Detección temprana de alumnos en riesgo académico mediante técnicas de clasificación de minería de datos, en estudiantes de la Facultad de Ciencias de la Universidad de Valparaíso

Trabajo final presentado por:
Felipe S. Quezada Barría

Trabajo de titulación para optar al título de:
Ingeniero Estadístico

Profesores guía:
Harvey Rosas Quintero, PhD

Diciembre 2018, Valparaíso, Chile.

Agradecimientos

En primer lugar, quiero agradecer a mi profesor guía, Dr. Harvey Rosas por su apoyo incondicional, compromiso y dedicación desde el principio de mi formación como estudiante. Estuvo presente en toda esta etapa universitaria, tanto en motivación como en palabras de aliento en momentos difíciles.

También agradezco a mi familia que estuvo presente en todo momento, en especial a mis hermanos Jorge Vyhmeister y Mariela Quezada, a mis tías Ana María Barría y Laura Rosales, y a mi padre Eric Quezada.

Dedicado en especial a mi madre Nancy Barría y mi abuela Elena Carvallo, que si bien no están presente en vida en este momento, fueron quienes me enseñaron y me formaron desde pequeño con grandes valores y lo necesario para llegar hasta esta etapa.

Índice general

Índice de figuras

Índice de cuadros

Resumen

El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes en el área de la educación, los cuales son difíciles de comprender y analizar sin la ayuda de software apropiados. Por lo anterior, recientemente ha surgido la necesidad de poder utilizar esta información para identificar aquellos estudiantes en situación de riesgo de abandono, la cual conlleva al descubrimiento de nuevos métodos para su detección temprana.

El presente trabajo de titulación explora y analiza diversas técnicas de minería de datos las cuales puedan ser aplicadas en el entorno educacional. Con el objetivo de poder detectar estudiantes en riesgo académico de manera temprana, para así, intentar disminuir las tasas de abandono de la Universidad de Valparaíso.

Los modelos de clasificación de minería de datos logran clasificar de manera eficiente a los alumno que se encuentran en riesgo académico en la Facultad de Ciencias de la Universidad de Valparaíso. Con esto se logra desarrollar intervenciones focalizadas en aquellos estudiantes que se encuentran en situación de riesgo.

Palabras clave: clasificación logística, máquinas de soporte vectorial, redes neuronal, *KDD*, sistema de alerta temprana, técnicas de minería de datos.

Abreviaturas

- ***DM*: *Data mining***
- ***EDM*: *Educational data mining***
- ***DW*: *Data wharehouse***
- **SIAT: Sistema de alerta temprana**
- **SIES: Servicio de información de educación superior**
- ***KDD*: *Knowledge discovery in databases***
- ***ANN*: *Artificial neural networks***
- ***SVM*: *Support vector machines***
- ***MLP*: *MultiLayer perceptron***
- **NEM: Notas de enseñanza media**

Capítulo 1

Introducción

La deserción académica es una problemática que se presenta en todas las instituciones educativas, tanto en Chile como en el resto del mundo. Este fenómeno se presenta esencialmente en niveles de educación media y educación superior, donde hay considerables cantidades de estudiantes que abandonan sus estudios. En los últimos años, las instituciones se han preocupado por abordar este problema y así, poder identificar información que contribuya a determinar sus causas.

La definición de deserción académica difiere entre investigadores. Tinto (1989) explica que el estudio de la deserción es extremadamente complejo, ya que implica consideraciones no sólo desde la perspectiva de los estudiantes, sino también una variedad de tipos de abandono. Adicionalmente, sugiere que los investigadores seleccionen la definición que mejor se adapte a sus objetivos y al problema a investigar.

En Chile, el problema de la deserción también ha sido abordada y discutida por distintos autores, entre los cuales se destaca Himmel (2018), definiendo la deserción académica como el abandono prematuro de un programa de estudios antes de alcanzar el título o grado, considerando un tiempo suficientemente largo como para descartar la posibilidad que el estudiante se reincorpore. Siendo esta última, la base teórica para definir la deserción académica en este trabajo.

