punto en el ROC como el mejor método de predicción posible. La mejor posible predicción produciría un punto en la esquina superior izquierda de coordenadas (0,1) del espacio del gráfico ROC, puesto que representa el 100% de TP/P y 0% de PF/N (o desde otra perspectiva como 100% de TN/N y 0% de FN/P). Este punto también se denomina como clasificación perfecta. Una suposición completamente al azar daría un punto a lo largo de una línea diagonal desde la parte inferior izquierda de las esquinas superiores derecha.

La diagonal divide el gráfico ROC y los puntos por encima de esta diagonal representan buenos resultados de la clasificación, mientras que los puntos por debajo de la diagonal representan los malos resultados. Por lo tanto, el mejor umbral es aquel que los puntos de mala clasificación los convierte en buena clasificación, siendo el óptimo aquél que invierte la mayor cantidad de puntos.

2.4.4 Desbalance

El problema del desbalance hace referencia a los trabajos realizados con base de datos en donde la proporción entre las distintas clases que hay en estudio no están balanceadas, lo que genera efectos negativos a los resultados y desempeño de modelo. El principal problema es que el modelo tiende a *aprender más de la clase mayoritaria que de la minoritaria,* puesto que los algoritmos reciben mayor información de la primera clase (Provost & Fawcett, 2013). En nuestro caso el modelo generado tendería a aprender más de los registros que reflejan que un alumno no deserte y menos de los registros que reflejan a los desertores. Para solucionar este problema, en la literatura se han planteado diversas técnicas basadas principalmente en el muestreo (Thammasiri et al., 2014). En esta tesis se utilizarán y compararán principalmente 2 técnicas:

1. ***Random Under-Sampling*:** Este es un método de muestreo que selecciona aleatoriamente registros de clase mayoritaria para removerlos de la base de datos final hasta que la cantidad se equipare con la clase minoritaria.

2. ***Random Over-Samping*:** Este método también es de muestreo, pero a diferencia de la anterior, los registros seleccionados aleatoriamente son de la clase minoritaría y no se remueven de la base, sino que se agregan al nuevo conjunto de base de datos. Este proceso se repite hasta que la cantidad equipare la clase mayoritaria.

Es importante destacar que en ambas técnicas se puede generar sesgo, sin embargo, para disminuir este efecto, tales técnicas serán aplicadas solamente a la base de entrenamiento, por lo que la de testeo mantendrá su distribución original.

2.5 Interpretación y Evaluación

En esta etapa se evalúa el desempeño de los modelos aplicados en la etapa anterior. Adicionalmente, se visualiza e interpreta los *patrones* que el(los) mejor(es) model(os) entregan. En esta etapa el juicio experto juega un rol fundamental, ya que el investigador deberá evaluar si los patrones extraídos tienen sentido en el contexto que fueron aplicados. Cabe destacar que en esta etapa, al igual que las anteriores, existe la posibilidad que se decida volver al primero paso o a una etapa previa según corresponda.

Los desempeños de los modelos pueden variar por muchos factores, tales como las variables escogidas y el manejo que se hizo a los datos. Por lo tanto se hace imperante utilizar mecanismos para evaluar el desempeño de cada uno de ellos. De esta manera el patrón identificado tendrá un sustento más objetivo. En este sentido, la literatura ha planteado distintas métricas para medir el desempeño predictivo de los modelos. Los más comunes son el *Error de clasificación* y la *Precisión de Predicción* (*Accuracy*). Sin embargo estas métricas miden el desempeño general de los modelos, asumiendo que todos los tipos de errores tienen el mismo costo, lo que no siempre es así en un contexto organizacional. Un claro ejemplo es el fenómeno estudiado en este documento, donde predecir que un estudiante NO DESERTA cuando finalmente DESERTA es mucho más costoso que predecir que DESERTA cuando NO DESERTA (Conceptos Error Tipo I y Error Tipo II). De esta manera, la evaluación de los desempeños de cada modelo tendrá considerado la importancia de cada una de las clases.

En conclusión, los desempeños de los modelos serán evaluados principalmente en dos criterios: Exactitud (*accuracy)* y Costo de Clasificación. Para medir ambas métricas se hará uso de la matriz de confusión, herramienta bastante utilizada en las aplicaciones de minería de datos de clasificación, dada su facilidad de uso y la calidad de información que entrega.

Validación Cruzada

La validación cruzada se utiliza para medir el desempeño de los modelos de minería de datos aplicados en cualquier proyecto de esta disciplina. El objetivo principal es utilizar la misma base de datos para generar el modelo de predicción y posteriormente evaluarlo y así obtener una proyección del desempeño de predicción.

La forma de validar el modelo se realiza a través de la división de la base en dos, siendo una de ellas la de entrenamiento y la otra de testeo. En la mayoría de los casos, aproximadamente el 70% del total de los registros es utilizados para el entrenamiento, mientras que el 30% restante es para la medición del desempeño. La selección se hace principalmente de manera aleatoria.

Los pasos de la validación cruzada es la siguiente:

1. Se aplica una técnica de minería de datos, principalmente máquina de aprendizaje, a los registros que forman parte de la base de entrenamiento.
2. Una vez obtenido el modelo, este es aplicado a los registros de la base de testeo y es comparado la predicción con la clase real.
3. Este desempeño es obtenido y se utiliza como input para ajustar el modelo según corresponda.
4. Se realiza nuevamente del paso uno al paso tres hasta completar x validaciones tomando en cuenta que los registros formen parte al menos una vez de la base de testeo.

En **Figura 11** muestra de manera gráfica los pasos anteriormente descritos.

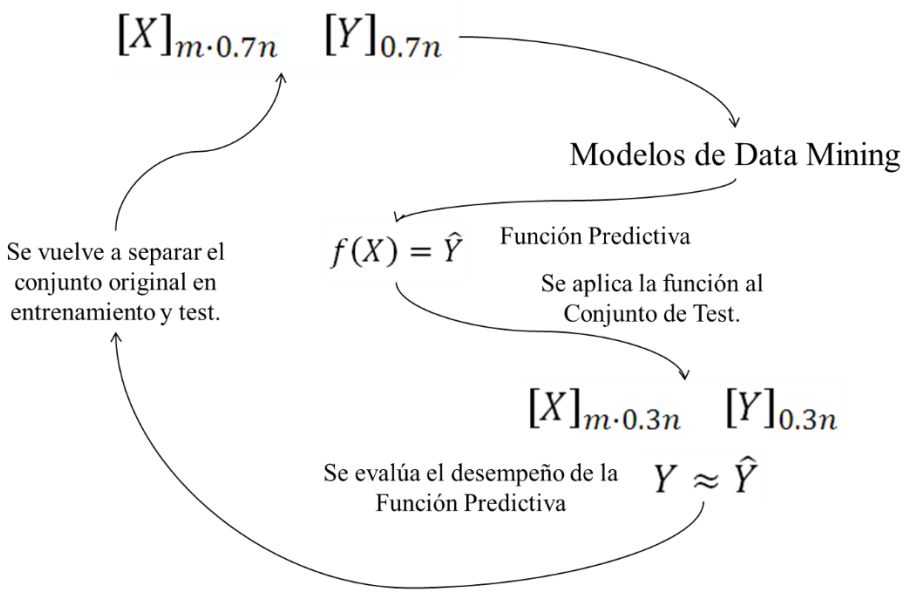

Figura 12. Representación gráfica de la validación cruzada.

La validación cruzada será utilizada en esta tesis para implementar los algoritmos de técnicas de minería de datos.

Matriz de Confusión

La matriz de confusión es una tabla compuesta mayoritariamente por dos filas y dos columnas, las cuales contienen información sobre el desempeño de las clasificaciones predichas por un modelo de clasificación. Usualmente, las filas representan las instancias que el modelo predijo, mientras que las columnas las instancias observadas reales.

En el caso de una clasificación dicotómica, es decir, clasificación en dos clases, se generan dos tipos de clases nombradas positivas y negativas. Para cada observación se realizan predicciones de ambas clases, a través de la implementación de un algoritmo de técnica de minería de datos, y estas se comparan con el valor real de la clase. Aquellas observaciones en que se predijo como clase positiva y efectivamente era de esa clase, son denominadas como *True Positives* **(TP),** o Verdadero Positivo en español;

mientras que si no lo eran, se evalúan como *False Positive* **(FP)**, o Falso Positivo. Ocurre lo mismo para las clases negativas, asignando como *True Negative* **(TN)**, o Verdadero Negativo, las predichas como negativas y efectivamente lo eran; mientras que *False Negative* **(FN)**, o Falso Negativo, las con predicción de clase negativa pero efectivamente positivas (Shmueli, Patel, & Bruce, 2011).

		Clase Verdadera	
		+	**-**
Predicción Clases	**+**	True Positive	False Positive
	-	False Negative	True Negative

Tabla 2. Ejemplo de matriz de confusión.

De la matriz presentada en la **Tabla 2** se pueden obtener indicadores de desempeño, siendo el más común el índice que mide la exactitud de predicción. Adicionalmente, según la importancia de cada una de las clases, se pueden generar otros indicadores como el Ratio Verdadero Positivo (*True Positive Rate* en su versión internacional en inglés).

Para esta tesis, ambos indicadores serán usados para medir el desempeño de los modelos desarrollados. Las fórmulas para sus cálculos se muestran a continuación:

$$Precisión = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Ratio\ Verdadero\ Positivo\ (TPR) = \frac{TP}{TP + FN}$$

En resumen, para medir el desempeño de los modelos predictivos generados se utilizarán dos indicadores de acuerdo a las ecuaciones anteriormente establecidas.

CAPITULO 3 – CASO DE ESTUDIO

3.1 Descripción del Caso Real

El experimento fue realizado en las escuelas de pregrado de la Facultad de Economía y Negocios de la Universidad de Chile (FEN). Actualmente, FEN imparte cuatro programas de estudios, los cuales son gestionados por dos escuelas. La carrera de Ingeniería Comercial mención Administración e Ingeniería Comercial mención Economía son gestionadas por la Escuela de Economía y Administración, mientras que Contador Auditor e Ingeniería en Información y Control de gestión lo son por la Escuela de Sistemas de Información y Auditoría. En esta tesis se trabajará con el programa de Ingeniería en Información y Control de Gestión, cuya creación en el 1989 fue motivada por las demandas de profesionales en áreas de sistemas de información y control de gestión.

El programa de IICG contempla 10 semestres y cada uno contempla entre cinco a siete cursos. Adicionalmente, los estudiantes deben realizar 2 prácticas profesionales, las cuales se realizan después de los semestres ocho y diez. La práctica está contemplada para ser realizada en los meses estivales, en el que comúnmente el sistema educacional chileno concentra las vacaciones para los estudiantes. Una vez cumplido los 8 semestres del diseño curricular del programa, más otros requisitos según decreto universitario, los estudiantes obtienen el grado de licenciado. Luego de aprobar los cursos del noveno y décimo semestre, los estudiantes deben rendir un examen de grado para optar al grado de título universitario.

Para el ingreso de estudiantes existen 5 tipos cupos: **(1)** Ingreso vía Prueba Selección Universitaria (PSU), **(2)** Extranjeros, **(3)** Deportistas Destacados, **(4)** Cupos Supernumerarios Beca Excelencia Académica (BEA), y **(5)** Cupos Sistema de Ingreso Prioritario de Equidad Educativa (SIPEE). De acuerdo a la información pública de la FEN, los cupos para cada tipo de ingreso en el 2016 se distribuyen de la siguiente manera:

Tipo de Ingreso	Vacantes
PSU	100
Extranjeros	5
Deporitstas Destacados	3
Cupos Supernumerarios BEA	5
Cupos SIPEE	5

Tabla 3 Vacantes por tipo de ingreso. Fuente <http://admisionfen.cl>, consultada el 12 de Mayo de 2016.

Tal como lo muestra la **Tabla 3**, la mayor cantidad de ingresos es a través de la PSU. Esto es importante tomarlo en cuenta, puesto que las características de los estudiantes que ingresan por PSU y los que ingresan por otra vía, no son comparables por los requisitos y los requisitos especiales que deben cumplir para ser aceptados. Adicionalmente, la cantidad de información disponible de los alumnos que ingresaron vía PSU es mayor debido al número de cupos existente para este tipo de ingresos. Por lo tanto, en esta tesis se trabajará con los estudiantes que ingresaron vía PSU.

3.2 Análisis de la deserción

Para determinar que estudiantes desertaron en cada grupo de año de ingreso, se utilizó el estado académico registrado por la FEN a inicios del semestre de otoño del año 2016. El sistema identifica siete tipos de estados académicos:

1. **Regular**: El estudiante está matriculado y activo académicamente
2. **Postergado**: El estudiante está matriculado pero congelado académicamente
3. **Egresado**: El estudiante aprobó todos los ramos del programa y cumple las condiciones para rendir el examen de grado
4. **Titulado**: El estudiante aprobó todos los ramos del programa y los exámenes de grado
5. **Transferido**: El estudiante renunció al programa y se transfirió a otro programa de la misma facultad
6. **Renuncia**: El estudiante renunció al programa como también a la FEN, y finalmente
7. **Eliminado**: El estudiante ha sido eliminado del programa por motivos disciplinarios o académicos.

Los primeros cuatros estados indican que el estudiante no ha desertado del programa, mientras que los últimos tres indican que sí lo ha hecho, siendo la transferencia y renuncia como deserciones voluntarias y eliminado como deserción no voluntaria.

Reclasificando los siete estados a Deserta y No deserta, la distribución de la deserción por tipo de ingreso se refleja en la **Tabla 4**.

Año Ingreso	No Deserta	% del Total del Año	Deserta No Voluntario	% del Total del Año	Deserta Voluntario	% del Total del Año	Total general
2007	68	58.1%	14	12.0%	35	29.9%	117
2008	74	67.3%	10	9.1%	26	23.6%	110
2009	77	69.4%	7	6.3%	27	24.3%	111
2010	73	58.9%	18	14.5%	33	26.6%	124
2011	62	53.0%	13	11.1%	42	35.9%	117
2012	73	58.9%	16	12.9%	35	28.2%	124
2013	75	75.0%	6	6.0%	19	19.0%	100
2014	93	74.4%	5	4.0%	27	21.6%	125
Total general	**595**	**64.1%**	**89**	**9.6%**	**244**	**26.3%**	**928**

Tabla 4. Número de tipo de deserciones por año de ingreso.

Según la **Tabla 4**, la deserción alcanza un promedio de 35,6% por año. Para los años 2013 y 2014 existe una disminución, tanto en cantidad como proporción de estudiantes que desertan. La mayoría de las deserciones son del tipo Voluntario, siendo anualmente en promedio un 74.1% de las deserciones por año de ingreso. Si bien en el caso de los años 2013 y 2014 las deserciones son un 20% del total de alumnos, las voluntarias son aproximadamente un 80% de todas las deserciones, al igual que para todos los grupos de estudiantes con mismo año de ingreso.

La explicación de la disminución de la deserción en los años 2013 y 2014, puede ser explicada porque en el año 2013 el programa IICG experimentó cambios del diseño curricular, los que incluyeron la implementación de una serie de talleres en los primeros semestres con el objetivo de contextualizar al estudiante sobre la carrera en la que está matriculado y fortalecer lazos con su grupo de generación. Adicionalmente, se separó el ingreso de estudiantes por programas, siendo desde el 2013 el ingreso separado para los programas Contador Auditor e Ingeniería en Información y Control de Gestión.

Las deserciones del tipo no voluntarias son ocasionadas por el bajo rendimiento de los alumnos, relacionado por factores psicológicos, médicos, familiares y otros del tipo extra-académicos. Por lo tanto, su detección se hace gestionable por la escuela, puesto que ya tienen implementado sistema de alertas, según cantidad de cursos reprobados, y protocolos que indican tanto al alumno como a las autoridades cuando un alumno está en amenaza de ser eliminado del programa. Sin embargo, no ocurre lo mismo para el

caso de las deserciones voluntarias, ya que las principales variables aún no están identificadas.

Por lo tanto, dada la imposibilidad de la escuela por identificar tempranamente estas decisiones a través de un sistema de alerta temprana, esta tesis estará enfocada en las deserciones voluntarias.

Análisis de la Deserción por Semestre Académico

Es importante tomar entender la descripción temporal de la deserción. En otras palabras, aquellos estudiantes que renunciaron en su $i-ésimo$ semestre, significa que el estudiante renunció una vez terminado i semestres desde que comenzó sus estudios en el programa, sin importar las postergaciones. La justificación de utilizar esta definición es porque la decisión de desertar puede ser ejecutada por el alumno en cualquier momento del año académico, pudiéndose presentar solicitudes al inicio, mitad o final de un semestre en particular. De esta manera, aquellos estudiantes que elevan solicitud durante un mismo semestre, serán agrupados como desertores en el $i-ésimo$ semestre, donde i es el último semestre cronológico del programa.

Por ejemplo, todas las solicitudes de deserciones voluntarias realizadas durante el segundo semestre del 2012, serán etiquetadas como deserciones del primer semestre 2012, y serán catalogados como deserción en el $i-ésimo$ semestre de acuerdo al año de ingreso. De esta manera, si un estudiante ingresó el 2011, entonces la deserción manifestada durante el segundo semestre del 2012 se considerará como deserción en el semestre 3.

La descripción anterior facilita la agrupación de solicitudes de renuncia elevadas en un mismo periodo semestral. De esta manera, al momento de cuantificar la frecuencia de las deserciones por semestre de avance, se obtiene el **Gráfico 1**:

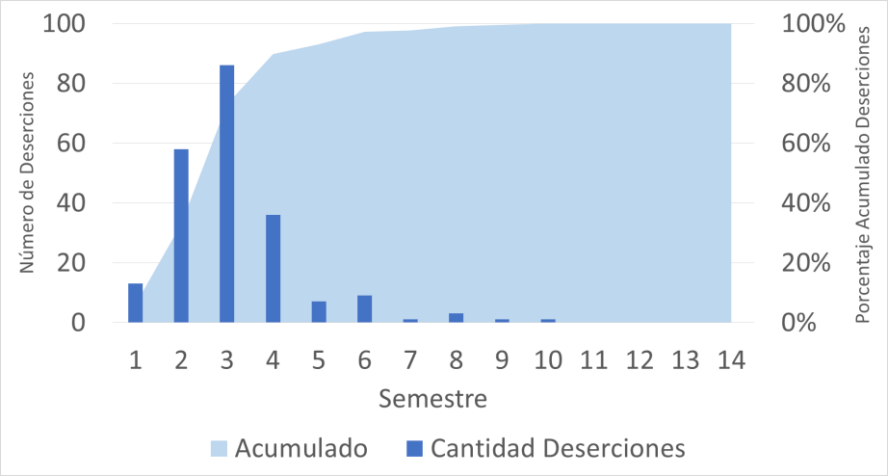

Gráfico 1. Comportamiento deserción por semestre.

Tal como se puede observar, las deserciones se concentran en los 3 primeros años del programa, siendo aproximadamente el 97% del total.

Las cantidades de deserciones crecen hasta el tercer semestre, alcanzando su punto más alto ese mismo semestre. Luego bajan considerablemente en los siguientes tres, siendo casi nula para el séptimo semestre y posteriores.

Por lo tanto, los tres primeros años del programa son cruciales para la detección de las decisiones de dejar la carrera, por lo que en esta tesis las deserciones de los 6 primeros semestres serán parte del objeto de estudio.

4. EXPERIMENTO COMPUTACIONAL

4.1 Descripción de la Muestra de Estudio

La muestra utilizada en esta tesis corresponde a todos los alumnos registrados en la carrera de Ingeniería en Información y Control de Gestión que ingresaron por vía PSU entre los años 2007 al 2014, de la Facultad de Economía y Negocios de la Universidad de Chile. Bajo esta definición, el total de estudiantes abarcados son 928.

De acuerdo a la información recopilada de la muestra de estudios, el número de ingresos por año varía entre 100 y 124 observaciones, cuya distribución por cada año se muestra en la **Tabla 5**.

Año Ingreso	Cantidad	Distribución
2007	117	12,61%
2008	110	11,85%
2009	111	11,96%
2010	124	13,36%
2011	117	12,61%
2012	124	13,36%
2013	100	10,78%
2014	125	13,47%
Total	**928**	**100,00%**

Tabla 5. Cantidad y Distribución de ingresos por año.

4.1 Bases de Datos

Los datos utilizados en esta tesis se recopilaron desde tres bases de datos: **(1)** Sistema de Administración Docente (SAD), **(2)** Becas y Créditos y **(3)** Base DEMRE.

1. **Sistema de Administración Docente (SAD)**: Este sistema es considerado como estratégico por la FEN, puesto que resguarda toda la información de los

estudiantes respecto a las actividades académicas realizadas durante su transcurso por la facultad.

SAD almacena y mantiene segura toda la información de los estudiantes, docentes y ayudantes respecto de la inscripción de ramos, cátedras cursadas, homologación de ramos, desempeño académico, solicitudes estudiantiles y evaluación docente.

2. **Becas y Créditos**: Esta base es gestionada por la unidad de Bienestar Social de la facultad y es usada como apoyo en el proceso de recopilación de información socioeconómica de los alumnos para su posterior evaluación de financiamiento.

   La información almacenada en el sistema está relacionada con todas las ayudas financieras, tanto el monto como el tipo de ayuda (becas, créditos y mantención) que los estudiantes reciben todos los años a través del sistema “Becas y Créditos” del Ministerio de Educación, como también la recibida de manera interna en la Universidad de Chile y/o en la FEN.