El Servicio de Información de Educación Superior (SIES), identifica que los costos de la deserción, incluyen factores de tipo económico como la inversión del estado, las instituciones y las familias, así como factores psicosociales relativos a la frustración que afecta a los estudiantes que abandonan sus estudios (SIES, 2014).

A nivel internacional, según un informe de la Organización para la Cooperación y el Desarrollo Económico (OCDE, 2007), presenta información de 18 países. Encontró que las tasas de abandono académico, son alrededor de un 31 % en los estudiantes de educación universitaria. Esto considerando tanto estudiantes que se retiran al primer

año como aquellos que lo hacen en años posteriores. Las tasas de deserción difieren ampliamente entre los países de la OCDE; Italia, Estados Unidos, Nueva Zelanda y Hungría, las cuales superan el 40 % de deserción en estudiantes. Por otro lado, Japón, Dinamarca y Bélgica, presentan tasas inferiores al 20 %. En España, según Michavila and Martínez (2016), al cabo del tercer año de estudios superiores, la tasa de deserción alcanzada corresponde a un 32 % en estudiantes que abandonan el programa de estudio. En Colombia, según Orozco (2016), las tasas de deserción en estudiantes universitarios, se han mantenido durante los años 2010 al 2014, alcanzando en promedio 45,4 % de abandono en los programas de estudio.

En Chile, las estadísticas del Servicio de Información de Educación Superior (SIES, 2014), muestran que las tasas de deserción al cabo del primer año en universidades chilenas, han disminuido desde el año 2012 con una tasa del 25,5 % al año 2016, con una tasa cercana al 22 % de estudiantes que abandonan el programa de estudio. Se debe considerar que, para el caso de Chile sólo se tienen estadísticas al cabo del primer año de ingreso y no de los años posteriores, por lo cual, estos valores aumentan al transcurrir el segundo y tercer año de estudio. Por otro lado, Santelices et al. (2013), estiman que la deserción al primer año es cercana al 25 % y cerca del 40 % al cabo del tercer año, con una gran variación en las tasas según el tipo de institución.

En las instituciones de educación superior existen distintas fuentes de información, llamadas almacén de datos (*data warehouse*, en Inglés), éstas requieren de grandes volúmenes de almacenamiento y una estructura definida. A su vez, esto conlleva a una problemática, dado que en muchas ocasiones no se sabe qué datos elegir a la hora de analizar e interpretar la información sobre los alumnos. En los almacén de datos se cuenta con distinta información proveniente de los estudiantes, entre los que se destacan: datos personales, socio-económicos, antecedentes académicos del desempeño antes y durante su permanencia en la institución, información demográfica, entre otras.

El crecimiento en la disponibilidad de información digital en el mundo, ha generado muchas oportunidades de aplicación de distintos métodos para su análisis, donde el área de la educación no es una excepción.

A comienzos de este siglo, surgió una comunidad de la investigación usando herramientas estadísticas, matemáticas y computacionales para la finalidad de analizar datos en el entorno educativo, siendo esta llamada minería de datos educacional (*Educational Data Minnig*, en Inglés), la cual nace de la minería de datos (*Data minning*, en Inglés), con la finalidad de utilizar los grandes almacenes de datos que poseen las instituciones de educación superior, permitiendo mejorar los sistemas de evaluación, el entendimiento en los procesos educativos, y la priorización y diseño de intervenciones educativas (Siemens and Baker, 2012).

Así, el objetivo general de este trabajo titulación es: proponer un modelo eficiente de clasificación para la detección temprana de alumnos en riesgo académico de la Universidad de Valparaíso, específicamente en la Faculad de Ciencias. De lo dicho anteriormente, se desprenden los siguientes objetivos específicos:

(i) Explorar y analizar diversas técnicas de minería de datos que se puedan aplicar en el entorno educacional.

(ii) Consolidar un conjunto de datos homogéneo con la totalidad de la información presente de los estudiantes de la Facultad de Ciencias.

(iii) Determinar un modelo eficiente para la predicción de estudiantes en riesgo académico.

(iv) Articular un modelo predictivo como apoyo a un sistema de toma decisiones para los estudiantes y docentes, como también para las unidades académicas.

Con estos antecedentes se puede plantear que, utilizando técnicas de minería de datos es posible diseñar, adaptar y proponer un modelo eficiente de clasificación, para la detección temprana de alumnos en riesgo académico en la Facultad de Ciencias de la Universidad de Valparaíso.

Con este trabajo se busca motivar la aplicación de la minería de datos en el entorno educativo, haciendo uso de estos repositorios de información que se tiene en las instituciones académicas. Y así, poder detectar alumnos en riesgo académico y de este modo, brindar herramientas de apoyo a los estudiantes y con ello, poder disminuir las tasas de deserción en la Universidad de Valparaíso, puntualmente en la Facultad de Ciencias. Por otra parte, estos métodos son beneficiosos tanto para alumnos, como profesores y de la misma manera para la unidades académicas de la universidad.

Capítulo 2

Estado del arte

El avance de las tecnologías de la información y comunicación (TIC), se han posicionado como una de las fuentes principales de innovación, crecimiento y desarrollo a nivel mundial, lo que trae consigo ventajas competitivas en los sectores que se han implementado, principalmente en el área de los negocios, salud y educación (Martelo et al., 2016).

El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes; los cuales son, por lo general, difíciles de comprender y analizar. En muchas ocasiones, más que pensar en nuevas tecnologías a futuro para obtener mayor información, lo que se debe realizar es invertir esfuerzo para construir modelos de enseñanza para obtener el máximo partido a las tecnologías que se tienen actualmente en las instituciones de educación. La innovación no se consigue por la novedad de la aplicación tecnológica, sino por la aplicación de criterios para conseguir nuevos escenarios formativos y comunicativos (Cabero Almenara, 2015).

En las dos últimas décadas, y de forma paralela al desarrollo de los sistemas de información, se ha implementado el conocimiento como recurso estratégico y los conjuntos de datos con información relacionada han pasado a convertirse en valiosos repositorios para la utilización de técnicas de minería de datos, que pasan a ser herramientas fundamentales en cualquier ámbito científico o empresarial, ya que permiten obtener información útil, permitiendo procesar, analizar y extraer patrones en cualquier proyecto de investigación (Gutiérrez, 2016).

2.1. Minería de datos en el entorno educativo

La necesidad de poder detectar a estudiantes en situación de riesgo académico, ha llevado al descubrimiento de nuevos métodos para su detección. En los últimos años, ha nacido un área en la disciplina educativa llamada minería de datos educacional (*Educational Data Mining*, en Inglés). La *EDM* ha aparecido como una nueva área relacionada con el desarrollo, investigación y la aplicación de métodos computacionales, para detectar patrones en grandes colecciones de datos educativos, que de otro modo sería muy difícil o casi imposible de analizar debido al enorme volumen de información que existe (Romero and Ventura, 2013).

Hechos que dan cuenta del aumento de interés en esta área, son las conferencias internacionales de minería de datos educacionales que se realizan a partir de 2008. Además, en el año 2009 se publica una revista anual de minería de datos educacional (*Journal of Educational Data Mining*). Por otro lado, en julio de 2011 se fundó la Sociedad International de Minería de Datos Educacional (*International Educational Data Mining Society*) , cuyo objetivo es apoyar la colaboración y el desarrollo científico en esta nueva disciplina, a través de la organización de conferencias de *EDM*, y así, apoyar el intercambio de datos y técnicas utilizadas en esta nueva área.

Lo anterior se ha realizado con el propósito de, entre otras cosas, predecir el desempeño educativo de estudiantes y la permanencia o retención de éstos en la escuela, o para clasificar a estudiantes en grupos de acuerdo a sus características. En el caso de la deserción, el desafío está no solo en predecirla, sino también realizar la predicción información anterior al evento, idealmente "lo antes posible"(Márquez-Vera et al., 2016).