   Es importante destacar que cualquier otra ayuda financiera otorgada por otras organizaciones privadas o públicas no suscritas en este sistema no son almacenados en esta base.

3. **Base de Datos DEMRE**: Anualmente el Departamento de Evaluación, Medición y Registro Educacional (DEMRE), administrado por la Universidad De Chile, envía a cada institución educacional superior suscrita al proceso de admisión nacional, toda la información relacionada con las postulaciones de los estudiantes que rindieron la PSU ese año.

   La base de datos está contienen información sociodemográfica, rendimiento académico preuniversitaria, puntajes PSU e historial de postulación de cada estudiante. Adicionalmente adjunta un documento con el diccionario de datos de la base.

Es importante destacar que toda la información entregada por el DEMRE corresponde a información proporcionada por el estudiante al momento de inscribirse para rendir la PSU, por lo que mientras avanza académicamente el estudiante en la carrera de estudio, el estado declarado antes de ingresar puede cambiar. Lamentablemente en esta tesis no se tiene información de las variables para cada semestre, por lo que será discutido este efecto en el siguiente capítulo.

Número de registros de la base de datos Final

El total de estudiantes por semestre se presenta en la **Tabla 6**. Del total de registros solamente se utilizaron aquellos que estuvieran completos, puesto que el eventual reemplazo de valores vacíos y *outliers* puede generar ruido en el desempeño de los modelos. Por lo tanto, solamente se utilizaron datos que reflejaran fielmente el comportamiento de la deserción de los estudiantes, en otras palabras, fueron eliminados de la base los registros que contenían al menos una variable nula y fueron identificados como *outliers* a través del análisis estadístico de rangos.

De esta manera, el 20% de los registros contenía datos vacíos o fue identificado como *outliers* y solamente un 80% no tenía problemas de datos nulos o se alejaba de la concentración de los datos.

Semestre	Deserta	No Deserta	Total
Sem 1	13	595	608
Sem 2	58	591	649
Sem 3	86	591	677
Sem 4	36	578	614
Sem 5	7	486	493
Sem 6	9	577	586
Total general	**209**	**3418**	**3627**

Tabla 6. Número de registros por clase y por semestre según bases de datos SAD, DEMRE y Bienestar,

.

4.2 Variables Escogidas

Variables Sociodemográficas

Las variables sociodemográficas fueron obtenidas desde la base DEMRE. Los datos que contiene esta base son los siguientes:

1. **Grupo familiar**: Los datos disponibles sobre el grupo familiar corresponden principalmente al número de integrantes del grupo, el nivel educacional de ambos padres, el número de padres vivos, cuantos integrantes del grupo trabajan y cuantos integrantes de la familia estudian en los distintos niveles del sistema educacional chileno (prebásica, básica, secundaria, superior y otros institutos como 2x1).

2. **Caracterización del Estudiante**: Adicionalmente la base de datos DEMRE entrega información respecto del género del estudiante, las horas que dedicaba al trabajo hasta antes de ingresar a la universidad y la preferencia con la que fue aceptado al ingresar al programa de estudio de esta tesis.

Variables Institución Educacional Preuniversitaria

Las variables relacionadas con la institución educacional del estudiante del cual egresó, son obtenidas desde la base DEMRE. Los datos obtenidos son explicados a continuación:

1. **Rama Educacional:** De acuerdo a la información contenida en la base DEMRE, es posible rescatar la Rama Educacional del colegió del cual egresó el estudiante. Las ramas definidas en la base son dos: Técnico y Humanista.

2. **Régimen Educacional**: Adicionalmente es posible obtener el régimen educacional del colegio, el cual puede ser Femenino; es decir solamente estudiantes mujeres estudiaban en el establecimiento educacional; Masculino, establecimiento solamente para estudiantes hombres; o Coeducacional, tanto hombres como mujeres eran aceptados como estudiantes.

3. **Grupo Dependencia:** La base DEMRE también entrega información respecto a la dependencia del colegio, el cual está relacionado con el tipo de financiamiento que recibía. En la base existen solamente tres: Municipal, el colegio financiado totalmente de manera estatal; Particular Subvencionado, en otras palabras, el establecimiento educacional es financiado tanto por los apoderados como por el estado; y finalmente Particular Pagado, donde el establecimiento no recibe fondos del estado y es financiado completamente por privados.

4. **Preferencia Postulación:** De acuerdo al proceso actual de postulación a las instituciones de educación superior, los alumnos tienen la facultad de elegir la preferencia de postulación. De esta manera, es posible capturar en qué orden de preferencia el estudiante fue seleccionado a la carrera en estudio, siendo el número 1 como la primera preferencia y el número 4 como la última.

Variables de Desempeño y Actividades Académicas

Las variables relacionadas con el desempeño académico del estudiante es obtenida tanto desde la base DEMRE como del SAD. El desempeño preuniversitario es recopilado por el DEMRE, mientras que el SAD recopila toda la información de actividades y desempeño académico del estudiante durante su transcurso por el programa. Las variables obtenidas son:

1. **Desempeño Preuniversitario**: El DEMRE permite obtener el desempeño preuniversitario medido como el promedio de las notas de enseñanza media, el cual es utilizado para postular a los distintos programas de la educación superior.

2. **Desempeño PSU:** Adicionalmente, el DEMRE contiene información relacionada con el puntaje obtenido en cada sección de la PSU. Las variables de desempeño son el puntaje logrado en las secciones de Lenguaje y Comunicación, Matemáticas, Ciencias e Historia.

Es importante destacar que los alumnos pueden postular con uno de las secciones de Ciencias e Historia, el cual regularmente es el más alto entre los dos. Para poder trabajar con estos datos, se generó una variable de equivalencia que contiene el máximo entre los dos puntajes.

3. **Desempeño Académico Universitario:** El SAD entrega información detallada respecto al desempeño académico del estudiante en cada semestre activo académicamente. El nivel de detalle permite obtener la nota final conseguida en cada semestre y su equivalencia en créditos universitarios por cada uno, lo que finalmente permite calcular el promedio ponderado de notas por semestre (según nota y créditos del curso), el promedio ponderado de notas del semestre anterior (desde el segundo semestre en adelante), el promedio ponderado de notas acumulados hasta el presente semestre (incluye todas las notas y sus respectivos créditos de los semestres anteriores y el semestre en estudio), el porcentaje de créditos reprobados, el porcentaje de créditos reprobados en el semestre anterior y el porcentaje acumulados de créditos reprobados hasta el presente semestre (incluye todas los créditos de los cursos reprobados en los semestres anteriores y el semestre en estudio).

Variables Ambientales

Se entiende como variables ambientales como todas aquellas que capturan la relación del estudiante con su ambiente universitario. Actualmente en FEN algunas de estas son capturadas, principalmente las relacionadas con los distintos tipos de solicitudes y su satisfacción con el cuerpo docente.

1. **Solicitudes**: El sistema de administración docente actualmente centraliza de manera digital todas las solicitudes tanto académicas como no académicas de los estudiantes. Dentro de estas están las relacionadas con la postergación de estudios y renuncia y permanencia al programa de estudios. El primer tipo de solicitud permite identificar si el alumno congela temporalmente sus estudios en un semestre en particular, como también la cantidad de semestres en que los

postergó. Los de segundo tipo de solicitud permite identificar cuando un estudiante expresa formalmente su intención de dejar (permanecer en) la carrera, lo que permite identificar el semestre en que un alumno deserta del (se mantiene en el) programa.

2. **Semestres Veranos:** Adicional al desempeño académico obtenido en los semestres de verano por los estudiantes, el SAD permite obtener la variable que indica si un estudiante participó voluntariamente en el último semestre estival, como también la cantidad de veces que lo ha hecho.

3. **Evaluación Docente:** Como política de retroalimentación al cuerpo docente, al final de cada semestre todos los alumnos deben responder una encuesta que recopila información respecto a su apreciación del desempeño del docente de cátedra. Los alumnos deben responder una encuesta (denominada como Evaluación Docente) que abarca la evaluación del curso, el cuerpo docente y los ayudantes del alumno. En el año 2012 esta encuesta sufrió un cambio en su estructura, lo que no permite la comparación de los elementos evaluados en los años 2007 al 2011. Por lo tanto, para efectos de esta tesis, solamente se tomará en cuenta la nota final que el alumno asigna a sus profesores cada semestre, generando así una variable que permite calcular el promedio de notas asignado a todos los docentes para un semestre, para un semestre posterior y el promedio acumulado hasta el semestre en estudio. Se considera esta variable como un *proxy* de satisfacción con la institución educacional.

Variables Financiamiento

Las variables de financiamiento se obtienen desde las bases DEMRE y BECAS. En el caso de la primera se recupera la información respecto al financiamiento potencial dado el ingreso familiar, mientras que desde la base BECAS se obtiene el monto total que el alumno obtiene anualmente como financiamiento, sin importar el tipo (beca o crédito).

Por lo tanto, desde DEMRE se genera una variable que captura de forma categórica el nivel de ingreso bruto familiar y desde BECAS el nivel de financiamiento anual.

Definición de la Variable dependiente

La deserción es considerada dentro de la disciplina de minería de datos del tipo categórico. La variable que almacena este comportamiento describe dos estados finales del estudiante: (1) Estudiante que deserta y (2) Estudiante que no deserta. Por lo tanto, la variable dependiente será generada por dos clases que representen cada uno de los estados finales de los estudiantes. Para aquellos que desertan, la clase será DESERTA, mientras que para los que no deserta será NO DESERTA.

El diccionario de datos de cada una de las variables anteriormente descritas se encuentra detalladamente en el **Anexo 1.**

4.3 Preprocesamiento y Transformación

Las variables categóricos tales como, dependencia educacional, régimen educacional, rama educacional y género fueron transformados a números para ser trabajados en las técnicas de minería de datos. De esta manera, para cada atributo se generaron n nuevas columnas, donde n era la cantidad de distintas categorías únicas del atributo. De estas n nuevas columnas se seleccionaron $n-1$, con el objetivo de eliminar problemas de multicolinealidad.

Las variables del tipo numéricos tienen distintos rangos, como es el caso de nivel de financiamiento y las notas del estudiante. Esta diferencia puede generar problemas de desempeño y ruido al momento de aplicar los algoritmos de las técnicas de minería de datos. Por lo tanto, todas las variables numéricas fueron escaladas en un rango de 0 a 1 utilizando la siguiente fórmula:

$$x'_{ij} = \frac{x_{ij} - \min(\mathbf{x_j})}{\max(\mathbf{x_j}) - \min(\mathbf{x_j})}$$

Donde x'_{ij} es la nueva escala para el registro la variable $j-ésima$ del registro $i-ésimo$ de la base de datos y $\mathbf{x_j}$ el vector de datos de la variable $j-ésima$. Esta fórmula es denominada técnicamente como normalización de rango 0-1.

4.4 Modelos Predictivos

El objetivo general de esta tesis es la creación de modelos predictivos de la deserción. Bajo este enfoque, se generaron 48 modelos predictivos, los cuales combina cuatro máquinas de aprendizaje, aplicación y no aplicación de clasificador, la aplicación y no aplicación de técnicas de balanceo y la aplicación y no aplicación de *clusterización*. Este enfoque es considerado como Enfoque de Modelos Híbridos, puesto que combina distintos tipos de técnicas de minería de datos.

Para cada tipo de técnicas se aplicaron las siguientes configuraciones paramétricas.

Desbalance

Se aplicaron dos técnicas de muestreo para balancear las bases de datos, siendo una de ellas el aumento de registros según muestreo ROS y la otra una reducción de la base según muestreo RUS. Al igual que en las técnicas anteriores, también se generarán modelos sin la aplicación de muestreos.

Balanceo	Descripción
NoBalanced	No se aplica ninguna de las técnicas de balance.
ROS	Se aplica la técnica de balance según muestreo ROS
RUS	Se aplica la técnica de balance según muestreo RUS

Tabla 7. Identificador y descripción de la aplicación de técnicas de desblance.

Clusterización

Antes de aplicar las máquinas de aprendizaje, umbral de clasificación y muestras de desbalance, se aplicaron algoritmos que identificaron el número de centroides y la asignación de cada registro a estos centroides. Adicionalmente, se generaron modelos sin la aplicación de esta técnica.

La descripción de la clasterización se muestra en la **Tabla 8.**

Clusterización	Descripción
n-Clusters	*Clustring* con la identificación de n-centrodies y la asignación de cada registro a uno de los n-centroides.
NoCluster	No se aplica *Clustring* .

Tabla 8. Identificador y descripción de la aplicación de clusterización.

Máquinas de Aprendizaje

En total se aplicaron cuatro máquinas de aprendizajes: Support Vector Machine, Árbol de Decisión, Regresión Logística y Redes Neuronales Artificiales. Para cada una se optimizó la configuración paramétrica a través de la comparación del desempeño de las siguientes combinaciones de parámetros[2]:

Máquina de Aprendizaje	Identificador	Grilla de Parámetros
Support Vector Machine	SVM	C: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.
Árboles de Decisión	DT	Máxima Profundidad: 5, 10, 15, 20. Mínimo Tamaño para Separación: 2, 5, 10, 15, 20, 25 Poda: Sí, No. PrePoda: Sí, No.
Redes Neuronales	NN	Tamaño Capa Oculta: 1, 5, 10, Automático. Ciclos de Entrenamiento: 500, 1000, 1500.
Regresión Logística	LR	C: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

Tabla 9. Grilla por Máquina de Aprendizaje

Para los 48 modelos de cada semestre, se realizaron optimizaciones de los parámetros a través de la implementación de las combinaciones según la grilla. Una vez obtenido el desempeño de cada uno, se escogió la combinación de parámetros con el desempeño más alto.

Umbral de Clasificador

Para cada modelo, posterior a la aplicación de una máquina de aprendizaje, se aplicó un operador que identificó el umbral óptimo del clasificador bajo el análisis ROC. Adicionalmente, se evaluó el desempeño de los modelos sin la aplicación de este operador.

[2] Para mayor detalle de los parámetros, revisar Capítulo 2.

La descripción de los umbrales de clasificador se muestra en la **Tabla 10**.

Clasificador	Descripción
ThresNo	No se aplica ninguna de las técnicas de clasificador
ThresSí	Se aplica la técnica del clasificador a través del análisis ROC e identificación del Umbral óptimo

Tabla 10. Identificador y descripción de la aplicación de clasificador con umbral.

Tomando las posibles aplicaciones de cada técnica, se obtiene un total de 48 modelos distintos aplicados por cada semestre. En total, se implementaron un total de 288 modelos. Los distintos modelos por semestre se muestran en las siguientes tablas.

4.5 Implementación de Modelos Predictivos en Software

El software utilizado en esta tesis fue el programa gratuito llamado RapidMiner. La aplicación de cada técnica fue realizada según la descripción de los siguientes párrafos.

Paso 0 – Preprocesamiento y Transformación

Para cada base de datos semestral, con el objetivo de utilizar solamente registros, se consideró como registro incompleto aquel que tuviera atributos con valores vacíos pero que no eran explicados por la naturaleza del negocio. Por ejemplo, los estudiantes que tenían promedio de notas del semestre anterior vacío cuando había congelado estudios ese semestre, no era considerado como registro incompleto (en ese caso se usó el promedio de notas acumulado). Pero si tenía vacío en el atributo grupo dependencia del colegio en que egresó, sí era considerado como registro incompleto.

Los atributos categóricos o polinominales fueron transformados a binominales. De esta manera, para cada atributo se generaron n nuevas columnas, donde n es la cantidad de distintas categorías únicas del atributo. De estas n nuevas columnas se seleccionaron

$n-1$ para evitar problemas de multicolinealidad. Por otro lado, los atributos numéricos fueron normalizados en un rango de 0 a 1.

Los operadores utilizados fueron **Filter Examples**, **Nominal to Numerical**, **Select Attributes** y **Normalize.**

Paso 1 - Clusterización

Luego de obtener las bases por semestre transformadas se aplicó para cada base un algoritmo de *Clusterización* que identificó los n centroides. Adicionlamente, se asigó cada registro a uno de los centroides. Este algoritmo es nombrado en el software como el operador **X-means.**

El operador llamado **X-means** es un algoritmo heurístico iterativo que analiza el indicador *Bayesian Information Criteria* (BIC), el cual balancea los costos asociados entre la precisión y complejidad del modelo (Pelleg & Moore, 2000). El resultado de este operador es la base de datos original más un atributo adicional que identifica a cuál centroide pertenece cada registro.

Paso 2 - Máquinas de Aprendizaje

Una vez identificado el número de centroides y a cuál pertenece cada registro, la base es replicada $n+1$ veces, donde se aplicará la validación cruzada para cada grupo de registros con el mismo clúster, como también a la base sin separar por clúster.

RapidMiner cuenta con una gran librería de algoritmos de máquinas de aprendizaje. Para una misma máquina existen distintos algoritmos, cuya configuración está descrita en la sección de ayuda de la plataforma y pueden ser usados según la necesidad del usuario.

Todos los algoritmos reciben como entrada la base de datos con la que aprenderán y validarán el modelo generado. Sin embargo, para que tengan un correcto funcionamiento, deben ser configurados a través del ingreso de parámetros que varían según el algoritmo de la máquina de aprendizaje. Es aquí donde se aplicó las distintas

configuraciones de los parámetros anteriormente descritas y se obtuvo el conjunto que obtenía el mayor desempeño de los modelos.

Paso 3 - Umbral de Clasificación

El umbral de clasificador es aplicado en el proceso de testeo de las máquinas de aprendizaje. Rapidminer cuenta con algoritmos en su librería que permite identificar el mejor umbral dado los costos de errores de clasificación para cada una de las clases. De esta manera, después de obtener el modelo según máquina de aprendizaje aplicada, se obtiene el *confidence* para cada registro, el cual es utilizado como input del operador que identifica el mejor umbral.

Por lo tanto, luego de que el algoritmo generara el modelo en la sección de entrenamiento del operador, se aplicó el algoritmo que identificara el mejor umbral y se procedió a la evaluación de este clasificador. La aplicación de este algoritmo se realizó en la sección de evaluación del algoritmo de las máquinas de aprendizaje.

Paso 4 - Desbalance

Las técnicas de desbalance se aplicaron en el proceso de entrenamiento, antes de que cada máquina de aprendizaje recibiera la base de entrenamiento. De esta manera, en la zona de entrenamiento, se generará un subproceso que formó una nueva base de datos balanceada aplicando según las técnicas de muestreo ROS y RUS.

El resultado entregado del subproceso era una nueva base que en el caso de ROS, era más grande que la original puesto que se repetían observaciones de la clase minoritaria (estudiantes que desertan) hasta igualar la cantidad de registros de la clase mayoritaria (estudiantes que no desertan); y en el caso de RUS, era más pequeña puesto que se eliminaban aleatoriamente registros de la clase mayoritaria hasta igualar la minoritaria.

4.6 Desempeño de los Modelos

En cada semestre se generó un *ranking* de los modelos considerando la precisión de predicción. Para la medición de los indicadores de desempeño se definió como **positiva** la clase DESERTA y **negativa** la clase NO DESERTA. De esta manera, el Ratio Clase Positiva (TPR, por su siglas en inglés) es la precisión del modelo de predecir correctamente cada registro de la clase DESERTA.

En base a estos indicadores, se muestra a continuación el desempeño obtenido de los mejores 10 modelos de 48 por semestre.

Semestre 1

Para este semestre la mejor combinación de técnicas fue en las que se utilizó SVM, *Clusterización*, umbral de clasificador y ninguno de las técnicas de balanceo. Cabe destacar que en las primeras 10 combinaciones, en siete de ellos se utiliza SVM, en cinco *Clusterización* identificando 3 centroides, en cinco un umbral de clasificador y en seis ninguna técnica de balanceo

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	SVM	Clustering	ThresSí	NoBalanced	90,69%	100,00%
2	SVM	NoCluster	ThresSí	NoBalanced	87,72%	100,00%
3	SVM	Clustering	ThresSí	ROS	87,33%	100,00%
4	NN	Clustering	ThresSí	NoBalanced	86,34%	100,00%
5	SVM	Clustering	ThreshNo	NoBalanced	97,03%	40,00%
6	SVM	Clustering	ThreshNo	ROS	94,65%	50,00%
7	NN	NoCluster	ThresSí	ROS	85,54%	100,00%
8	SVM	NoCluster	ThreshNo	ROS	96,04%	40,00%
9	SVM	NoCluster	ThreshNo	NoBalanced	97,62%	30,00%
10	NN	NoCluster	ThreshNo	NoBalanced	97,43%	30,00%

Tabla 11 Ranking de 10 modelos con la mejor precisión para el primer semestre.

Tomando en cuenta el desempeño según promedio del TPR, los modelos más bajo fueron en aquellos que se utilizó las máquinas de aprendizaje del tipo Regresión Logística, con *clusterización,* sin aplicación de umbral de clasificador y sin técnica de balance, tal como lo muestra la siguiente tabla:

	PROMEDIO	
	Precisión	**TPR**
DT	88,6%	14,2%
LR	78,5%	66,7%
NN	82,4%	67,5%
SVM	86,1%	73,3%
Clustering	84,7%	55,0%
NoClustering	83,1%	55,8%
ThreshNo	84,6%	37,5%
ThresSí	83,2%	73,3%
NoBalanced	92,2%	49,4%
ROS	88,8%	50,6%
RUS	70,7%	66,3%

Tabla 12. Desempeño promedio por técnica para el primer semestre.

Es importante destacar la casi nula diferencia en el desempeño entre los modelos donde se aplica clusterización y en los que no. En términos de Precisión aumenta un poco más de 1 punto y en el caso del TPR disminuye en 0.8 puntos. Esto significaría que al menos la distribución de las observaciones que representan estudiantes que desertan, no se concentra en estudiantes con similares características.

Semestre 2

Para este semestre el mejor modelo nuevamente fue en el que se aplicó SVM, *clusterización*, un umbral de clasificador y ninguno de las técnicas de balanceo. Cabe destacar que en los mejores 10 modelos, en solo tres de ellos se utilizó SVM y en cuatro

LR, en cinco *clusterización* identificando 3 centroides, en todos se aplicó un umbral de clasificador y en seis se aplicó alguna técnica de balanceo (3 ROS y 3 RUS).

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	SVM	Clustering	ThresSí	NoBalanced	80,08%	77,27%
2	NN	Clustering	ThresSí	NoBalanced	77,44%	77,27%
3	NN	Clustering	ThresSí	ROS	75,28%	78,57%
4	SVM	NoClustering	ThresSí	RUS	68,98%	88,64%
5	SVM	NoClustering	ThresSí	ROS	68,05%	86,36%
6	LR	NoClustering	ThresSí	ROS	69,36%	84,09%
7	LR	Clustering	ThresSí	NoBalanced	75,19%	75,00%
8	NN	NoClustering	ThresSí	NoBalanced	76,13%	72,73%
9	LR	NoClustering	ThresSí	RUS	66,54%	86,36%
10	LR	Clustering	ThresSí	RUS	73,87%	75,00%

Tabla 13 Ranking de 10 modelos con la mejor precisión para el segundo semestre.

El desempeño promedio en general fue más alto para los modelos con SVM. En el caso de *clusterización*, la no aplicación de esta técnica generó mejores resultados en promedio, mientras que la ejecución de un umbral de clasificador sí generó un efecto positivo.

	PROMEDIO	
	Precisión	**TPR**
DT	79,6%	29,2%
LR	71,9%	59,3%
NN	72,2%	59,8%
SVM	72,0%	60,4%
Clustering	71,5%	52,7%
NoClustering	76,3%	51,6%
ThresSí	75,0%	64,3%
ThresNo	72,9%	40,1%
NoBalanced	84,7%	32,1%
ROS	78,6%	52,4%
RUS	58,5%	72,0%

Tabla 14 Desempeño promedio por técnica para el segundo semestre.

Llama la atención que aunque el mejor modelo no está compuesta por alguna aplicación de técnica de balance, en general el muestreo ROS generó en promedio mejores desempeños en los modelos. Adicionalmente, al igual que en el primer semestre, las técnicas de balance aumentan el TPR promedio, pero disminuyen la precisión promedio de los modelos. Esto se explica principalmente porque al tanto en las técnicas ROS y RUS los algoritmos reciben menor cantidad relativa de observaciones de la clase de NO DESERTA, pero sí más de la clase DESERTA.

Semestre 3

En el caso del tercer semestre el mejor modelo tuvo algunas semejanzas con el del semestre anterior. Nuevamente las técnicas SVM, umbral de clasificador y base de datos sin balancear aparecen en el mejor modelo, pero se genera una diferencia en el no uso de *clusterización*. Cabe destacar que en los mejores 10 modelos sólo aparecen las máquinas de aprendizaje SVM y LR; y adicionalmente en seis de los top diez no se hace uso de la *clusterización* y en todos nuevamente sí se aplica un umbral de clasificador. Al

igual que el semestre pasado, en seis de los modelos se utiliza alguna técnica de balanceo (3 ROS y 3 RUS).

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	SVM	NoClustering	ThresSí	NoBalanced	72,66%	76,00%
2	LR	Clustering	ThresSí	RUS	71,04%	76,00%
3	LR	NoClustering	ThresSí	RUS	68,88%	77,33%
4	SVM	Clustering	ThresSí	NoBalanced	77,88%	69,33%
5	SVM	NoClustering	ThresSí	RUS	64,93%	80,00%
6	LR	NoClustering	ThresSí	NoBalanced	75,54%	70,67%
7	LR	NoClustering	ThresSí	ROS	68,71%	76,00%
8	LR	Clustering	ThresSí	ROS	77,70%	68,00%
9	LR	Clustering	ThresSí	NoBalanced	75,90%	69,33%
10	SVM	NoClustering	ThresSí	ROS	76,44%	68,00%

Tabla 15 Ranking de 10 modelos con la mejor precisión para el tercer semestre.

Nuevamente el desempeño promedio en general fue más alto para los modelos con la máquina de aprendizaje SVM. En el caso de la técnica *clusterización*, la diferencia es mínima entre aplicar la técnica o no, tal como se discutió en los desempeños de los modelos del primer semestre. Nuevamente la aplicación de un umbral de clasificador arrojó mayores desempeños según TPR, al igual que las técnicas de balance, siendo el RUS el mayor. El rendimiento por técnica se muestra en la siguiente tabla.

	PROMEDIO	
	Precisión	**TPR**
DT	74,8%	25,8%
LR	75,1%	54,4%
NN	71,9%	54,3%
SVM	72,5%	58,0%
Clustering	74,3%	47,4%
NoClustering	72,8%	48,8%
ThresNo	74,1%	37,8%
ThresSí	73,1%	58,5%
NoBalanced	80,9%	31,6%
ROS	74,2%	51,0%
RUS	65,7%	61,8%

Tabla 16 Desempeño promedio por técnica para el tercer semestre.

Semestre 4

En el caso del cuarto semestre el mejor modelo se compone por SVM, *clusterización*, umbral de clasificador y balanceo del tipo ROS. Entre los 10 modelos con mayor precisiones se puede destacar que la máquina de aprendizaje de Redes Neuronales (NN) aparece 3 veces con uno de ellos al tope de la tabla y los otros dos al final; en siete modelos se usa *clusterización*; en siete de ellos se aplica alguna técnica de balance (6 ROS y 1 RUS) y nuevamente en todos se hace uso del umbral de clasificador. Esto último podría significar una tendencia de que esta técnica genera efectos positivos en los modelos de deserción utilizados.

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	SVM	Clustering	ThresSí	ROS	84,17%	85,19%
2	LR	Clustering	ThresSí	ROS	81,56%	88,89%
3	NN	Clustering	ThresSí	ROS	83,17%	85,19%
4	SVM	NoClustering	ThresSí	ROS	78,36%	92,59%
5	SVM	NoClustering	ThresSí	RUS	75,55%	92,59%
6	LR	Clustering	ThresSí	NoBalanced	80,76%	81,48%
7	LR	Clustering	ThresSí	RUS	75,95%	85,19%
8	SVM	Clustering	ThresSí	NoBalanced	73,35%	88,89%
9	NN	NoClustering	ThresSí	ROS	71,54%	92,59%
10	NN	Clustering	ThresSí	NoBalanced	78,56%	77,78%

Tabla 17 Ranking de 10 modelos con la mejor precisión para el cuarto semestre.

Al igual que los tres primeros semestres el desempeño en general fue mejor para los modelos con la máquina de aprendizaje SVM. En el caso de *clusterización,* la aplicación de esta técnica entregó un mejores resultados notorios para la precisión, pero no muestra una mejora y diferencia significativa para el TPR, el cual muestra la precisión del modelo para la clase DESERTA. Al igual que todos los semestres anteriores, la aplicación de un umbral de clasificador arrojó mayores desmepños en promedio. Mientras que los modelos con muestreo RUS es el que arrojó mayor desempeño según TPR, pero a la vez el más bajo para la precisión general.

	PROMEDIO	
	Precisión	**TPR**
DT	83,6%	23,8%
LR	77,7%	59,6%
NN	78,6%	58,0%
SVM	76,3%	65,4%
Clustering	80,4%	51,1%
NoClustering	77,7%	52,3%
ThresNo	80,2%	36,3%
ThresSí	77,9%	67,1%
NoBalanced	85,9%	32,4%
ROS	82,7%	55,6%
RUS	68,6%	67,1%

Tabla 18 Desempeño promedio por técnica para el cuarto semestre.

Hasta el momento, los desempeños promedios obtenidos en las técnicas de máquinas de aprendizaje, umbral de clasificación y balance mantienen las mismas tendencias, lo que se espera que se mantenga para los modelos mostrados en los siguietnes dos semestres.

Semestre 5

El modelo compuesto por LR, *clusterización*, umbral de clasificador y sin balanceo fue el que obtuvo mayores desempeños de precisión y TPR. En este caso también existe un segundo modelo con exactamente los mismos resultados, el cual se compone por la máquina de aprendizaje Redes Neuronales, *clusterización,* umbral de clasificador y balance ROS.

Cabe destacar que en los primeros 10 modelos el SVM, LR y NN aparecen casi la misma cantidad de veces; el *clusterización* con dos centroides en prácticamente todos los modelos; el umbral de clasificador en siete y las técnicas de balance en seis (4 ROS y 2 RUS).

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	LR	Clustering	ThresSí	No Balanced	94,47%	100,00%
2	NN	Clustering	ThresSí	ROS	94,47%	100,00%
3	SVM	Clustering	ThresSí	ROS	93,97%	100,00%
4	LR	Clustering	ThresSí	ROS	93,47%	100,00%
5	SVM	Clustering	ThresSí	No Balanced	92,71%	100,00%
6	LR	Clustering	ThresSí	RUS	92,71%	100,00%
7	LR	Clustering	ThresNo	No Balanced	99,25%	40,00%
8	SVM	Clustering	ThresSí	RUS	91,71%	100,00%
9	NN	Clustering	ThresNo	ROS	98,49%	40,00%
10	SVM	NoClustering	ThresNo	No Balanced	98,24%	40,00%

Tabla 19 Ranking de 10 modelos con la mejor precisión para el quinto semestre.

En este semestre, el promedio de la precisión y TPR fue más alto en aquellos donde se aplicó la técnica NN. A diferencia de los semestres anteriores, los modelos con *clusterización* obtuvieron mejores resultados. La aplicación del umbral de clasificador arrojó mayores precisiones y TPR promedios, aunque cabe destacar que en el ranking de los 10 mejores modelos por primera vez aparecen modelos en donde no se utilizó esta técnica. El muestreo ROS es el mejor rendimiento entregó en los modelos, ya que en promedio los modelos que usaron esta técnica obtuvieron precisiones más altas y solamente se encuentra a 7 puntos de diferencia con la técnica RUS en el indicador TPR. Sin embargo, esta última generó a los modelos rendimientos muy bajos en la precisión.

	PROMEDIO	
	Precisión	**TPR**
DT	89,3%	10,0%
LR	85,7%	58,3%
NN	85,8%	66,7%
SVM	86,0%	63,3%
Clustering	89,2%	50,8%
NoClustering	84,1%	48,3%
ThresNo	82,8%	25,8%
ThresSí	90,6%	73,3%
No Balanced	94,9%	41,3%
ROS	95,0%	50,0%
RUS	70,1%	57,5%

Tabla 20 Desempeño promedio por técnica para el quinto semestre.

Cabe destacar a diferencia de los primeros 4 semestres los modelos con redes neuronales obtuvieron en promedio los mejores resultados en TPR y una diferencia de solo 0.2 puntos con los SVM. Sin embargo, esta última máquina de aprendizaje sigue generando buenos reusltados en los modelos, tendencia que se ha identificado en los primeros cuatro semestres.

Semestre 6

Para el sexto semestre el mejor modelo se compone por LR, *clusterización*, umbral de clasificador y balance ROS. En los tres mejores modelos se usa LR; el *clusterización* es aplicado en siete de los diez modelos; las técnicas de balance aparecen en siete (4 ROS y 3 RUS) y el umbral de clasificador, al igual que en los primeros cuatro semestres, aparece en todos los modelos.

La siguiente tabla muestra el desempeño de las 10 mejores modelos que tienen los indicadores de desempeño más alto.

Modelo	Técnica DM	Clusterización	Clasificador	Balanceo	Precisión	TPR
1	LR	Clustering	ThresSí	ROS	95,43%	87,50%
2	LR	NoClustering	ThresSí	No Balanced	92,39%	100,00%
3	LR	NoClustering	ThresSí	ROS	92,13%	100,00%
4	SVM	Clustering	ThresSí	RUS	93,65%	87,50%
5	NN	Clustering	ThresSí	ROS	90,86%	100,00%
6	NN	Clustering	ThresSí	No Balanced	90,36%	100,00%
7	LR	Clustering	ThresSí	RUS	96,19%	62,50%
8	SVM	Clustering	ThresSí	ROS	91,37%	87,50%
9	SVM	Clustering	ThresSí	No Balanced	88,32%	100,00%
10	LR	NoClustering	ThresSí	RUS	87,82%	100,00%

Tabla 21. Ranking de 10 modelos con la mejor precisión para el sexto semestre.

Los modelos en que se usó las máquinas de aprendizaje de Regresiones Lineales obtuvieron desempeños mayores, siendo la principal diferencia de la tendencia generada en los primeros cuatro semestres. En el caso de *clusterización*, nuevamente la aplicación de esta técnica entregó mejores resultados en la precisión, pero peores en el caso del desempeño TPR, lo que podría significar que tal como se discutía anteriormente, esta técnica no generaría beneficios significantes respecto a los desempeños de los modelos.

	PROMEDIO	
Técnica	**Precisión**	**TPR**
DT	89,6%	12,5%
LR	90,1%	58,3%
NN	86,8%	57,3%
SVM	85,7%	56,3%
Clustering	89,7%	42,2%
NoClustering	86,4%	50,0%
ThresNo	86,9%	24,0%
ThresSí	89,2%	68,2%
No Balanced	94,3%	35,9%
ROS	90,8%	43,0%
RUS	79,0%	59,4%

Tabla 22. Desempeño promedio por técnica para el sexto semestre.

Al igual que todos los semestres anteriores, la aplicación de un umbral de clasificador arrojó mejores resultados, siendo muy notoria la diferencia en la predicción de la clase DESERTA.

Finalmente, al igual que semestres anteriores, el no uso de algún muestreo es el que mejores desempeños de precisión, sin embargo, esto no se refleja en los primeros cinco lugares del ranking, puesto que los resultados obtenidos en modelos con TPR son muy bajos.

TÉCNICAS		SEMESTRES						
Clusterización		**Semestre 1**	**Semestre 2**	**Semestre 3**	**Semestre 4**	**Semestre 5**	**Semestre 6**	**Número de Usos**
Clustering		✓	✓		✓	✓	✓	5
NoClustering				✓				1
Balance								
NoBalanced		✓	✓	✓		✓		4
ROS					✓		✓	2
RUS								0
Máquina Aprendizaje								
SVM		✓	✓	✓	✓			4
DT								0
NN								0
LR						✓	✓	2
Umbral								
ThreshNo								0
ThresSí		✓	✓	✓	✓	✓	✓	6
	Precisión	**90,69%**	**80,08%**	**72,66%**	**84,17%**	**94,47%**	**95,43%**	
	TPR	**100,00%**	**77,27&**	**76,00%**	**85,19%**	**100,00%**	**87,50%**	

Tabla 23. Desempeño promedio por técnica para el sexto semestre.

En resumen, tal como lo muestra la **Tabla 23**, las máquinas de aprendizaje de regresión logística (LR) y *support vector machine* (SVM) son los que mejores desempeños obtuvieron, al igual que el umbral de clasificación. Éste último fue usado en todos los mejores modelos. Para el caso de las técnicas de balance no fue tan claro su impacto en el desempeño de los modelos, lo cual puede inducir a que no siempre es conveniente su aplicación, como es el caso también de la *clusterización,* aun cuando éste último es usado en cinco de los mejores modelos para los 6 semestres.

CAPÍTULO 5 - RESULTADOS

En este capítulo se discutirán los resultados entregados por los mejores modelos de cada semestre. Tal como se planteó al inicio de este documento, se identificarán los predictores para las dos clases estudiadas, como también un perfil para cada semestre.

5.1 Predictores

Uno de los objetivos planteados al inicio de este documento establecía la identificación de los predictores del comportamiento de abandonar el programa. De acuerdo al análisis realizado anteriormente, se definió que este comportamiento tiene carácter semestral, por lo que eventualmente las variables que determinen la deserción pueden variar de un semestre a otro.

Una vez analizado el desempeño de todos los modelos realizados a cada semestre, se obtuvo el peso que otorga cada algoritmo. Este peso permite identificar aquellos variables que podrían ser considerados como predictores, como también los que no lo serían por su bajo o nulo peso.

Utilizando este indicador, se realizó una clasificación de los variables para cada semestre. En cada semestre se decidió generar los siguientes 3 grupos:

1. **Predictores Primer Grupo**: Serán considerados como predictores primer grupo las variables que estarán en el 25% más alto del ranking según peso. Para determinar el 25% superior, se utilizarán solamente los que tengan pesos mayores a 0.

2. **Predictores Segundo Grupo**: Serán considerados como predictores segundo grupo las variables que se encuentren bajo el 25% superior y sobre el 50% del ranking según peso. Para determinar el grupo que se encuentre bajo el 25% y sobre el 50% se utilizarán solamente los que tengan pesos mayores a 0.

3. **No Predictores:** serán considerados como no predictores aquellas variables en que el peso sea igual a 0.

De esta manera, y según los pesos de los mejores modelos, se mostrará a continuación los predictores para cada semestre.

Semestre 1

Para el primer semestre, el mejor modelo era el compuesto por SVM, *clusterización*, aplicación de un umbral de clasificador y sin ninguna técnica de balanceo.

Los pesos para cada variable del primer clúster se muestran en la **Tabla 24**. En ella se muestran que las variables consideradas como predictores del primer grupo son si el estudiante es mujer, cuántos de los dos padres están vivos, el promedio de las notas asignadas al profesor por parte del estudiante y el rendimiento obtenido en la PSU. Por otro lado, los predictores del segundo grupo son el rendimiento del estudiante en el colegio, el financiamiento otorgado, el nivel educacional de la madre, el promedio de notas que obtuvo en el semestre y el nivel educacional del padre.

Los predictores listados muestran que en un principio las variables más importantes son principalmente los relacionados con el rendimiento en la PSU y las características propias del estudiante como el género. En el segundo grupo está el nivel educacional del padre y de la madre, como también el financiamiento y el promedio obtenido en el semestre por el estudiante, lo que podría significar alguna fuerte incidencia de los padres en el estudiante para tomar la decisión final de desertar en el primer semestre.

Para este grupo de clúster, las variables relacionadas con el colegio en que se graduaron no tienen impacto con la decisión de desertar, puesto que según el modelo sus pesos son iguales a cero.

CLUSTER 1 Variables	Peso	Cuartil
SEXO = MUJER	32,533	Q1
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	32,533	Q1
EvalProf_Sem	30,406	Q1
PROM_LM_FINAL	25,923	Q1
PTJE_MATEMATICA_FINAL	25,892	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	24,715	Q1
PTJE_NEM	24,619	Q2
FINANCIAMIENTO	24,207	Q2
EDUCACION_DELA_MADRE_EQUIVALENCIA	23,116	Q2
NotaSem	22,982	Q2
EDUCACION_DEL_PADRE_EQUIVALENCIA	22,964	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	16,911	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	15,621	Q3
GRUPO_FAMILIAR	11,947	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	10,844	Q3
Pref	10,414	Q3
%Reprobadas_Sem	7,206	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	4,648	Q4
INGRESO_BRUTO_FAMILIAR_GRUPO	2,958	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	2,755	Q4
REGIMEN = Femenino	1,29	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
REGIMEN = Masculino	0	
GRUPO_DEPENDENCIA = Particular Pagado	0	
GRUPO_DEPENDENCIA = Municipal	0	
PostergaSem	0	
HORAS_QUE_DEDICA_TRABAJO	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0	

Tabla 24 Peso de las variables para el Cluster 1 del primer semestre.

Para el segundo grupo de clúster, los pesos asignados a cada variable se muestran en la **Tabla 25**. En ella se puede ver que nuevamente el número de padres vivos, el rendimiento en la PSU y la evaluación a los profesores en el semestre son considerados dentro del primer grupo de predictores. El nivel educacional de los padres también es considerado dentro de este grupo, a diferencia del primer clúster que es considerado en el segundo.

El segundo grupo de predictores son los relacionados con el financiamiento, el ingreso bruto familiar, su desempeño en el colegio antes de ingresar a la universidad, su desempeño en el semestre y antecedentes familiares tales como número de integrantes y cuantos estudian en cuarto medio antes de ingresar a la universidad.

Nuevamente las variables relacionadas con el colegio en que estudió antes de ingresar a la universidad no son consideradas importantes, a excepción del grupo dependencia del colegio, que si bien no está dentro de los dos cuartiles más importantes, está en el tercero más importante. Adicionalmente, las variables relacionados con las horas de trabajo y el número de integrantes que estudian en la prebasica tampoco tienen algún peso para predecir si el estudiante pertenece a la clase DESERTA o NO DESERTA.

CLUSTER 2		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	132,914	Q1
PROM_LM_FINAL	114,1	Q1
EvalProf_Sem	113,894	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	113,013	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	112,636	Q1
PTJE_MATEMATICA_FINAL	110,179	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	105,03	Q1
PTJE_NEM	80,998	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	80,039	Q2
NotaSem	74,996	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	65,142	Q2
FINANCIAMIENTO	62,272	Q2
GRUPO_FAMILIAR	59,293	Q2

Tabla 25 Peso de las variables para el Cluster2 del primer semestre.

Variables	Peso	Cuartil
REGIMEN = Masculino	52,493	Q3
GRUPO_DEPENDENCIA = Particular Pagado	41,811	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	41,246	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	40,925	Q3
Pref	33,795	Q3
GRUPO_DEPENDENCIA = Municipal	30	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	23,273	Q4
SEXO = MUJER	19,318	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	14,314	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	6,439	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,877	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
REGIMEN = Femenino	0	
PostergaSem	0	
HORAS_QUE_DEDICA_TRABAJO	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	

Tabla 25 (Continuación). Peso de las variables para el Cluster2 del primer semestre. Construcción Propia.

Semestre 2

En el caso del segundo semestre, el mejor modelo también fue compuesto por SVM, *clusterización*, aplicación de un umbral de clasificador y sin ninguna técnica de balanceo.

Los pesos para cada variable del primer clúster se muestran en la **Tabla 26**. En él se muestran que las variables consideradas como predictores del primer cuartil son los que reflejan cuántos de los dos padres están vivos, la evaluación a los profesores en el segundo semestre y en el anterior; y el rendimiento obtenido en la PSU. Por otro lado, los predictores del segundo nivel son el financiamiento, su rendimiento en el colegio antes de ingresar en la universidad, el género, el nivel educacional de sus padres y el porcentaje de créditos reprobados acumulados al final del segundo semestre.

Analizando el listado de las variables podemos ver que para el segundo semestre las variables relacionadas con el ambiente universitario del estudiante toman importancia.

Principalmente los relacionados con la evaluación que realiza el estudiante a los profesores. Adicionalmente, al igual que en el primer semestre, las variables sobre el rendimiento en la PSU se mantienen en el primer grupo; mientras que respecto del segundo grupo, las variables son muy parecidos al del primer clúster en el primer semestre, con la diferencia que las reprobaciones aparecen en este grupo.

Para este primer clúster, las variables relacionados con el colegio en que se graduaron, la postergación y la cantidad de integrantes de la familia que estudia en la básica, media o cuarto medio no tienen algún impacto en determinar si el estudiante pertenece a la clase DESERTA o NO DESERTA.

CLUSTER 1		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	11,599	Q1
EvalProf_Sem_Anterior	10,656	Q1
EvalProf_Sem	10,324	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	9,945	Q1
PROM_LM_FINAL	9,942	Q1
PTJE_MATEMATICA_FINAL	9,484	Q1
EvalProf_Sem_Acum	9,42	Q1
FINANCIAMIENTO	9,334	Q2
PTJE_NEM	8,232	Q2
SEXO = MUJER	7,78	Q2
EDUCACION_DEL_PADRE_EQUIVALENCIA	7,128	Q2
%Reprobadas_Sem_Acum	6,551	Q2
EDUCACION_DELA_MADRE_EQUIVALENCIA	5,948	Q2
NotaSem	5,509	Q3
%Reprobadas_Sem	5,496	Q3
NotaSem_Anterior	4,954	Q3
NotaSem_Acum	4,52	Q3
GRUPO_FAMILIAR	3,151	Q3
%Reprobadas_Sem_Anterior	2,867	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	1,91	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	1,34	Q4
Pref	1,273	Q4

Tabla 26. Peso de las variables para el Cluster 1 del segundo semestre.

Variables	Peso	Cuartil
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	1,253	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	1,108	Q4
INGRESO_BRUTO_FAMILIAR_GRUPO	0,707	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	0,546	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
REGIMEN = Femenino	0	
REGIMEN = Masculino	0	
GRUPO_DEPENDENCIA = Particular Pagado	0	
GRUPO_DEPENDENCIA = Municipal	0	
PostergaSem	0	
PostergaSem_Anterior	0	
PostergaSem_Acum	0	
HORAS_QUE_DEDICA_TRABAJO	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0	

Tabla 26 (Continuación) Peso de las variables para el Cluster 1 del segundo semestre.

Para el segundo grupo de clúster los pesos asignados a cada variable se muestran en la **Tabla 27**. En ella se observa que nuevamente el número de padres vivos, el rendimiento en la PSU y la evaluación a los profesores en el semestre son considerados dentro del primer grupo de predictores. Adicionalmente, las reprobaciones aparecen en este grupo, no en el segundo cuartil de predictores como lo es para el primer grupo de clúster.

El segundo grupo de predictores son los relacionados con la educación de los padres, la evaluación a los profesores en el semestre anterior y el acumulado, género del estudiante y las características del colegio que egresó. En este clúster, a diferencia de los anteriores, la dependencia y régimen sí son considerados como importantes para predecir a qué clase pertenece cada estudiante, siempre y cuando estos sean para indicar si el colegio era de régimen femenino y si era de dependencia municipal.

Las variables relacionadas con el colegio y que indican si el estudiante pertenece a la rama educacional técnico, régimen masculino y dependencia particular pagado no tienen algún peso en este clúster. Adicionalmente, las variables que indican la cantidad de estudiantes que estudian en algún nivel educacional bajo el universitario y las relacionadas con la postergación de estudios en la universidad se repiten dentro del grupo de no predictores.

CLUSTER 2		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	3,288	Q1
%Reprobadas_Sem_Acum	3,101	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	3,028	Q1
PROM_LM_FINAL	2,946	Q1
EvalProf_Sem	2,801	Q1
%Reprobadas_Sem	2,741	Q1
PTJE_MATEMATICA_FINAL	2,732	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	2,728	Q2
EvalProf_Sem_Anterior	2,63	Q2
EvalProf_Sem_Acum	2,172	Q2
SEXO = MUJER	1,928	Q2
REGIMEN = Femenino	1,928	Q2
GRUPO_DEPENDENCIA = Municipal	1,928	Q2
EDUCACION_DELA_MADRE_EQUIVALENCIA	1,908	Q2
PTJE_NEM	1,584	Q3
GRUPO_FAMILIAR	1,371	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	1,286	Q3
%Reprobadas_Sem_Anterior	1,266	Q3
NotaSem_Anterior	1,155	Q3
FINANCIAMIENTO	0,91	Q3
INGRESO_BRUTO_FAMILIAR_GRUPO	0,825	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,643	Q4
NotaSem	0,531	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	0,47	Q4
Pref	0,453	Q4
NotaSem_Acum	0,42	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	0,314	Q4

Tabla 27. Peso de las variables para el Cluster 2 del segundo semestre.

Variables	Peso	Cuartil
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
REGIMEN = Masculino	0	
GRUPO_DEPENDENCIA = Particular Pagado	0	
PostergaSem	0	
PostergaSem_Anterior	0	
PostergaSem_Acum	0	
HORAS_QUE_DEDICA_TRABAJO	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0	

Tabla 27 (Continuación). Peso de las variables para el Cluster 2 del segundo semestre.

Para el tercer grupo de clúster, los pesos asignados a cada variable se muestran en la **Tabla 28**. Nuevamente el número de padres vivos y el rendimiento en la PSU son considerados dentro del primer grupo de predictores. Sin embargo, esta vez aparecen más variables relacionados con los padres, las cuales son el nivel educacional del padre y de la madre. Adicionalmente, al igual que en el primer grupo de clúster, las variables relacionados con la evaluación a los profesores son identificados como parte del primer grupo de predictores.

El segundo grupo de predictores está más relacionado con su rendimiento académico, tanto en el semestre, en el semestre anterior y el rendimiento acumulado. Recordar que los dos últimos variables aparecían dentro del segundo grupo del clúster anterior. Las variables relacionadas con los antecedentes familiares aparecen en este grupo, principalmente aquellos que indican indirectamente como el estudiante financiaría sus estudios, asumiendo que mientras mayor nivel de ingresos bruto familiares, mayor probabilidad que el estudiante pague los estudios usando esta fuente como financiamiento. A diferencia del segundo clúster, el variable que indica la dependencia del colegio como particular pagado sí es importante en este clúster.

Nuevamente las variables relacionadas con el colegio en que estudió antes de ingresar a la universidad no son consideradas importantes, pero solamente las que indican si el estudiante pertenece a la rama educacional técnico y régimen femenino. El variable que indica la cantidad de integrantes del grupo familiar que estudian en la prebasica nuevamente aparece como no importante. Finalmente, las variables relacionadas con la postergación no tienen algún impacto para predecir la clase final del estudiante, al igual que los clúster y semestres anteriores.

CLUSTER 3		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	2310	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	2114,118	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	2068,235	Q1
EvalProf_Sem_Anterior	1998,771	Q1
PROM_LM_FINAL	1991,661	Q1
EvalProf_Sem	1962,429	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	1947,828	Q1
PTJE_MATEMATICA_FINAL	1942,306	Q1
EvalProf_Sem_Acum	1647,654	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	1581,818	Q2
GRUPO_DEPENDENCIA = Particular Pagado	1560	Q2
PTJE_NEM	1447,886	Q2
NotaSem	1353,184	Q2
NotaSem_Anterior	1270,221	Q2
NotaSem_Acum	1233,313	Q2
GRUPO_FAMILIAR	1058,182	Q2
REGIMEN = Masculino	840	Q3
Pref	840	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	840	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	780	Q3
%Reprobadas_Sem_Acum	703,729	Q3
SEXO = MUJER	600	Q3
%Reprobadas_Sem	568,8	Q3
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	514,286	Q3

Tabla 28. Peso de las variables para el Cluster 3 del segundo semestre

Variables	Peso	Cuartil
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	420	Q4
FINANCIAMIENTO	414,171	Q4
%Reprobadas_Sem_Anterior	370,986	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	300	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	267,956	Q4
GRUPO_DEPENDENCIA = Municipal	180	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	100	Q4
HORAS_QUE_DEDICA_TRABAJO	46	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
REGIMEN = Femenino	0	
PostergaSem	0	
PostergaSem_Anterior	0	
PostergaSem_Acum	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	

Tabla 28 (Continuación). Peso de las variables para el Cluster 3 del segundo semestre.

Semestre 3

En el caso del tercer semestre, el mejor modelo también fue compuesto por SVM, aplicación de un umbral de clasificador y sin ninguna técnica de balance. Sin embargo, a diferencia de los primeros dos semestre, la técnica de *clusterización* no es parte del mejor modelo.

Para tercer semestre los pesos asignados a cada variable se muestran en la **Tabla 29**, y que al igual que en los dos semestres anteriores, el variable relacionado con la cantidad de padres vivos y el rendimiento en la PSU son considerados dentro del primer grupo de predictores. Adicionalmente, al igual que en algunos de los clúster del primer y segundo semestre, las variables relacionados con la evaluación a los profesores son identificados como parte del primer grupo de predictores.

El segundo grupo de predictores está más relacionado con su rendimiento académico, tanto en el semestre como en el semestre anterior. Llama la atención que las variables relacionados con el grupo familiar (número de integrantes del familiar y cuántos estudian

en cuarto medio), el género y ciertas características del colegio que egresó el estudiante (por ejemplo, si es de dependencia particular apagado) aparecen en el segundo grupo, puesto que tales variables eran considerados como no predictores por su peso 0 en los clústeres de semestres anteriores.

A diferencia de los semestres uno y dos, las variables relacionados con el colegio que estudiaron antes de ingresar a la universidad ya no aparecen dentro del grupo de variables con peso cero. Solamente se mantienen los relacionados con las postergaciones y el número de integrantes que estudia en la prebásica. Por lo tanto, esto podría demostrar un primer cambio en los tipos de predictores semestrales.

Variables	Peso	Cuartil
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	367,5	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	325,294	Q1
EvalProf_Sem_Anterior	323,743	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	320,294	Q1
EvalProf_Sem	314,557	Q1
PROM_LM_FINAL	313,13	Q1
PTJE_MATEMATICA_FINAL	309,835	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	301,583	Q1
NotaSem_Acum	249,856	Q1
EvalProf_Sem_Acum	246,109	Q2
NotaSem	242,746	Q2
NotaSem_Anterior	217,686	Q2
PTJE_NEM	212,936	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	206,364	Q2
GRUPO_DEPENDENCIA = Particular Pagado	185	Q2
SEXO = MUJER	150	Q2
GRUPO_FAMILIAR	148,75	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	140	Q2

Tabla 29. Peso de las variables para el tercer semestre.

Variables	Peso	Cuartil
Pref	118,333	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	93,333	Q3
GRUPO_DEPENDENCIA = Municipal	90	Q3
REGIMEN = Masculino	85	Q3
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	82,143	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	72,5	Q3
FINANCIAMIENTO	71,146	Q3
Sem_Verano_Ant	70	Q3
%Reprobadas_Sem_Acum	51,338	Q3
%Reprobadas_Sem_Anterior	48,05	Q4
REGIMEN = Femenino	45	Q4
%Reprobadas_Sem	42,4	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	42,175	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	31,667	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	26,667	Q4
HORAS_QUE_DEDICA_TRABAJO	9,833	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	5	Q4
PostergaSem_Acum	5	Q4
PostergaSem	0	
PostergaSem_Anterior	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	

Tabla 29 (Continuación). Peso de las variables para el tercer semestre.

Semestre 4

En el caso del cuarto semestre el mejor modelo está compuesto por SVM, *clusterización* y aplicación de un umbral de clasificador. Cabe destacar que a diferencia de los mejores modelos anteriores, el de este semestre sí está compuesto por alguna técnica de balance y esta es del tipo ROS.

Para el primer grupo de clúster, el peso de cada variable se muestra en la **Tabla 30**. Nuevamente, al igual que en los tres primeros semestres, el variable relacionado con la cantidad de padres vivos, el rendimiento en la PSU y los relacionados con la evaluación a los profesores son identificados como parte del primer grupo de predictores.

El segundo grupo de predictores está más relacionado con el rendimiento académico del estudiante, tanto en el semestre como en el semestre anterior. Nuevamente se destaca

que las variables relacionados con el grupo familiar (número de integrantes del familiar y cuántos estudian en cuarto medio), el género y del colegio que egresaron (si es de dependencia particular apagado) aparecen en el segundo grupo. Esto aportaría nuevamente la idea planteada en el análisis de los semestres anteriores de que desde el segundo año se comienzan a detectar cambios en las variables para determinar la clase a la que pertenecería cada estudiante, pasando a tomar mayor importancia las variables relacionadas con el desempeño y relación con el ambiente universitario (evaluación a profesores).

En el grupo de no predictores se mantienen los relacionados con las postergaciones y se agrega la variable de si tuvo participación en el semestre verano.

CLUSTER 1		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	11,47	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	10,41	Q1
EvalProf_Sem_Anterior	10,007	Q1
PROM_LM_FINAL	9,924	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	9,879	Q1
EvalProf_Sem	9,846	Q1
PTJE_MATEMATICA_FINAL	9,789	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	9,598	Q1
NotaSem	8,449	Q1
NotaSem_Anterior	8,223	Q2
NotaSem_Acum	8,102	Q2
EvalProf_Sem_Acum	7,982	Q2
PTJE_NEM	7,795	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	5,498	Q2
GRUPO_DEPENDENCIA = Particular Pagado	5,361	Q2
SEXO = MUJER	5,115	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	5,063	Q2
GRUPO_FAMILIAR	4,912	Q2

Tabla 30. Peso de las variables para el primer cluster del cuarto semestre.

Variables	**Peso**	**Cuartil**
Pref	3,443	Q3

CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	3,018	Q3
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	2,352	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	1,875	Q3
FINANCIAMIENTO	1,755	Q3
%Reprobadas_Sem_Acum	1,27	Q3
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	1,211	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	1,184	Q3
REGIMEN = Masculino	0,92	Q4
%Reprobadas_Sem	0,819	Q4
%Reprobadas_Sem_Anterior	0,672	Q4
GRUPO_DEPENDENCIA = Municipal	0,552	Q4
REGIMEN = Femenino	0,46	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,286	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0,276	Q4
HORAS_QUE_DEDICA_TRABAJO	0,054	Q4
PostergaSem_Acum	0,046	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
PostergaSem	0	
PostergaSem_Anterior	0	
Sem_Verano_Ant	0	

Tabla 30 (Continuación). Peso de las variables para el primer cluster del cuarto semestre.

En la **Tabla 31** se muestran los pesos para cada variable del Cluster 2. En ella se puede ver que al igual que en el semestre tres y el primer clúster del semestre cuarto, el variable relacionado con la cantidad de padres vivos, el rendimiento en la PSU y los relacionados con la evaluación a los profesores son identificados como parte del primer grupo de predictores. También se encuentran en este grupo el nivel educacional de los padres, tal como aparece para algunos clúster del segundo semestre y en todos los clúster del tercer semestre.

El segundo grupo de predictores está más relacionado con su rendimiento académico, tanto en el semestre cuatro como en el semestre tres. Adicionalmente, las variables relacionadas con el número de integrantes del grupo familiar, el ingreso bruto familiar y financiamiento aparecen en el segundo grupo.

En este clúster las variables que indican si la dependencia del colegio del que egresó es particular pagado y si participó en el semestre verano también pertenecen al segundo grupo de predictores.

Se mantienen los relacionados con las postergaciones como parte de las variables sin pesos. Adicionalmente se repite si la rama educacional del colegio que egresó es técnica, al igual que en el primer clúster.

CLUSTER 2		
Variables	**Peso**	**Cuartil**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	113,121	Q1
EvalProf_Sem_Anterior	100,548	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	99,867	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	99,072	Q1
EvalProf_Sem	96,975	Q1
PTJE_MATEMATICA_FINAL	96,755	Q1
PROM_LM_FINAL	96,545	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	93,985	Q1
EvalProf_Sem_Acum	80,369	Q1
NotaSem	69,313	Q2
NotaSem_Acum	68,545	Q2
NotaSem_Anterior	67,771	Q2
PTJE_NEM	65,532	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	51,601	Q2
GRUPO_FAMILIAR	47,571	Q2
GRUPO_DEPENDENCIA = Particular Pagado	43,867	Q2
Sem_Verano_Ant	40,473	Q2
FINANCIAMIENTO	40,036	Q2

Tabla 31. Peso de las variables para el tercer clúster del cuarto semestre.

Variables	**Peso**	**Cuartil**

%Reprobadas_Sem	35,685	Q3
%Reprobadas_Sem_Acum	35,24	Q3
SEXO = MUJER	34,099	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	33,545	Q3
Pref	31,904	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	29,317	Q3
%Reprobadas_Sem_Anterior	28,34	Q3
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	23,491	Q3
REGIMEN = Masculino	22,556	Q3
GRUPO_DEPENDENCIA = Municipal	19,627	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	17,465	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	13,992	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	7,771	Q4
REGIMEN = Femenino	6,627	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	6,104	Q4
HORAS_QUE_DEDICA_TRABAJO	3,177	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	3	Q4
PostergaSem_Acum	1	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	
PostergaSem	0	
PostergaSem_Anterior	0	

Tabla 31 (Continuación). Peso de las variables para el tercer cluster del cuarto semestre.

Semestre 5

En el caso del quinto semestre el mejor modelo está compuesto por LR, *clusterización,* aplicación de un umbral de clasificador y ninguna técnica de balanceo. Cabe destacar que a diferencia de los cuatro semestres anteriores, aquí la máquina de aprendizaje LR es parte del mejor modelo.

Según la **Tabla 32** en este semestre y clúster, existe una notoria diferencia respecto en los predictores comparado con los identificados en los semestres y clúster anteriores. Las variables relacionados con el desempeño del estudiante, su participación en semestres veranos y postergaciones aparecen dentro del primer grupo de predictores. Adicionalmente, las relacionadas con el nivel educacional de la madre y el número de integrantes del grupo familiar siguen siendo identificadas dentro del primer grupo.

El segundo grupo de predictores se compone primeramente por su rendimiento acumulado hasta al quinto semestre. Luego, las relacionadas con el rendimiento en la PSU también son parte del grupo, pero solamente el promedio entre la prueba matemática y lenguaje y comunicación. Recordar que en los semestres anteriores, estas variables aparecían en el primer grupo de predictores.

Se mantienen los relacionados con las postergaciones como parte de las variables sin pesos, a excepción de las postergaciones acumuladas. Adicionalmente aparecen en este grupo el género del estudiante y si el colegio era régimen masculino.

CLUSTER 1			
Variables	**Peso**	**Abs(Peso)**	**Cuartil**
PostergaSem_Acum	1,916	1,916	Q1
%Reprobadas_Sem	0,43	0,43	Q1
Sem_Verano_Acum	0,279	0,279	Q1
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	-0,217	0,217	Q1
NotaSem_Anterior	0,189	0,189	Q1
Pref	0,189	0,189	Q1
GRUPO_FAMILIAR	-0,188	0,188	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	-0,135	0,135	Q1
EDUCACION_DELA_MADRE_EQUIVALENCIA	-0,131	0,131	Q1
NotaSem_Acum	0,124	0,124	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	-0,119	0,119	Q2
PROM_LM_FINAL	-0,111	0,111	Q2
EvalProf_Sem_Acum	-0,104	0,104	Q2
PTJE_NEM	0,101	0,101	Q2
REGIMEN = Femenino	-0,096	0,096	Q2
FINANCIAMIENTO	0,088	0,088	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	-0,086	0,086	Q2
EvalProf_Sem	-0,08	0,08	Q2

Tabla 32. Peso de las variables para el primer cluster del quinto semestre**.**

Variables	**Peso**	**Cuartil**	**Variables**
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	-0,067	0,067	Q3
%Reprobadas_Sem_Anterior	-0,066	0,066	Q3

INGRESO_BRUTO_FAMILIAR_GRUPO	0,062	0,062	Q3
GRUPO_DEPENDENCIA = Municipal	-0,059	0,059	Q3
NotaSem	-0,059	0,059	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	-0,054	0,054	Q3
EDUCACION_DEL_PADRE_EQUIVALENCIA	0,053	0,053	Q3
HORAS_QUE_DEDICA_TRABAJO	-0,045	0,045	Q3
%Reprobadas_Sem_Acum	-0,044	0,044	Q3
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	0,041	0,041	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0,04	0,04	Q4
EvalProf_Sem_Anterior	0,03	0,03	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	-0,029	0,029	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0,029	0,029	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	-0,024	0,024	Q4
GRUPO_DEPENDENCIA = Particular Pagado	0,008	0,008	Q4
Sem_Verano_Ant	0,001	0,001	Q4
PTJE_MATEMATICA_FINAL	-0,001	0,001	Q4
SEXO = MUJER	0	0	
REGIMEN = Masculino	0	0	
PostergaSem	0	0	
PostergaSem_Anterior	0	0	

Tabla 32. Peso de las variables para el primer cluster del quinto semestre.

Para el segundo grupo de clúster, el peso de cada variable se muestra en la **Tabla 33**. Al igual que en el primer clúster la gran mayoría de las variables en el primer grupo están relacionados con el rendimiento y ambiente universitario del estudiante. Cabe destacar la aparición en este grupo las variables relacionadas con el número del grupo familiar, ingreso bruto familiar y si el colegio del que egresó era de régimen femenino. Sin embargo, la mayoría de los predictores siguen siendo los relacionados con su ambiente y rendimiento académico universitario.

El segundo grupo de predictores se compone inicialmente por variables de evaluación del estudiante a los profesores en el semestre anterior, su rendimiento en la PSU de matemáticas y la cantidad de horas que trabajaba antes de ingresar a la universidad. Variables relacionados con el grupo familiar tales como cuantos estudian en cuarto medio y otras instituciones, cuantos trabajaban antes de ingresar del ingreso del

estudiante universidad y el nivel educacional del padre también forman parte de este grupo. Es importante este cambio, puesto que algunos de estos variables aparecían en el primer grupo de predictores.

Nuevamente las variables relacionadas con la postergación en el semestre y el anterior aparecen como significativas para el clúster con peso igual a cero.

CLUSTER 2			
Variables	**Peso**	**Abs(Peso)**	**Cuartil**
%Reprobadas_Sem	0,152	0,152	Q1
%Reprobadas_Sem_Acum	0,113	0,113	Q1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0,066	0,066	Q1
NotaSem_Anterior	-0,063	0,063	Q1
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0,046	0,046	Q1
%Reprobadas_Sem_Anterior	0,043	0,043	Q1
NotaSem	-0,035	0,035	Q1
INGRESO_BRUTO_FAMILIAR_GRUPO	-0,031	0,031	Q1
NotaSem_Acum	-0,027	0,027	Q1
REGIMEN = Femenino	-0,025	0,025	Q1
EvalProf_Sem_Anterior	0,022	0,022	Q2
PTJE_MATEMATICA_FINAL	-0,021	0,021	Q2
HORAS_QUE_DEDICA_TRABAJO	0,02	0,02	Q2
PostergaSem_Acum	-0,019	0,019	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	-0,019	0,019	Q2
GRUPO_DEPENDENCIA = Municipal	-0,018	0,018	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0,018	0,018	Q2
EDUCACION_DEL_PADRE_EQUIVALENCIA	0,018	0,018	Q2
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	0,017	0,017	Q2

Tabla 33. Peso de las variables para el segundo cluster del quinto semestre.

Variables	**Peso**	**Abs(Peso)**	**Cuartil**

Sem_Verano_Acum	-0,015	0,015	Q3
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	0,015	0,015	Q3
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	0,015	0,015	Q3
REGIMEN = Masculino	-0,013	0,013	Q3
GRUPO_DEPENDENCIA = Particular Pagado	0,013	0,013	Q3
SEXO = MUJER	0,012	0,012	Q3
EDUCACION_DELA_MADRE_EQUIVALENCIA	-0,012	0,012	Q3
PTJE_NEM	0,011	0,011	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	-0,01	0,01	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0,009	0,009	Q4
FINANCIAMIENTO	0,009	0,009	Q4
Sem_Verano_Ant	-0,007	0,007	Q4
GRUPO_FAMILIAR	0,006	0,006	Q4
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	0,006	0,006	Q4
EvalProf_Sem_Acum	0,005	0,005	Q4
Pref	-0,005	0,005	Q4
PROM_LM_FINAL	-0,003	0,003	Q4
EvalProf_Sem	0,002	0,002	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	-0,001	0,001	Q4
PostergaSem	0	0	
PostergaSem_Anterior	0	0	

Tabla 33 (Continuación). Peso de las variables para el segundo cluster del quinto semestre.

Semestre 6

Para el caso del sexto semestre el mejor modelo está compuesto nuevamente por LR, *clusterización* y aplicación de un umbral de clasificador. Sin embargo, se diferencia del semestre anterior en que esta vez la aplicación de la técnica de balance ROS es parte del mejor modelo.

En el primer grupo de clúster los pesos para cada variable mostrados en la **Tabla 34** muestran un cambio fuerte respecto al quinto semestre. Las variables que pertenecen al primer grupo de predictores están fuertemente relacionados con antecedentes preuniversitarios. Por ejemplo, las variables que indican el número de integrantes del grupo familiar, cuántos de ellos estudiaban en la básica o en otros niveles educacionales,

si el régimen del colegio en que estudiaron era masculino, cuántos del grupo familiar trabajaban al momento de ingresar a la universidad y su rendimiento en la PSU, específicamente en historia y lenguaje y comunicación, son todas variables que se obtienen antes de ingresar a la universidad. Solamente dos predictores del primer cuartil son variables capturados durante su avance académico del programa, que son el porcentaje de créditos reprobados acumulados al finalizar el semestre y el total de semestres postergados acumulados.

El segundo grupo de predictores se compone también principalmente por variables relacionados con antecedentes preuniversitarios. Ejemplo de algunos son el número de padres vivos, el rendimiento en la PSU, el ingreso bruto familiar y el número de integrantes del grupo familiar estudiando. Solamente tres de las variables se relacionan con características más actuales al semestre seis y que son el nivel de financiamiento, el número de veces que ha participado en los semestres de verano y el rendimiento académico obtenido en el semestre anterior.

Se mantienen las variables del semestre anterior dentro del grupo no predictores con peso cero. Adicionalmente aparecen en este grupo el variable que indica si la rama educacional del colegio era técnico y cuantos integrantes de la familia estudiaban en alguna institución de la educación superior.

CLUSTER 1			
Variables	**Peso**	**Abs(Peso)**	**Cuartil**
GRUPO_FAMILIAR	0,084	0,084	Q1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,074	0,074	Q1
REGIMEN = Masculino	0,067	0,067	Q1
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	-0,067	0,067	Q1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0,061	0,061	Q1
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	-0,06	0,06	Q1

Tabla 34. Peso de las variables para el primer clúster del sexto semestre.

Variables	Peso	Abs(Peso)	Cuartil
%Reprobadas_Sem_Acum	-0,059	0,059	Q1
PostergaSem_Acum	-0,059	0,059	Q1

CUANTOS_TRABAJAN_GRUPO_FAMILIAR	0,056	0,056	Q1
PTJE_MATEMATICA_FINAL	0,054	0,054	Q2
Sem_Verano_Acum	0,051	0,051	Q2
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	-0,05	0,05	Q2
FINANCIAMIENTO	-0,05	0,05	Q2
INGRESO_BRUTO_FAMILIAR_GRUPO	0,047	0,047	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	-0,046	0,046	Q2
NotaSem_Anterior	0,043	0,043	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0,042	0,042	Q2
GRUPO_DEPENDENCIA = Municipal	0,04	0,04	Q2
NotaSem	0,033	0,033	Q3
PTJE_NEM	0,033	0,033	Q3
EvalProf_Sem_Anterior	0,031	0,031	Q3
%Reprobadas_Sem_Anterior	-0,03	0,03	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	-0,03	0,03	Q3
REGIMEN = Femenino	0,027	0,027	Q3
GRUPO_DEPENDENCIA = Particular Pagado	-0,027	0,027	Q3
SEXO = MUJER	-0,026	0,026	Q3
PROM_LM_FINAL	-0,019	0,019	Q3
HORAS_QUE_DEDICA_TRABAJO	0,018	0,018	Q4
%Reprobadas_Sem	-0,014	0,014	Q4
NotaSem_Acum	0,012	0,012	Q4
EvalProf_Sem	0,011	0,011	Q4
EDUCACION_DEL_PADRE_EQUIVALENCIA	-0,01	0,01	Q4
EvalProf_Sem_Acum	0,007	0,007	Q4
Pref	-0,006	0,006	Q4
EDUCACION_DELA_MADRE_EQUIVALENCIA	0,004	0,004	Q4
Sem_Verano_Ant	-0,001	0,001	Q4
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0	0	
PostergaSem	0	0	
PostergaSem_Anterior	0	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	0	0	

Tabla 34 (Continuación). Peso de las variables para el primer cluster del sexto semestre.

En el caso del segundo clúster, el grupo de variables se repite a la situación presentada en el clúster anterior. Según **Tabla 35** Seis de los ocho variables son características definidas previas al ingreso, las cuales se relacionan fuertemente con el rendimiento en

la PSU, el grupo familiar y la rama del colegio donde estudió. Es interesante que como variable más importante sea el puntaje obtenido en la sección de historia o ciencias de la PSU. Por otro lado, el nivel educacional del padre nuevamente aparece dentro de este grupo, la cual estuvo casi siempre dentro de los dos primeros grupos de predictores en todos los semestres anteriores. Esto podría ayudar a la detección temprana de los estudiantes con alto potencial de renunciar en el largo plazo al programa.

Al igual que en el primer clúster, el segundo grupo de predictores se compone también principalmente por variables relacionados con características preuniversitarias. Dos de los primeros 4 variables se encuentran relacionados con el rendimiento en la PSU, lo que podría ayudar a identificar qué grupo de estudiantes desertaría en este semestre al inicio de la carrera universitaria del estudiante. Finalmente, el nivel educacional de la madre aparece en este grupo, mientras que el del padre aparece en el primer grupo.

La cantidad de variables con peso igual a cero es mucho mayor en este semestre, las cuales son nuevamente las relacionadas con la postergación, género, régimen del colegio, participación en semestre de verano y el número de padres vivos.

CLUSTER 2			
Variables	**Peso**	**Abs(Peso)**	**Cuartil**
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	-127,937	127,937	Q1
GRUPO_FAMILIAR	122,609	122,609	Q1
EDUCACION_DEL_PADRE_EQUIVALENCIA	-95,355	95,355	Q1
PTJE_NEM	60,96	60,96	Q1
%Reprobadas_Sem_Acum	51,22	51,22	Q1
%Reprobadas_Sem_Anterior	46,457	46,457	Q1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	41,721	41,721	Q1
RAMA_EDUCACIONAL_EQUIVALENCIA = T	-39,505	39,505	Q1

Tabla 35. Peso de las variables para el segundo cluster del sexto semestre.

Variables	**Peso**	**Abs(Peso)**	**Cuartil**
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	38,536	38,536	Q2
NotaSem_Anterior	-33,347	33,347	Q2

NotaSem_Acum	-29,696	29,696	Q2
%Reprobadas_Sem	28,867	28,867	Q2
EDUCACION_DELA_MADRE_EQUIVALENCIA	-26,727	26,727	Q2
GRUPO_DEPENDENCIA = Municipal	-20,209	20,209	Q2
EvalProf_Sem	-18,035	18,035	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	-16,865	16,865	Q2
Sem_Verano_Acum	-16,832	16,832	Q3
INGRESO_BRUTO_FAMILIAR_GRUPO	-12,897	12,897	Q3
EvalProf_Sem_Acum	-10,907	10,907	Q3
EvalProf_Sem_Anterior	10,892	10,892	Q3
FINANCIAMIENTO	-8,578	8,578	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	8,43	8,43	Q3
PROM_LM_FINAL	7,963	7,963	Q3
PTJE_MATEMATICA_FINAL	7,859	7,859	Q3
GRUPO_DEPENDENCIA = Particular Pagado	-7,619	7,619	Q4
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	-7,321	7,321	Q4
REGIMEN = Femenino	-5,994	5,994	Q4
PostergaSem_Acum	-5,637	5,637	Q4
NotaSem	-4,677	4,677	Q4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	-4,424	4,424	Q4
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	3,776	3,776	Q4
Pref	-3,55	3,55	Q4
SEXO = MUJER	0	0	
REGIMEN = Masculino	0	0	
PostergaSem	0	0	
PostergaSem_Anterior	0	0	
Sem_Verano_Ant	0	0	
HORAS_QUE_DEDICA_TRABAJO	0	0	
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0	0	
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	0	0	

Tabla 35 (Continuación). Peso de las variables para el segundo cluster del sexto semestre.

Para el tercer grupo de clúster, el peso de cada variable se muestra en la **Tabla 36**. En este clúster, al igual que en los anteriores, las variables relacionadas con características previas al ingreso aparecen nuevamente en el primer grupo de predictores, aunque en

este caso es más equiparado con la cantidad de variables de desempeño. Cuatro de los ocho variables son de antecedentes preuniversitarios, mientras que tres son de rendimiento universitario y una relacionada con la evaluación que realiza el estudiante a los profesores. En relación a las características previas al ingreso del estudiante a la universidad, dos de ellas son respecto a la configuración familiar, una respecto a su desempeño en la PSU sección ciencias/historia y otra de la rama educacional del estudiante. Eventualmente esto podría ayudar a la identificación temprana de la clase del estudiante una vez ingresado a la universidad.

Al igual que en el primer clúster y el primer grupo de predictores, el segundo grupo se compone también principalmente por variables relacionados con características preuniversitarias. La mayoría de estos variables se relacionan con la configuración familiar antes de ingresar a la universidad tales como, cuantos integrantes trabajaban, cuantos estudiaban en otros establecimientos educacionales y el nivel educacional de los padres. Los otros predictores de este grupo se relacionan con los antecedentes del colegio que egresó el estudiante.

Las variables con peso igual a cero son nuevamente los relacionados con la postergación y su participación en semestres de verano. Adicionalmente, variables respecto al género y régimen femenino del colegio aparecen como valor cero de peso. También lo es el ingreso bruto familiar del estudiante y prácticamente también el financiamiento, por lo que la decisión de desertar del programa de manera voluntaria en este clúster y semestre no es explicada por aspectos financieros.

CLUSTER 3			
Variables	**Peso**	**Abs(Peso)**	**Cuartil**
NotaSem_Acum	0,076	0,076	Q1
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	-0,064	0,064	Q1
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	0,059	0,059	Q1
EvalProf_Sem_Acum	-0,05	0,05	Q1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,05	0,05	Q1
NotaSem_Anterior	0,047	0,047	Q1
%Reprobadas_Sem_Acum	-0,045	0,045	Q1
RAMA_EDUCACIONAL_EQUIVALENCIA = T	-0,044	0,044	Q1
EvalProf_Sem_Anterior	-0,041	0,041	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	-0,041	0,041	Q2
EDUCACION_DEL_PADRE_EQUIVALENCIA	0,041	0,041	Q2
GRUPO_DEPENDENCIA = Particular Pagado	0,038	0,038	Q2
HORAS_QUE_DEDICA_TRABAJO	0,038	0,038	Q2
NotaSem	0,03	0,03	Q2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	0,03	0,03	Q2
EDUCACION_DELA_MADRE_EQUIVALENCIA	-0,026	0,026	Q2
PTJE_MATEMATICA_FINAL	0,026	0,026	Q2
GRUPO_FAMILIAR	-0,023	0,023	Q3
Sem_Verano_Acum	0,02	0,02	Q3
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	0,019	0,019	Q3
PROM_LM_FINAL	0,018	0,018	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0,017	0,017	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	-0,013	0,013	Q3
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	-0,013	0,013	Q3
REGIMEN = Masculino	0,012	0,012	Q4
PTJE_NEM	0,012	0,012	Q4
GRUPO_DEPENDENCIA = Municipal	-0,011	0,011	Q4
%Reprobadas_Sem	0,01	0,01	Q4
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	0,009	0,009	Q4
Pref	-0,008	0,008	Q4
PostergaSem_Acum	0,004	0,004	Q4
EvalProf_Sem	-0,003	0,003	Q4
%Reprobadas_Sem_Anterior	-0,002	0,002	Q4
FINANCIAMIENTO	0,002	0,002	Q4

Tabla 36. Peso de las variables para el tercer cluster del sexto semestre.

Variables	Peso	Abs(Peso)	Cuartil
SEXO = MUJER	0	0	
REGIMEN = Femenino	0	0	
PostergaSem	0	0	
PostergaSem_Anterior	0	0	
Sem_Verano_Ant	0	0	
INGRESO_BRUTO_FAMILIAR_GRUPO	0	0	

Tabla 36 (Continuación). Peso de las variables para el tercer cluster del sexto semestre.

La obtención de los predictores por semestre permite analizar la diferencia que se genera entre semestre. Es decir, mientras para el primer y tercer año las variables relacionados con las características preuniversitarias del estudiante son considerados como predictores importantes, para el segundo año los relacionados con el rendimiento universitario y participación en semestres de verano son los más importantes. Esto se puede deber principalmente a que en el segundo año existe la transferencia interna, donde la facultad permite a los estudiantes transferirse de un programa a otro. Esta transferencia pueden realizarla solamente algunos alumnos, principalmente aquellos que obtienen buenos rendimientos académicos.

5.2 Perfiles por Semestre

En esta sección se identificarán los perfiles de cada una de las clases estudiadas. El perfil se generará a través del promedio de cada predictor sobre la clase de los registros que el modelo con mejor desempeño.

Semestre 1

La **Tabla 37** muestra el valor para cada predictor de los primeros grupos del primer clúster. En general podemos describir la clase DESERTA como alumnas todas mujeres que realizan evaluación a sus profesores en promedio altas, pero que en tienen un peor rendimiento en la PSU que la clase NO DESERTA. También tienen un peor rendimiento académico en el semestre, pero son estudiantes con buen NEM (Promedio de Notas

Enseñanza Media). Eventualmente la identificación de estos estudiantes al inicio del semestre y el posterior apoyo académico en sus estudios podría disminuir la probabilidad de que finalmente deserte.

Cluster 1		
Predictor	**DESERTA**	**NO DESERTA**
SEXO = MUJER	1,00	1,00
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	2,00	1,89
EvalProf	6,46	5,96
PROM_LM_FINAL	6.342,50	6.804,18
PTJE_MATEMATICA_FINAL	647,50	696,11
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	621,00	665,13
PTJE_NEM	754,00	696,13
FINANCIAMIENTO	3.393.750,00	2.920.679,24
EDUCACION_DELA_MADRE_EQUIVALENCIA	13,00	12,46
NotaSem	4,59	5,06
EDUCACION_DEL_PADRE_EQUIVALENCIA	12,00	12,45

Tabla 37 Perfiles primer clúster para el primer semestre.

En el caso del segundo clúster, podemos identificar que respecto al desempeño en la PSU ocurre lo contrario con el clúster anterior. Según **Tabla 38**, en este grupo los estudiantes de la clase DESERTA tienen un mejor desempeño en la PSU que sus pares de la clase NO DESERTA. Adicionalmente, el nivel de ingresos familiares brutos son altos y reciben más financiamiento que los NO DESERTA. Finalmente, al igual que el **Cluster 1**, el rendimiento académico es peor en los estudiantes desertores que los no desertores.

Cluster 2		
Predictor	**DESERTA**	**NO DESERTA**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	1,98	1,91
PROM_LM_FINAL	6.968,14	6.721,26
EvalProf	5,84	5,97
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	685,12	649,89
EDUCACION_DEL_PADRE_EQUIVALENCIA	14,07	14,44
PTJE_MATEMATICA_FINAL	710,81	701,96
EDUCACION_DELA_MADRE_EQUIVALENCIA	13,74	14,19
PTJE_NEM	658,65	670,42
INGRESO_BRUTO_FAMILIAR_GRUPO	8,09	5,97
NotaSem	4,45	5,08
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0,86	0,63
FINANCIAMIENTO	1.758.668,30	1.136.260,84
GRUPO_FAMILIAR	4,72	4,46

Tabla 38 Perfiles segundo clúster para el primer semestre.

Dada las descripciones de los perfiles para los dos grupos, se puede concluir que el problema mayor de los estudiantes se relaciona con el rendimiento académico preuniversitario y universitario. Existen dos grupos, siendo uno de ellos identificado como estudiantes que tienen buenas notas en el colegio, evalúan bien a sus profesores, pero que obtienen un bajo rendimiento en la PSU. El otro grupo son los que tienen altos rendimientos en la PSU, vienen de familias con ingresos altos y que la mayoría tiene al menos un familiar estudiando en cuarto medio al ingresar a la universidad. Eventualmente programas de apoyo académico a estos tipos de alumnos podría ayudar a disminuir la deserción voluntaria.

Semestre 2

Para el segundo semestre se identificaron tres clúster, por lo que se discutirán en esta sección 6 perfiles en total, uno por clase para cada clúster.

La **Tabla 39** muestra que en promedio los estudiantes de la clase DESERTA tienen una mejor percepción de los profesores que los NO DESERTA. Adicionalmente, reciben mayor financiamiento y en su mayoría son mujeres. Si bien el rendimiento obtenido en la PSU es bastante similar en las dos clases, la diferencia se genera en el porcentaje de créditos reprobados acumulados al finalizar el segundo semestre, el cual alcanza un 33% en promedio para los DESERTA y 10% en promedio para los NO DESERTA.

Respecto a la configuración familiar, los padres para ambas clases tienen un nivel educacional muy parecido, pero se diferencian fuertemente en el de la madre, el cual alcanza en promedio 8 años de estudios, lo que significa que las madres de este grupo llegan hasta la educación básica completa.

Se concluye que este clúster está compuesto por estudiantes con buen rendimiento en la PSU y puntaje NEM, con alto nivel de financiamiento y sus padres con relativo bajo nivel de educación (no más que educación media completa). Sin embargo, las clases se diferencian principalmente en que los DESERTA realizan evaluaciones a los profesores más alta, el porcentaje de créditos reprobados acumulados es del 33% en promedio y la educación de la madre ronda la básica completa; mientras que los estudiantes de la clase NO DESERTA evalúan peor a los profesores, tienen tasas de reprobación acumulada en un promedio de 10% y la educación de la madre alcanza la media completa.

Cluster 1		
Predictor	**DESERTA**	**NO DESERTA**
EvalProf_Sem_Anterior	6,43	6,12
EvalProf_Sem	6,23	5,78
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	697,33	666,85
PROM_LM_FINAL	6.956,67	6.815,42
PTJE_MATEMATICA_FINAL	694,00	697,61
EvalProf_Sem_Acum	6,33	5,99
FINANCIAMIENTO	3.580.833,33	3.318.197,27
PTJE_NEM	702,33	682,51
SEXO = MUJER	0,67	0,31
EDUCACION_DEL_PADRE_EQUIVALENCIA	10,33	10,88
%Reprobadas_Sem_Acum	0,33	0,10
EDUCACION_DELA_MADRE_EQUIVALENCIA	8,67	11,12

Tabla 39 Perfiles primer clúster para el segundo semestre.

En la **Tabla 40** se muestran los perfiles para cada clase del segundo clúster. En él se puede ver que en general el clúster está compuesto por estudiantes con buen rendimiento en la sección de matemática de la PSU, con evaluación a los profesores rondando el 5,8 y en promedio la mitad del grupo se graduó de un colegio municipal. Las principales diferencias entre ambas clases son detectadas en los predictores del rendimiento de PSU sección Lenguaje y Comunicación, porcentaje de créditos reprobados en el semestre y acumulado, nivel educacional de los padres y el género.

Los estudiantes que desertan finalizado el segundo semestre tienen mayores créditos reprobados tanto en el semestre como en el acumulado. Si bien, el clúster se compone por estudiantes con alto porcentaje de créditos reprobados (sobre el 20%), los estudiantes desertores son los que tienen en promedio mayores créditos no aprobados.

Adicionalmente, los estudiantes que desertan tienen mejor rendimiento en la PSU, principalmente en la sección de Lenguaje y Comunicación. Otra gran diferencia se genera en el género, en que al menos la mitad de los DESERTORES son hombres, mientras que en los NO DESERTORES no supera el 15%.

Finalmente, el nivel educacional de los padres nuevamente es una notoria diferencia entre las dos clases. Para los desertores sus padres tienen en promedio mayor nivel educacional, mientras que en los no desertores sus madres tienen mayor nivel educacional.

Cluster 2		
Predictor	**DESERTA**	**NO DESERTA**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	2,00	1,92
%Reprobadas_Sem_Acum	0,56	0,27
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	755,50	665,27
PROM_LM_FINAL	7.330,00	6.842,31
EvalProf_Sem	5,90	5,72
%Reprobadas_Sem	0,85	0,39
PTJE_MATEMATICA_FINAL	710,50	703,19
EDUCACION_DEL_PADRE_EQUIVALENCIA	13,50	11,73
EvalProf_Sem_Anterior	5,61	5,71
EvalProf_Sem_Acum	5,76	5,91
SEXO = MUJER	0,50	0,88
REGIMEN = Femenino	0,50	0,85
GRUPO_DEPENDENCIA = Municipal	0,50	0,54
EDUCACION_DELA_MADRE_EQUIVALENCIA	9,00	11,65

Tabla 40 Perfiles segundo clúster para el segundo semestre.

Finalmente, en la **Tabla 41** se muestran los perfiles del tercer clúster. En general este clúster está compuesto por estudiantes que sus padres tienen niveles de educación altos (al menos unos dos años de estudios post educación media), con desempeño en la PSU relativamente bajo en lenguaje y comunicación y con rendimiento académico rondando entre el 4,5 y 5,0.

Las diferencias entre las clases DESERTA y NO DESERTA se encuentran claramente en el nivel educacional de sus padres, donde la cantidad de año de estudios de los padres y madres tiende a ser más alta en promedio para los desertores. Ocurre lo mismo en el rendimiento de la sección de Lenguaje y Comunicación de la PSU. En el ingreso los desertores se encuentran ubicados al menos 4 grupos más arriba que los no

desertores y la mayoría de los estudiantes clase DESERTA estudió en un colegio particular pagado, cuya distribución alcanza el 81% en promedio.

En resumen, estudiantes que tienen un rendimiento académico entre el 4,5 y 5,0, con nivel educacional de los padres altos, altos ingresos brutos familiares y graduados de colegios particulares pagados son potenciales desertores para este clúster.

Cluster 3		
Predictor	**DESERTA**	**NO DESERTA**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	2,00	1,92
EDUCACION_DEL_PADRE_EQUIVALENCIA	16,03	14,27
EDUCACION_DELA_MADRE_EQUIVALENCIA	15,16	14,11
EvalProf_Sem_Anterior	6,01	5,96
PROM_LM_FINAL	6.936,72	6.719,36
EvalProf_Sem	5,89	5,96
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	685,06	648,96
PTJE_MATEMATICA_FINAL	702,28	702,02
EvalProf_Sem_Acum	6,01	5,98
INGRESO_BRUTO_FAMILIAR_GRUPO	9,83	5,62
GRUPO_DEPENDENCIA = Particular Pagado	0,81	0,29
PTJE_NEM	672,19	674,53
NotaSem	4,51	4,64
NotaSem_Anterior	5,05	5,04
NotaSem_Acum	4,79	4,84
GRUPO_FAMILIAR	5,28	4,47

Tabla 41 Perfiles segundo clúster para el segundo semestre.

Semestre 3

Para el tercer semestre el mejor modelo no está compuesto por la técnica de culsterización, por lo que solamente existe un perfil por cada clase.

Según **Tabla 42,** la mayoría de las diferencias entre las clases se genera en la dependencia del colegio en que egresaron los estudiantes, el ingreso bruto familiar, cuantos integrantes estudiaban en cuarto medio antes de ingresar a la universidad y el género.

Es bastante notorio que al menos la mitad de los estudiantes que desertan egresaron de colegios particulares pagados, lo que se relaciona con el nivel de ingreso bruto familiar que es en promedio dos niveles más alto que el de los no desertores. Adicionalmente, los estudiantes que permanecen en la carrera son casi la mitad del género femenino y existe una pequeña diferencia en la educación de los padres, quienes en el caso de los desertores tienden a ser mayores que el de los NO DESERTA.

En general, para el tercer semestre los predictores principales pertenecen a variables relacionados con la configuración familiar o variables preuniversitarias.

Predictor	DESERTA	NO DESERTA
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	1,97	1,90
EDUCACION_DEL_PADRE_EQUIVALENCIA	14,81	13,75
EvalProf_Sem_Anterior	6,06	5,95
EDUCACION_DELA_MADRE_EQUIVALENCIA	14,68	13,54
EvalProf_Sem	5,95	5,89
PROM_LM_FINAL	6.760,82	6.752,20
PTJE_MATEMATICA_FINAL	698,82	702,49
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	653,35	654,16
NotaSem_Acum	5,06	4,80
EvalProf_Sem_Acum	6,01	5,97
NotaSem	4,77	4,78
NotaSem_Anterior	4,95	4,62
PTJE_NEM	663,18	677,01
INGRESO_BRUTO_FAMILIAR_GRUPO	7,20	5,22
GRUPO_DEPENDENCIA = Particular Pagado	0,52	0,26
SEXO = MUJER	0,39	0,45
GRUPO_FAMILIAR	4,63	4,46
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	0,74	0,61

Tabla 42 Perfiles para el tercer semestre.

Semestre 4

De acuerdo a la información mostrada en la **Tabla 43**, el primer clúster del semestre cuatro está compuesto por estudiantes con padres que tienen al menos la educación media completa, con evaluación a los profesores rondando el 5.8, con desempeño académico entre el 4.6 y 5.0, y con desempeño en la PSU relativamente alto en la sección de matemáticas.

Las principales diferencias entre ambas clases para este clúster se generan fuertemente en la dependencia del colegio que egresaron, el género y el nivel de ingreso bruto familiar. En el caso de los desertores es muy marcado que la gran mayoría proviene de colegios particulares pagados (el 90%), mientras que en los no desertores es solo al 20%. Adicionalmente, el grupo de los estudiantes que deserta está compuesto principalmente por hombres, cuya distribución alcanza un poco más del 85%. En el caso del ingreso bruto familiar, los estudiantes que no desertan están en promedio en el cuarto

grupo más bajo de ingresos familiares, mientras que los desertores lo hacen en el octavo. Finalmente, el nivel educacional de los padres nuevamente marca una gran diferencia, ya que en promedio los padres y madres tienen 16 y 15 años de estudios respectivamente para los desertores, mientras que este baja a 13 años aproximadamente en el caso de los no desertores.

En resumen, al igual que en el semestre anterior, los principales predictores con grandes diferencias están fuertemente relacionados con variables definidos previamente al ingreso del estudiante a la universidad. Esto podría servir para identificar desde el semestre 1 quienes serían los estudiantes a desertar en los próximos semestres.

Cluster 1		
Predictor	**DESERTA**	**NO DESERTA**
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	1,86	1,92
EDUCACION_DEL_PADRE_EQUIVALENCIA	16,00	13,72
EvalProf_Sem_Anterior	5,87	5,91
PROM_LM_FINAL	6.849,29	6.790,52
EDUCACION_DELA_MADRE_EQUIVALENCIA	15,24	13,42
EvalProf_Sem	5,80	5,84
PTJE_MATEMATICA_FINAL	707,90	701,84
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	661,95	656,57
NotaSem	4,70	5,04
NotaSem_Anterior	4,69	5,07
NotaSem_Acum	4,73	5,02
EvalProf_Sem_Acum	5,77	5,95
PTJE_NEM	662,62	690,63
INGRESO_BRUTO_FAMILIAR_GRUPO	7,81	4,68
GRUPO_DEPENDENCIA = Particular Pagado	0,95	0,20
SEXO = MUJER	0,14	0,50
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	1,05	0,74
GRUPO_FAMILIAR	5,00	4,41

Tabla 43 Perfiles primer clúster del cuarto semestre.

En la **Tabla 44** se pueden ver los perfiles para ambas clases, donde notoriamente el grupo está compuesto principalmente por estudiantes con peor rendimiento académico

que el primer clúster, con ingresos brutos familiares dentro del quinto y sexto grupo más bajo de ingresos, con número de integrantes entre 4 y 5, y educación de los padres rondando los 13 y 14 años.

Aunque no existen mayores diferencias en el valor promedio de los predictores para cada clase, las mayores se generan en la participación en el semestre verano, la distribución de estudiantes egresados de colegios particulares pagados, ingreso bruto familiar y el desempeño académico obtenido en el semestre. La diferencia más destacable se genera en este último predictor, donde los rendimientos en promedio están en bajo o igual a 4.0 para el caso de los desertores. Esto lo hace más interesante al mirar el rendimiento PSU y la evaluación que realizan tales estudiantes a los profesores en el semestre, ya que en promedio estos evalúan con peores notas a los profesores y tienen un rendimiento mayor en la PSU. Por lo tanto, se podría suponer un descontento por parte de los estudiantes, puesto que la evaluación de los profesores en el semestre anterior en promedio es de 6.1 y baja drásticamente a 5.4 en el semestre actual. Se concluye entonces que este hecho afectaría directamente su rendimiento académico y finalmente la decisión de abandonar el programa.

En resumen, este clúster se caracteriza principalmente por estudiantes con bajo rendimiento académico e ingresos familiares entre el quinto y sexto grupo más bajo. Adicionalmente, dentro de este grupo los desertores son aquellos que tienen peores notas y baja evaluación a los profesores. Esto podría reflejar un cambio en los tipos de predictores que comienzan a tomar mayor importancia para predecir el comportamiento de deserción de los estudiantes.

Cluster 2		
Predictor	**DESERTA**	**NO DESERTA**

¿VIVEN_SUS_PADRES?_CUANTOSPADRES	1,88	1,91
EvalProf_Sem_Anterior	6,10	5,90
EDUCACION_DELA_MADRE_EQUIVALENCIA	13,88	14,32
EDUCACION_DEL_PADRE_EQUIVALENCIA	13,91	14,34
EvalProf_Sem	5,45	5,83
PTJE_MATEMATICA_FINAL	711,07	700,02
PROM_LM_FINAL	6.809,19	6.669,98
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	652,09	648,53
EvalProf_Sem_Acum	5,99	5,94
NotaSem	3,57	4,54
NotaSem_Acum	4,07	4,64
NotaSem_Anterior	4,00	4,57
PTJE_NEM	661,47	657,62
INGRESO_BRUTO_FAMILIAR_GRUPO	5,00	6,62
GRUPO_FAMILIAR	3,98	4,58
GRUPO_DEPENDENCIA = Particular Pagado	0,28	0,38
Sem_Verano_Ant	0,26	0,42
FINANCIAMIENTO	1.957.851,88	1.616.334,34

Tabla 44 Perfiles segundo clúster del cuarto semestre.

Semestre 5

En este semestre el mejor modelo utiliza la técnica *clusterización*, identificando dos clúster para el semestre. Por lo tanto, en esta sección se mostrarán 4 perfiles finales, los que corresponden a dos por clúster.

Para el primer clúster del semestre los valores en promedio de cada predictor se muestran en la **Tabla 45**. En ella se puede ver que el clúster está compuesto principalmente por estudiantes con rendimiento del semestre anterior entre un 5.0 y 5.3 aproximadamente, con bajo rendimiento en la PSU, especialmente en las de ciencias/historia y lenguaje y comunicación, bajo nivel de financiamiento y con un 5.5 en promedio de evaluación al profesor.

Las principales diferencias se identifican en el porcentaje de créditos reprobados en el semestre, donde los estudiantes desertores alcanzan el 44% en promedio, muy lejos del

7% de los estudiantes que no desertan. En segundo lugar está el nivel de financiamiento, en donde los que deciden dejar voluntariamente la carrera reciben un poco más de 2 veces el financiamiento de los que se quedan. Llama la atención también la preferencia con que postularon al programa, donde en los estudiantes desertores tiende a ser menor de los que no desertan (la carrera más preferida es la con preferencia igual a 1). Esto se condice con las postergaciones acumuladas hasta este semestre, ya que mientras los estudiantes que se mantienen en la carrera al finalizar el quinto semestre no han postergado ningún semestre, los estudiantes que dejan la carrera voluntariamente sí lo han hecho una vez en promedio.

En resumen, los estudiantes de este clúster son en general aquellos con bajo nivel de financiamiento, seleccionados generalmente en su segunda o tercera preferencia y con bajo nivel de reprobación, a excepción de los estudiantes que desertan que alcanzan el 44%. Lo interesante es que los tres predictores más importantes son variables relacionados con información que se captura después del ingreso del estudiante a la universidad, principalmente el rendimiento académico universitario.

Cluster 1		
Predictor	**DESERTA**	**NO DESERTA**
PostergaSem_Acum	1,00	0,00
%Reprobadas_Sem	0,44	0,07
Sem_Verano_Acum	1,50	1,14
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	595,50	655,98
NotaSem_Anterior	5,26	4,95
Pref	2,50	1,79
GRUPO_FAMILIAR	4,00	4,75
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	628,00	660,52
EDUCACION_DELA_MADRE_EQUIVALENCIA	13,00	14,62
NotaSem_Acum	5,01	4,91
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0,00	0,45
PROM_LM_FINAL	6.655,00	6.805,29
EvalProf_Sem_Acum	5,83	5,97
PTJE_NEM	733,50	703,61
REGIMEN = Femenino	0,00	0,24
FINANCIAMIENTO	812.250,00	267.498,96
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,00	0,18
EvalProf_Sem	5,47	5,75

Tabla 45. Perfiles primer clúster del quinto semestre.

En el caso del segundo clúster, los integrantes tienen en general sobre el 10% de los créditos reprobados tanto en el semestre, siendo mayor para los estudiantes DESERTORES (59%). El ingreso familiar se encuentra entre el cuarto y quinto nivel más bajo, mientras que la educación del padre es al menos educación media completa y el rendimiento académico se encuentra bajo el 5,0 en el semestre. Adicionalmente, los estudiantes de este clúster dedicaban en promedio algunas horas semanales al trabajo antes de ingresar a la universidad.

Tal como se puede ver en la **Tabla 46**, los estudiantes desertores tienen marcadas diferencias con los que no. Las reprobaciones son mucho más altas, el rendimiento del semestre anterior es bastante más bajo (cercano al 4.0), la clase acumula más

estudiantes que egresaron de un colegio técnico (el 33% versus el 8%) y antes de ingresar a la universidad dedicaban mayores horas de trabajo.

Tales diferencias descritas en el párrafo anterior podrían indicar la importancia de la aplicación de algún apoyo académico en los estudiantes que cumplen las características del clúster, con el objetivo de disminuir la cantidad de créditos reprobados y así evitar la decisión final de desertar.

Cluster 2		
Predictor	**DESERTA**	**NO DESERTA**
%Reprobadas_Sem	0,59	0,12
%Reprobadas_Sem_Acum	0,33	0,11
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0,67	0,33
NotaSem_Anterior	4,11	4,82
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0,33	0,08
%Reprobadas_Sem_Anterior	0,33	0,12
NotaSem	3,28	4,83
INGRESO_BRUTO_FAMILIAR_GRUPO	4,33	4,47
NotaSem_Acum	4,16	4,87
REGIMEN = Femenino	0,00	0,15
EvalProf_Sem_Anterior	4,20	5,88
PTJE_MATEMATICA_FINAL	675,33	710,46
HORAS_QUE_DEDICA_TRABAJO	2,67	1,33
PostergaSem_Acum	0,00	0,02
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,00	0,14
GRUPO_DEPENDENCIA = Municipal	0,00	0,30
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	1,00	0,61
EDUCACION_DEL_PADRE_EQUIVALENCIA	14,67	13,37
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	1,67	1,28

Tabla 46 Perfiles segundo clúster del quinto semestre.

Semestre 6

En este semestre el mejor modelo también utiliza *clusterización*, por lo que fueron identificados 3 clúster, generando un total de 6 perfiles en total.

Los perfiles del primer semestre se pueden ver en la **Tabla 47**, en donde la mayoría de los estudiantes que componen este grupo tienen un rendimiento relativamente normal en la PSU (entre 670 a 700 puntos) a excepción de la sección de Lenguaje y Comunicación, el cual es el más bajo que las otras secciones. También en general ambos padres están vivos, el desempeño académico del semestre anterior es bajo del 5,0 y tienen bajo nivel de financiamiento.

Las mayores diferencias entre clases para este clúster se generan en la cantidad de postergaciones acumuladas de los estudiantes. El número indica que los estudiantes desertores tienden a desertar más que los no desertores (465% más). También existe una gran diferencia en el financiamiento, donde si bien en el clúster el nivel de financiamiento es bajo, para los estudiantes que desertan en es mucho más alto en promedio que los que se mantienen en el programa. Es importante destacar la diferencia que se genera en la configuración familiar, particularmente en el predictor que indica el número de integrantes de la familia que estudiaban en la básica y en la media al momento del ingreso del estudiante a la universidad. En el caso de los desertores, en promedio ningún integrante lo hacía, sin embargo, para los no desertores existen casos en que algunos sí lo hacían.

En resumen los estudiantes de este clúster tienden a desertar principalmente por las postergaciones y la cantidad de créditos reprobados. Esto podría aumentar el efecto por el hecho de que dos años anteriores, la cantidad de integrantes estudiando en la básica y media es más alto que los no desertores.

Cluster 1		
Predictor	**DESERTA**	**NO DESERTA**
GRUPO_FAMILIAR	3,40	4,62
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,00	0,20
REGIMEN = Masculino	0,00	0,18
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	699,00	657,55
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	0,00	0,37
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	705,60	664,60
%Reprobadas_Sem_Acum	0,23	0,12
PostergaSem_Acum	0,20	0,04
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	1,00	1,47
PTJE_MATEMATICA_FINAL	716,60	709,63
Sem_Verano_Acum	1,00	1,13
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	2,00	1,96
FINANCIAMIENTO	1.248.774,40	458.965,96
INGRESO_BRUTO_FAMILIAR_GRUPO	5,60	6,92
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	0,60	0,28
NotaSem_Anterior	4,26	4,79
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	0,00	0,04
GRUPO_DEPENDENCIA = Municipal	0,00	0,15

Tabla 47. Perfiles primer clúster del sexto semestre.

Para el caso del segundo clúster, los estudiantes que lo componen se caracterizan por tener un desempeño académico no muy alto tanto en el semestre anterior como en el acumulado, el nivel educacional de la madre en promedio es cercano a la educación media completa, la evaluación a los profesores está sobre el 5,1 de promedio y el puntaje NEM se encuentra sobre los 700 puntos.

Las diferencias entre clases se puede ver fácilmente en la **Tabla 48**, donde los desertores tienden a tener mayores integrantes estudiando en básica o media de primero a tercero al momento de ingresar a la universidad. Ocurre lo contrario respecto al número de integrantes estudiando en la educación superior, donde los no desertores tienden a tener más integrantes en la educación superior previo ingreso al programa. Es importante destacar la gran diferencia que existe en el nivel educacional del padre, donde los desertores tienden a tener padres con niveles muchos más bajos,

principalmente básica incompleta. Finalmente, otra gran diferencia se genera en el ingreso bruto familiar, donde los desertores se encuentran principalmente en el séptimo grupo más bajo de ingresos y los no desertores en el cuarto.

En resumen, al igual que el semestre anterior, la deserción de los estudiantes se puede identificar principalmente por la cantidad de créditos reprobados. Esto podría aumentar el efecto por el hecho de que dos años anteriores, la cantidad de integrantes estudiando en la básica y media es más alto que los no desertores, pero no ocurre así en el caso de integrantes estudiando en la educación superior al ingresar a la universidad. Al parecer, el efecto de tener integrantes en la misma situación estudiantil genera un efecto positivo para finalmente no desertar.

Cluster 2		
Predictor	**DESERTA**	**NO DESERTA**
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	531,00	659,46
GRUPO_FAMILIAR	7,00	4,11
EDUCACION_DEL_PADRE_EQUIVALENCIA	4,00	11,75
PTJE_NEM	764,00	703,49
%Reprobadas_Sem_Acum	0,22	0,10
%Reprobadas_Sem_Anterior	0,19	0,05
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	1,00	0,31
RAMA_EDUCACIONAL_EQUIVALENCIA = T	0,00	0,13
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	1,00	0,25
NotaSem_Anterior	4,67	5,10
NotaSem_Acum	4,49	4,98
%Reprobadas_Sem	0,25	0,11
EDUCACION_DELA_MADRE_EQUIVALENCIA	12,00	12,62
GRUPO_DEPENDENCIA = Municipal	0,00	0,41
EvalProf_Sem	5,21	5,72
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	0,00	0,44

Tabla 48 Perfiles segundo clúster del sexto semestre.

Finalmente el tecer clúster de este semestre está compuesto principalmente por estudiantes con bajo rendimiento acádemico (bajo el 5,0), con al menos 10% de reprobación promedio de los créditos cursados hasta el momento y con un nivel educacional de los padres de media incompleto o completa.

Las princpales diferencias entre las clases, tal como se muestra en la **Tabla 49**, se ve principalmente en el rendimiento académico acumulado, el puntaje obtenido en la sección de ciencia/historia de la PSU, la evaluación al profesor en el semestre anterior, las horas que dedicaba al trabajo antes de comenzar estudios universitarios y cuantos integrantes de la familia estudiaban en la educación superior antes de ingresar a la universidad. De todos estos variablees llama la atención nuevamente el último, puesto que al igual que en el segundo clúster, el tener un integrante en la familia en la educación superior podría generar un efecto positivo al momento de decidir si desertar o no. Nuevamente el desempeño académico es más bajo para los desertores, aún cuando la evaluación a los profesores es mayor en estos tipos de estudiantes.

En resumen, el bajo rendimiento de los estudiantes desertores podría estar reflejado por la configuración familiar de los estudiantes, ya que si bien evalúan en promedio mejor a los profesores, en general la cantidad de integrantes estudiando en la educación superior tiende a ser menor. Adicionalmente, el nivel educacional del padre es mucho mayor para los no desertores. Es importante también destacar que todos los estudiantes en la categoría desertora provienen de un colegio técnico, por lo tanto, estudiantes de estas características no tienen otros integrantes de la familia estudiando en la educación superior y el nivel educacional del padre tiende a ser más bajo en este clúster. En conclusión, se podría implementar un programa de apoyo académico el cual supla la ausencia de algún integrante familiar en la educación superior. De esta manera se podría aumentar el desempeño académico y evitar la deserción voluntaria del estudiante.

Cluster 3		
Predictor	**DESERTA**	**NO DESERTA**
NotaSem_Acum	4,31	4,93
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	790,00	673,54
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	1,00	1,15
EvalProf_Sem_Acum	6,41	5,93
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	0,00	0,07
NotaSem_Anterior	4,33	4,89
%Reprobadas_Sem_Acum	0,26	0,10
RAMA_EDUCACIONAL_EQUIVALENCIA = T	1,00	0,15
EvalProf_Sem_Anterior	6,60	5,77
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	1,00	0,37
EDUCACION_DEL_PADRE_EQUIVALENCIA	10,00	12,43
GRUPO_DEPENDENCIA = Particular Pagado	0,00	0,08
HORAS_QUE_DEDICA_TRABAJO	0,00	2,83
NotaSem	4,46	4,88
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	0,00	0,62
EDUCACION_DELA_MADRE_EQUIVALENCIA	12,00	11,79
PTJE_MATEMATICA_FINAL	674,00	707,42

Tabla 49 Perfiles tercer clúster del sexto semestre.

Análisis Transversal

Según lo mostrado anteriormente, variables relacionadas con el rendimiento académico y configuración familiar se repiten en todos los semestre como predictores de las deserciones y retenciones.

Analizando el uso de las variables en cada modelo generado por semestre, es posible identificar que las variables relacionadas con la PSU, desempeño académico universitario, evaluación a los profesores, desempeño académico preuniversitario, nivel educacional de los padres, número de integrantes que componen el grupo familiar, cantidad de integrantes de la familia trabajando, preferencia de ingreso, participación en semestres de verano y financiamiento son importantes dado su peso obtenido en los modelos.

En el caso de evaluar solamente aquellas variables que aparecen más veces dentro del primer cuartil de variables con mayores pesos, aparecen como principales los relacionados con la PSU, específicamente el puntaje en la prueba de Lenguaje y Comunicación, seguido por el nivel educacional de los padres y la evaluación a los profesores. Por otro lado, si extendemos el mismo análisis tomando en cuenta los dos primeros cuartiles, nuevamente el rendimiento en la PSU, principalmente en la prueba de Lenguaje y Comunicación y Matemáticas, aparece dentro de los primeros grupos de variables importantes. Junto a estos dos variables, se encuentran el nivel educacional de los padres y posteriormente, los relacionados con el desempeño académico universitario del semestre anterior, el número de padres vivos, desempeño preuniversitario, evaluación al profesor, desempeño académico universitario acumulado, número integrantes del grupo familiar, ingreso bruto familiar y nivel de financiamiento.

Los perfiles mostrados entregan una primera idea de los predictores más importante en donde se generan grandes diferencias entre clases. Analizando la lista de todos los predictores de cada semestre se puede constatar que aquellos relacionados con la configuración familiar y el rendimiento durante su transcurso por la carrera son los que más se repiten durante todos los semestres, lo que se condice con los modelos teóricos de Spady (1970a), Tinto y Cullen (1975) y Bean (1980), los cuales muestran también que los antecedentes familiares y el desempeño académico del estudiante son variables primordiales para explicar la deserción del estudiante.

En resumen, las variables consideradas importantes para todos los semestres son las relacionados con el puntaje PSU del alumno, seguido del nivel educacional de los padres y el desempeño académico universitario. En un segundo nivel de importancia destaca el financiamiento y la configuración familiar del estudiante.

A modo de resumen, en el **Anexo 5 – Resumen Variables por Semestre** se muestran por semestre las variables identificadas como predictores más importantes de acuerdo a los mejores modelos.

CAPÍTULO 6 – DISCUSIÓN Y CONCLUSIONES

Como primer resultado, y respondiendo a los objetivos planteados al inicio de este documento, se pudo identificar los predictores más importantes para cada semestre. Cabe destacar que entre los más importantes para todos los semestres se encuentran el rendimiento en la PSU, el número de padres vivos, la evaluación de los alumnos que realizaban a los profesores de manera semestral y el rendimiento académico universitario. Sin embargo, el conjunto total de predictores era distinto para todos los semestres, pudiéndose identificar una tendencia mientras se avanzaba en los semestres

Para los primeros cuatro semestres, predictores relacionados con características preuniversitarias del estudiante fueron identificadas dentro de los primeros dos cuartiles de predictores. En específico, tales predictores eran el rendimiento PSU en todas sus secciones, la cantidad de padres vivos, la evaluación que el estudiante realiza a los docentes y el nivel educacional de los padres. En otras palabras, tal como lo plantea Spady (1970b), las variables relacionadas con el potencial académico (rendimiento preuniversitario), antecedentes familiares y contexto educacional impactan la decisión del estudiante de permanecer en el programa. Esta relación identificada podría servir a la gestión de selección de candidatos por parte de las autoridades, como también, identificar qué estudiantes son potenciales desertores e implementar herramientas para evitar la salida temprana del estudiante durante los primeros cuatro semestres. Adicionalmente, para evitar la deserción en el segundo año, es recomendable la extensión del programa de apoyo académico hasta finales del segundo año, puesto que el desempeño acumulado luego del segundo año es un predictor de mayor peso que los antecedentes preuniversitarios.

Para el quinto semestre se identifica un fuerte cambio, puesto que tres predictores más importantes eran solamente características que se capturaban una vez ingresado el estudiante a la universidad. Esta tendencia se comenzó a detectar desde el segundo semestre, donde las variables de rendimiento y reprobación pasaban del segundo cuartil al primer cuartil de predictores más importantes y las variables preuniversitarias bajaban del primer al segundo. Sin embargo esta tendencia no continuó para el sexto semestre, donde los mismos variables que eran considerados como predictores importantes en los primeros dos semestres volvían a aparecer en el primer cuartil.

De acuerdo al análisis realizado en las secciones anteriores, los perfiles varían de un semestre a otro, lo que sustenta las teorías planteadas al inicio de esta tesis. Los alumnos están en constante evaluación de la decisión de desertar, el cual se ve afectado por características definidas previas al ingreso a la universidad, el desempeño obtenido durante el transcurso del programa y su relación con el ambiente universitario.

Cuando el estudiante ingresa a un programa educacional, según sus antecedentes familiares y escolares preuniversitarios, establece sus objetivos educacionales iniciales y su compromiso institucional. Posteriormente, después de un tiempo suficiente en que interactúa con su ambiente universitario, estos objetivos educacionales y compromiso institucional son ajustados, pudiéndose desencadenar en una deserción.

La descripción del párrafo anterior se puede observar claramente en los perfiles por semestre, donde al inicio del programa los alumnos con género femenino, provenientes en su mayoría de colegios particulares subvencionados o municipales, con integrantes de la familia estudiando y con altos niveles de financiamiento no desertan. Algunas de estas características en general se repiten en los tres años, como lo es el género, integrantes estudiando en otra institución educacional y el grupo dependencia del colegio.

Las variables que capturan la participación del estudiante con su ambiente académico comienzan a tomar importancia, principalmente para el segundo año, donde aquellos

estudiantes con bajo rendimiento acumulado, pero con alta evaluación a sus profesores desertan. Sin embargo, no es así en el caso de los estudiantes que se quedan en el programa, donde tiende a ocurrir lo contrario.

La configuración familiar es generalmente repetida en todo los perfiles de los distintos semestres. Por ejemplo, mientras mayor es el nivel educacional de al menos uno de los padres, la tendencia a desertar voluntariamente por parte del alumno es mayor. Ocurre lo mismo cuando aumenta el número de integrantes de la familia estudiando en algún nivel de la media. Sin embargo, este efecto no es el mismo cuando los integrantes están estudiando en otro nivel educacional. Eventualmente esto podría mostrar la dificultad de la familia en que dos integrantes estén estudiando en la universidad, aun cuando los niveles de financiamiento que estos estudiantes reciben son en promedio mayor que los no desertores.

En general, se puede describir un perfil desertor como aquellos estudiantes que estudiaron en colegios particulares pagados, son del género masculino, sus padres tienen niveles educacionales altos, el ingreso bruto familiar es alto, tienen integrantes de la familia estudiando en la media, la evaluación a los profesores es alta al menos hasta el tercer semestre y su desempeño en la PSU es ligeramente mayor, no así en la universidad. Para el caso de los no desertores, son aquellos alumnos que vienen de colegios principalmente particulares subvencionado, son del género femenino, sus padres tienden a tener a no superar el nivel educacional escolar completo, el ingreso bruto familiar es relativamente bajo y reciben financiamiento altos financiamientos, sus puntajes PSU no son muy latos al igual que el promedio de sus notas; sin embargo son los que menos créditos reprueban durante el transcurso del programa.

La tendencia antes mencionada permitiría la identificación temprana de los estudiantes que desertarían en algún semestre específico. Adicionalmente, dado que las variables de rendimiento van tomando mayor importancia durante los semestres, sobre todo en el semestre donde existe mayor número de deserciones, permitiría identificar el tipo de alumno que necesita apoyo académico con el objetivo de reducir la cantidad de créditos reprobados y aumentar el promedio de notas obtenido por semestre.

Es importante agrupar los alumnos según los antecedentes familiares y el rendimiento que tienen por semestre. El hecho de que 5 de los 6 mejores modelos estén compuestos por técnicas de *clusterización* responde a que los alumnos tienden a agruparse y dependiendo de sus características, algunos de ellos desertan y los otros se mantienen en la carrera.

Los seis modelos generados permitirán a la escuela predecir de manera semestral el conjunto de alumnos que podría desertar, y de esta manera aplicar políticas educacionales sobre este grupo de estudiantes para reducir los efectos negativos la deserción. En otras palabras, utilizar el conjunto de estudiantes que los modelos predicen como desertores para focalizar el grupo de individuos que deberían participar en programas de apoyo tales como, talleres de contextualización, programas de apoyo académico y programas de apoyo psicológico.

ANEXOS

Anexo 1 – Modelos por semestre

Las siguientes tablas muestran el total de combinatorias de técnicas de minerías de datos, los cuales generan un total de los 48 modelos por semestres.

Modelo	Máquina Apr	Clustering	Clasificador	Balanceo
1	SVM	Clustering	ThresNo	NoBalanced
2	SVM	Clustering	ThresNo	ROS
3	SVM	Clustering	ThresNo	RUS
4	SVM	Clustering	ThresSí	NoBalanced
5	SVM	Clustering	ThresSí	ROS
6	SVM	Clustering	ThresSí	RUS
7	SVM	NoClustering	ThreshNo	NoBalanced
8	SVM	NoClustering	ThreshNo	ROS
9	SVM	NoClustering	ThreshNo	RUS
10	SVM	NoClustering	ThresSí	NoBalanced
11	SVM	NoClustering	ThresSí	ROS
12	SVM	NoClustering	ThresSí	RUS
13	DT	Clustering	ThreshNo	NoBalanced
14	DT	Clustering	ThreshNo	ROS
15	DT	Clustering	ThreshNo	RUS
16	DT	Clustering	ThresSí	NoBalanced
17	DT	Clustering	ThresSí	ROS
18	DT	Clustering	ThresSí	RUS
19	DT	NoClustering	ThreshNo	NoBalanced
20	DT	NoClustering	ThreshNo	ROS
21	DT	NoClustering	ThreshNo	RUS
22	DT	NoClustering	ThresSí	NoBalanced
23	DT	NoClustering	ThresSí	ROS
24	DT	NoClustering	ThresSí	RUS
25	LR	Clustering	ThreshNo	NoBalanced
26	LR	Clustering	ThreshNo	ROS
27	LR	Clustering	ThreshNo	RUS
28	LR	Clustering	ThresSí	NoBalanced
29	LR	Clustering	ThresSí	ROS
30	LR	Clustering	ThresSí	RUS
31	LR	NoClustering	ThreshNo	NoBalanced

Modelo º	Máquina Apr	Clustering	Clasificador	Balanceo
32	LR	NoClustering	ThreshNo	ROS
33	LR	NoClustering	ThreshNo	RUS
34	LR	NoClustering	ThresSí	NoBalanced
35	LR	NoClustering	ThresSí	ROS
36	LR	NoClustering	ThresSí	RUS
37	NN	Clustering	ThreshNo	NoBalanced
38	NN	Clustering	ThreshNo	ROS
39	NN	Clustering	ThreshNo	RUS
40	NN	Clustering	ThresSí	NoBalanced
41	NN	Clustering	ThresSí	ROS
42	NN	Clustering	ThresSí	RUS
43	NN	NoClustering	ThreshNo	NoBalanced
44	NN	NoClustering	ThreshNo	ROS
45	NN	NoClustering	ThreshNo	RUS
46	NN	NoClustering	ThresSí	NoBalanced
47	NN	NoClustering	ThresSí	ROS
48	NN	NoClustering	ThresSí	RUS

Tabla 50. Modelos aplicados por semestre según combinación de técncias de mienría de datos.

Anexo 2 – Diccionario de datos de variables obtenidas de la Base de Datos SAD y Becas y Créditos

Nombre de variables y su descripción obtenidas de la base de datos Becas y SAD:

Variable	Base	Descripción
FINANCIAMIENTO	BECAS	Monto de financiamiento del alumno entregado en el año. Este cálculo se obtuvo para todos los años, el cual se mantiene por semestre
%Reprobadas_Sem	SAD	Porcentaje de créditos reprobados por el alumno al final del semestre. Esta variable se obtuvo para cada semestre del estudiante
%Reprobadas_Sem_Acum	SAD	Porcentaje de créditos reprobados acumulados hasta el final del semestre. Esta variable se obtuvo para cada semestre del estudiante
%Reprobadas_Sem_Anterior	SAD	Porcentaje de créditos reprobados por el alumno al final del semestr anterior. Esta variable se obtuvo para cada semestre del estudiante
EvalProf_Sem	SAD	Evaluación promedio de los docentes asignado por el alumno al final del semestre. Esta variable se obtuvo para cada semestre del estudiante
EvalProf_Sem_Acum	SAD	Evaluación promedio acumulada de los docentes asignado por el alumno al final del semestre. Esta variable se obtuvo para cada semestre del estudiante
EvalProf_Sem_Anterior	SAD	Evaluación promedio de los docentes asignado por el alumno al final del semestre anterior. Esta variable se obtuvo para cada semestre del estudiante
NotaSem	SAD	Promedio ponderado del alumno al final del semestre. Esta variable se obtuvo para cada semestre del estudiante

Variable	Base	Descripción
NotaSem_Acum	SAD	Promedio ponderado del alumno acumulado hasta el final del semestre. Esta variable se obtuvo para cada semestre del estudiante
NotaSem_Anterior	SAD	Promedio ponderado del alumno al final del semestre anterior. Esta variable se obtuvo para cada semestre del estudiante
PostergaSem	SAD	1 si el alumno postergó en el semestre,0 en caso contrario. Esta variable se obtuvo para cada semestre del estudiante
PostergaSem_Acum	SAD	Número de postergaciones acumuladas hasta el final del semestre. Esta variable se obtuvo para cada semestre del estudiante
PostergaSem_Anterior	SAD	1 si el alumno postergó en el semestre anterior, 0 en caso contrario. Esta variable se obtuvo para cada semestre del estudiante
Sem_Verano_Acum	SAD	Indica el número de veces que el alumno ha participado en los semestre verano. Esta variable se obtuvo para cada semestre del estudiante
Sem_Verano_Ant	SAD	Indica si el alumno participó en el semestre verano anterior. Esta variable se obtuvo para cada semestre del estudiante
Label	SAD	Estado académico final del estudiante una vez terminado el semestre. Para esta tesis se identificó como DESERTA los alumnos que renunciaron voluntariamente o NO DESERTA aquellos que siguieron sus estudios en el semestre siguiente. Variable a estudiar.

Tabla 51. Descripción y definición de los datos obtenidos desde base de datos Becas y SAD.

Anexo 3 – Diccionario de datos de variables obtenidas de la Base de DEMRE

Nombre de variables y su descripción obtenidas de la base de datos DEMRE:

Variable	Base	Descripción
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	DEMRE	Indica si los padres están vivos. Rango: 0 de los padres, 1 de los padres, 2 de los padres.
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	DEMRE	Número de integrantes de la familia en Básica
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	DEMRE	Número de integrantes de la familia en Media 1° a 3°
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	DEMRE	Número de integrantes de la familia en Media 4°
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	DEMRE	Número de integrantes de la familia en otras instituciones de estudio
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	DEMRE	Número de integrantes de la familia en Pre-Básica
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	DEMRE	Número de integrantes de la familia en Superior
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	DEMRE	Número de personas del grupo familiar que trabajan
EDUCACION_DEL_PADRE_EQUIVALENCIA	DEMRE	Nivel educacional del padre medida en años
EDUCACION_DELA_MADRE_EQUIVALENCIA	DEMRE	Nivel educacional de la madre medida en años

Variable	Base	Descripción
GRUPO_DEPENDENCIA = Municipal	DEMRE	Indica si del colegio que proviene el alumno es Municipal
GRUPO_DEPENDENCIA = Particular Pagado	DEMRE	Indica si del colegio que proviene el alumno es Particular Pagado
GRUPO_DEPENDENCIA = Particular Subvencionado	DEMRE	Indica si del colegio que proviene el alumno es Particular Subvencionado
GRUPO_FAMILIAR	DEMRE	Número de personas que componen el grupo familiar
HORAS_QUE_DEDICA_TRABAJO	DEMRE	Número de horas que dedicaba al trabajo
INGRESO_BRUTO_FAMILIAR_GRUPO	DEMRE	Nivel del ingreso bruto familiar
Pref	DEMRE	Preferencia de ingreso a la carrera
PROM_LM_FINAL	DEMRE	Promedio PSU puntaje en Lenguaje y Matemáticas
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	DEMRE	Puntaje PSU en Ciencias (Mejor entre Historia y Ciencias)
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	DEMRE	Puntaje PSU en Lenguaje y comunicación
PTJE_MATEMATICA_FINAL	DEMRE	Puntaje PSU en Matemática
PTJE_NEM	DEMRE	Puntaje PSU NEM

Variable	Base	Descripción
RAMA_EDUCACIONAL_EQUIVALENCIA = T	DEMRE	Indica si la rama educacional del colegio que proviene el alumno es Técnico
RAMA_EDUCACIONAL_EQUIVALENCIA = H	DEMRE	Indica si la rama educacional del colegio que proviene el alumno es Humanista
REGIMEN = Coeducacional	DEMRE	Indica si del colegio que proviene el alumno es Coeducacional
REGIMEN = Femenino	DEMRE	Indica si del colegio que proviene el alumno es Femenino
REGIMEN = Masculino	DEMRE	Indica si del colegio que proviene el alumno es Masculino
Sexo = HOMBRE	DEMRE	Indica si el género del alumno es HOMBRE
Sexo = MUJER	DEMRE	Indica si el género del alumno es MUJER

Tabla 52. Descripción y definición de los datos obtenidos desde base de datos DEMRE.

Anexo 4 – Procesos en Software

En las siguientes figuras se muestra el encadenamiento de los procesos generados en el desarrollo del experimento.

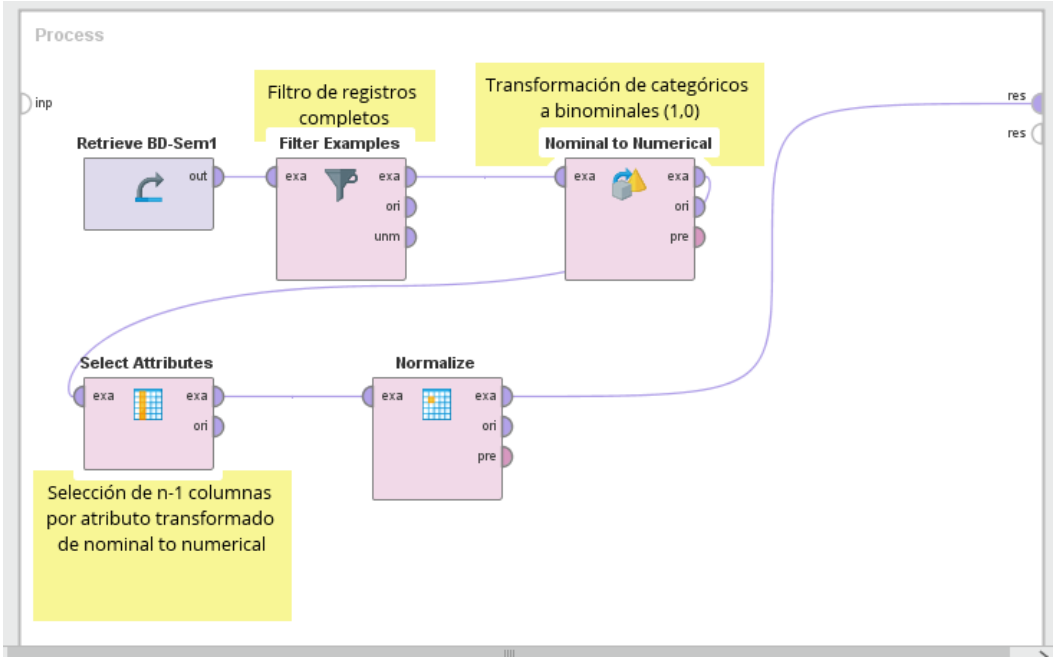

Figura 13. Ejemplo del semestre 1 para el proceso de transformación de los atributos aplicado en cada semestre.

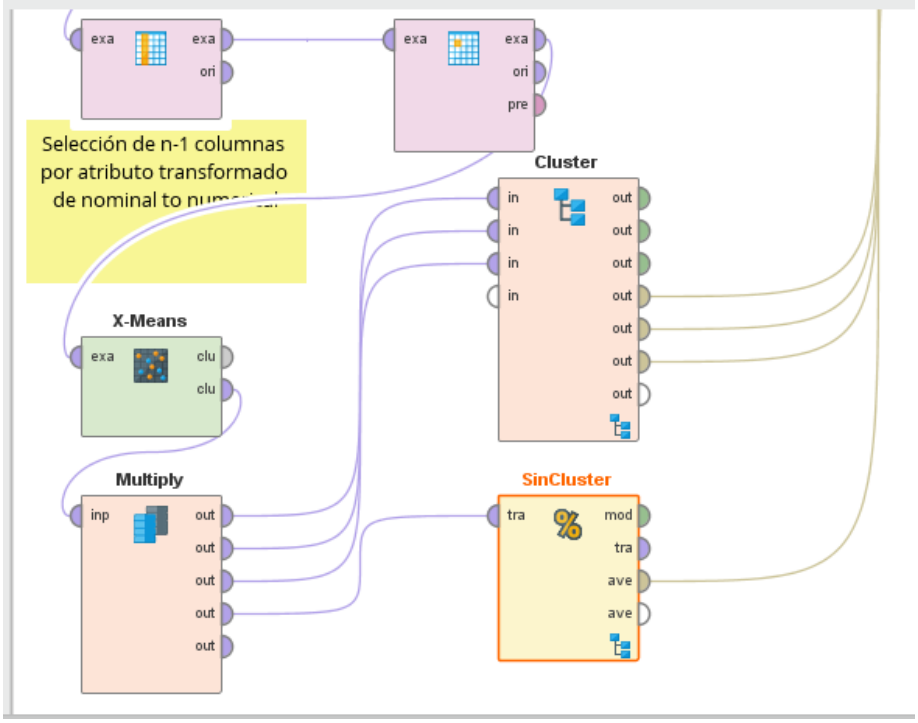

Figura 14. Ejemplo del semestre 1 para el proceso de clusterización y los subprocesos de validación cruzada.

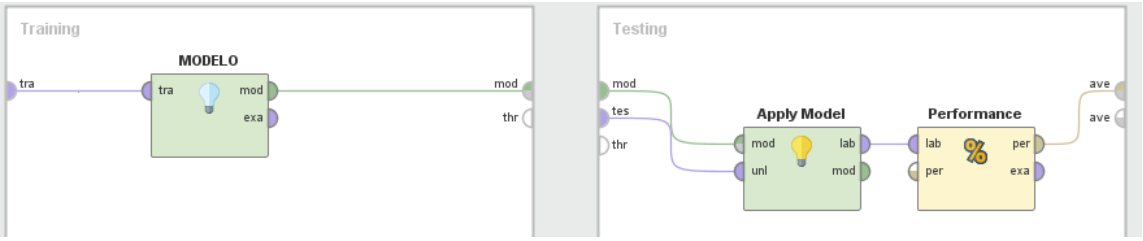

Figura 15. Ejemplo del encadenamiento de operadores dentro del operador X-Validations.

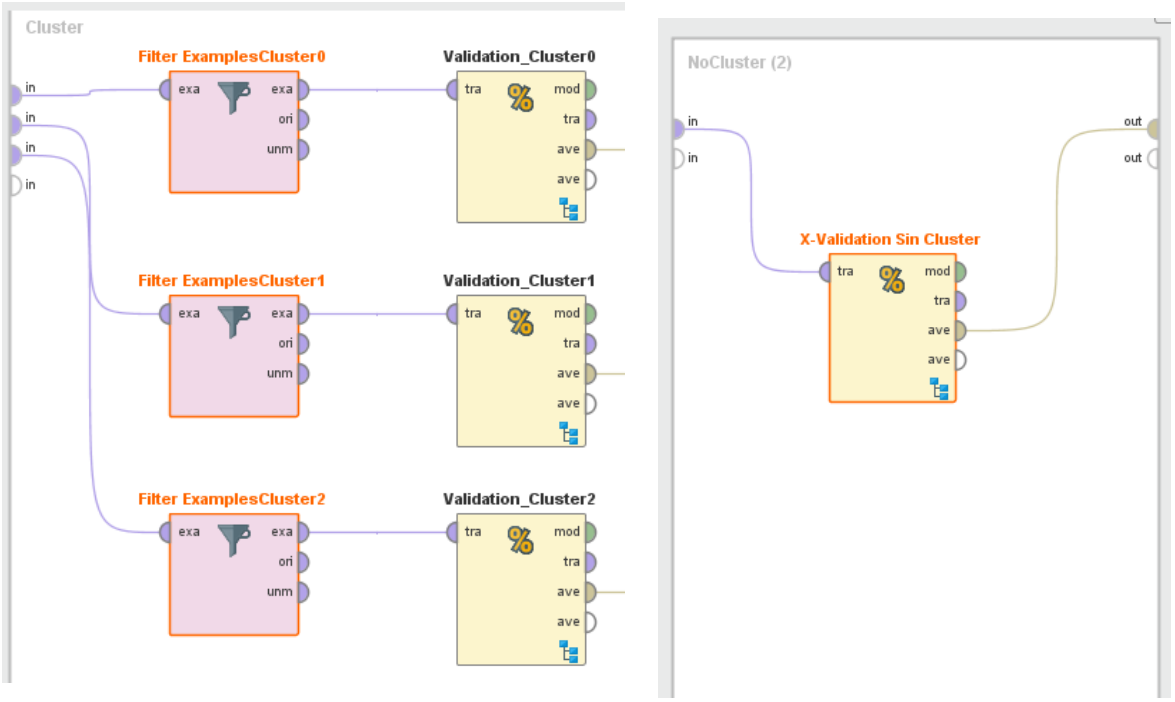

Figura 16. Encadenamiento de la validación cruzada para cada cluster (lado izquierdo) y sin cluster (lado derecho).

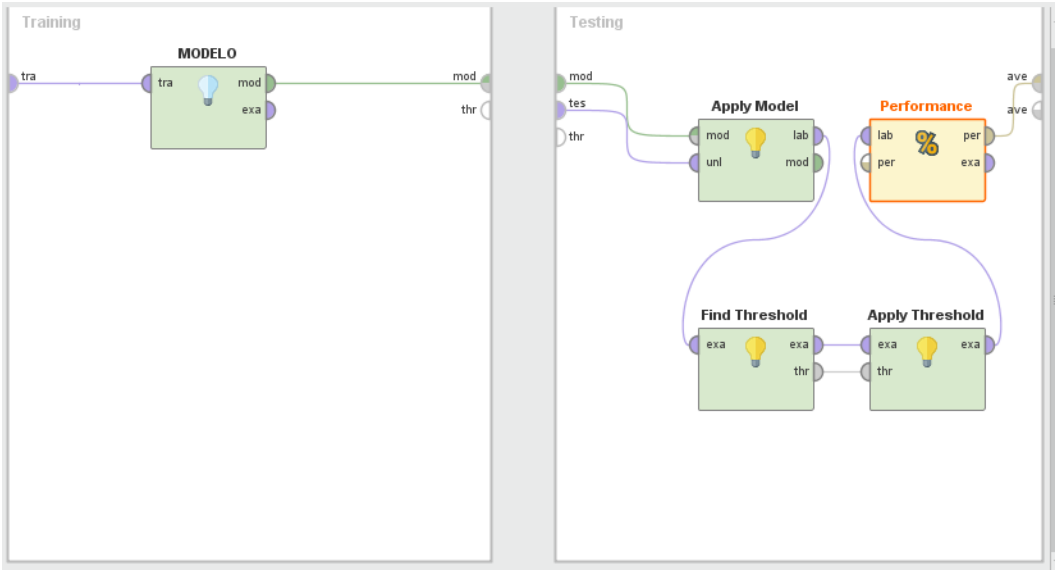

Figura 17. Encadenamiento de aplicación de umbral.

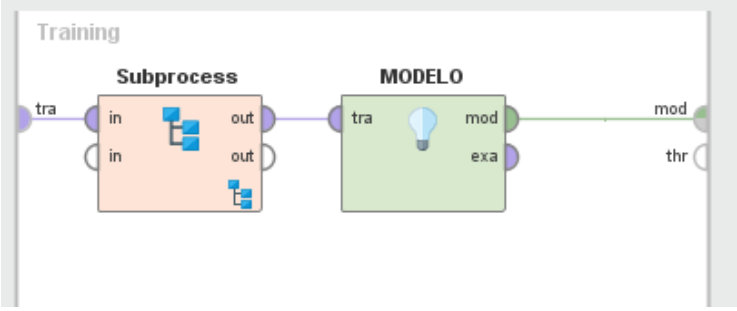

Figura 18. Ejemplo de encadenamiento para ROS con el subproceso antes del modelo.

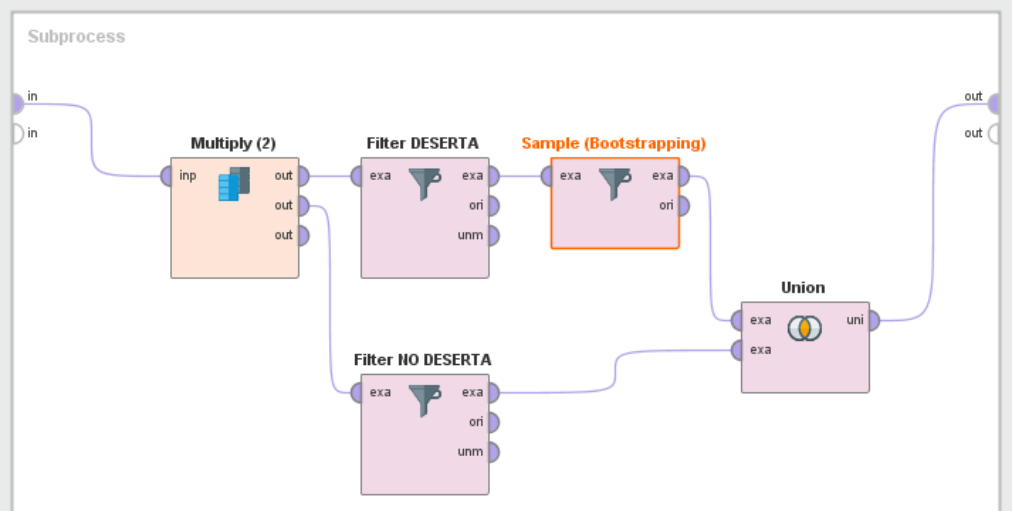

Figura 19. Ejemplo de encadenamiento al interior del subrpoceso para la aplicación de la técnica ROS.

Anexo 5 – Resumen Variables por Semestre

Variable	Base	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Número de Usos
FINANCIAMIENTO	BECAS	✓	✓		✓	✓	✓	5
%Reprobadas_Sem	SAD		✓			✓	✓	3
%Reprobadas_Sem_Acum	SAD		✓			✓	✓	3
%Reprobadas_Sem_Anterior	SAD					✓	✓	2
EvalProf_Sem	SAD	✓	✓	✓	✓	✓	✓	6
EvalProf_Sem_Acum	SAD		✓	✓	✓	✓	✓	5
EvalProf_Sem_Anterior	SAD		✓	✓	✓	✓	✓	5
NotaSem	SAD	✓	✓	✓	✓	✓	✓	6
NotaSem_Acum	SAD		✓	✓	✓	✓	✓	5
NotaSem_Anterior	SAD		✓	✓	✓	✓	✓	5
PostergaSem	SAD							0
PostergaSem_Acum	SAD					✓	✓	2
PostergaSem_Anterior	SAD							0
Sem_Verano_Acum	SAD						✓	1
Sem_Verano_Ant	SAD				✓			1
¿VIVEN_SUS_PADRES?_CUANTOSPADRES	DEMRE	✓	✓	✓	✓		✓	5
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_BASICA	DEMRE					✓	✓	2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA1a3	DEMRE						✓	1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_MEDIA4	DEMRE	✓		✓	✓	✓		4
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_OTRAS	DEMRE					✓	✓	2
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_PREBASICA	DEMRE						✓	1
CUANTOS_ESTUDIAN_GRUPO_FAMILIAR_SUPERIOR	DEMRE						✓	1
CUANTOS_TRABAJAN_GRUPO_FAMILIAR	DEMRE					✓	✓	2
EDUCACION_DEL_PADRE_EQUIVALENCIA	DEMRE	✓	✓	✓	✓	✓	✓	6
EDUCACION_DELA_MADRE_EQUIVALENCIA	DEMRE	✓	✓	✓	✓	✓	✓	6
GRUPO_DEPENDENCIA = Municipal	DEMRE		✓			✓	✓	3
GRUPO_DEPENDENCIA = Particular Pagado	DEMRE		✓	✓	✓		✓	4
GRUPO_DEPENDENCIA = Particular Subvencionado	DEMRE							0
GRUPO_FAMILIAR	DEMRE	✓	✓	✓	✓	✓	✓	6
HORAS_QUE_DEDICA_TRABAJO	DEMRE					✓	✓	2
INGRESO_BRUTO_FAMILIAR_GRUPO	DEMRE	✓	✓	✓	✓	✓	✓	6
Pref	DEMRE					✓		1
PROM_LM_FINAL	DEMRE	✓	✓	✓	✓	✓		5
PTJE_CIENCIASHISTORIA_EQUIVALENCIA	DEMRE					✓	✓	2
PTJE_LENGUAJE_Y_COMUNICACION_FINAL	DEMRE	✓	✓	✓	✓	✓	✓	6
PTJE_MATEMATICA_FINAL	DEMRE	✓	✓	✓	✓	✓	✓	6
PTJE_NEM	DEMRE	✓	✓	✓	✓	✓	✓	6
RAMA_EDUCACIONAL_EQUIVALENCIA = T	DEMRE					✓	✓	2
RAMA_EDUCACIONAL_EQUIVALENCIA = H	DEMRE							0
REGIMEN = Coeducacional	DEMRE							0
REGIMEN = Femenino	DEMRE		✓			✓		2
REGIMEN = Masculino	DEMRE						✓	1
Sexo = MASCULINO	DEMRE							0
Sexo = MUJER	DEMRE	✓	✓	✓	✓			4

Tabla 53. Resumen Variables identificadas como predictores según mejores modelos por Semestre.

REFERENCIAS

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques. *Journal of STEM Education: Innovations and Research*, *15*(3), 35.

Barrios, A. (2013). *Deserción universitaria en Chile: incidencia del financiamiento y otros factores asociados*. Revistacis. Recuperado a partir de http://www.techo.org/wp-content/uploads/2013/02/barrios.pdf

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.

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. En *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on* (pp. 103–106). IEEE.

Braxton, J. M., Shaw Sullivan, A. V., & Johnson, R. M. (1997). Appraising Tinto's theory of college student departure. *HIGHER EDUCATION-NEW YORK-AGATHON PRESS INCORPORATED-*, *12*, 107–164.

Byrd, G., Garza, C., & Nieswiadomy, R. (1999). Predictors of successful completion of a baccalaureate nursing program. *Nurse Educator*, *24*(6), 33–37.

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

Chien, C.-F., & Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with applications*, *34*(1), 280–290.

De MagalhaesL-Calvet, T. (2013). Estudio de los Factores Determinantes de la Deserción en el Sistema Universitario Chileno.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. http://doi.org/10.1016/j.dss.2010.06.003

Díaz, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios pedagógicos (Valdivia)*, *34*(2), 65–86.

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)*.

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

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.

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

Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969–976.

Himmel, E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad de la Educación*, *17*, 91–107.

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515–524.

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial intelligence in medicine*, *16*(1), 3–23.

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

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.

Morales, F., Fuentes, R., Riquielme, S., & Kraemer, H. (2011). Impacto de la intervención del programa de inducción, adaptación y vinculacón a la vida universitaria en la facultad de ciencias empresariales de universidad del Bío Bío. (Vol. 4, pp. 2730–2757). Presentado en ENEFA.

Morales, F., Riquelme, S., Bascuñan, E., & Navarrete, M. (2014). Estudio sobre el éxito académico de estudiantes de ciencias empresariales de la Universidad del Bío-Bío. Presentado en ENEFA.

Nye, J. S. (1976). Independence and interdependence. *Foreign Policy*, 130–161.

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.

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* (First Edition).

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.

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

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

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

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.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). *Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley and Sons.

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

Spady, W. G. (1970b). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330.

Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, *53*(6), 687.

Tinto, V. (2007). *Taking student retention seriously*. Syracuse University.

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

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

Vásquez, J., Ortega, C., Lee, M., & Silva, D. (2015). Carga Académica: Identificación de Factores claves en una escuela de economía y negocios. (Vol. 50). Presentado en CLADEA, Viña del Mar.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229.

Vercellis, C. (2009). Business intelligence: data mining and optimization for decision making. *Editorial John Wiley and Sons*.

Wilson, R., Eva, K., & Lobb, D. K. (2013). Student attrition in the Ontario midwifery education programme. *Midwifery*, *29*(6), 579–584.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, *8*(2), 307–325.