2.2. Sistemas para la detección temprana de riesgo académico

Una universidad que experimenta el fenómeno de la deserción académica, es una institución que tiene menor capacidad de retención sobre sus estudiantes. Esto a su vez, trae consecuencias, tanto para el alumno, como para las familias y las instituciones educativas (SIES, 2014).

Para tratar de reducir esta problemática, es necesario detectar a estudiantes que se encuentren en riesgo académico, así poder brindarles un Sistema de Alerta Temprana (SIAT). Seidman (1996) desarrolló una fórmula que se muestra en la Ecuación 2.1, con la cual explica que, para la retención de un estudiante, requiere ser identificado de manera temprana el riesgo académico de éste, e implementando oportunamente un sistema de

apoyo intensivo y continuo, logrando de esta manera disminuir las tasas de deserción en las instituciones educativas.

$$\text{Retención} = \text{Identificación Temprana} + \text{Intervención (Intensiva + Continua)} \quad (2.1)$$

Sin embargo, el desafió de una detección temprana de alumnos en riesgo académico, es una tarea difícil. Las técnicas de clasificación de minería de datos, no se adaptan bien a la naturaleza temporal de este tipo de información, porque normalmente se considera que todos los atributos están siempre disponibles (Antunes, 2010).

Márquez-Vera et al. (2016) explican que el desafío de generar SIAT ha pasado de ser un problema académico, a un desafío de políticas públicas, que ya ha sido implementado por gobiernos como Estados Unidos, México, Inglaterra, Austria y Croacia. Esto se ha realizado con la finalidad de predecir el desempeño académico del estudiante, su permanencia en la institución y la clasificación de alumnos en riesgo académico, entre otras cosas. Logrando obtener como resultado un modelo de clasificación suficientemente confiable para realizar una predicción temprana de abandono, antes de la mitad del curso.

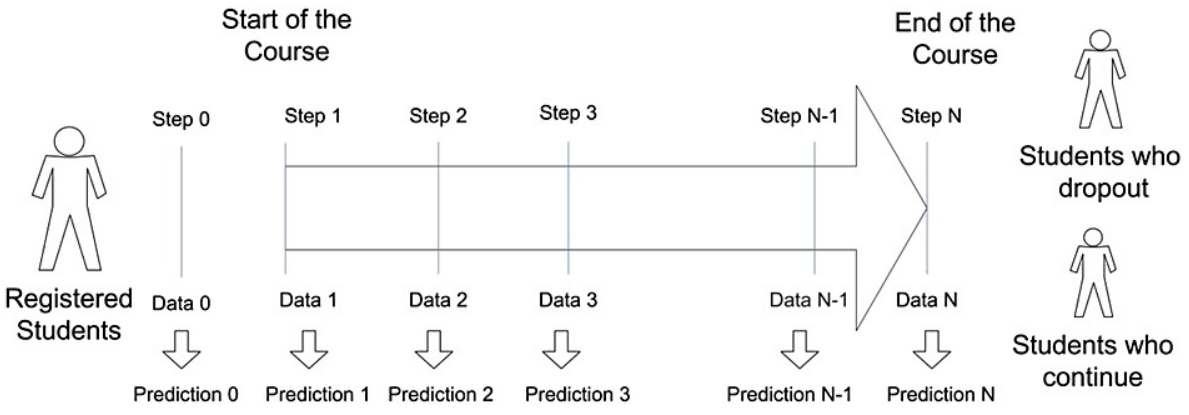

Figura 2.1: Metodología para una predicción temprana de abandono (Márquez-Vera et al., 2016).

En la Figura 2.1, los autores plantean que incluso al comienzo del curso, se puede hacer una predicción de abandono temprano utilizando solo los datos disponibles de la información personal y administrativa del alumno. A medida que avanza el curso, progresivamente se obtiene más información sobre las actitudes, actividades y el rendimiento de los estudiantes. Por lo tanto, no es necesario esperar hasta el final del

curso para predecir si un estudiante continuará al siguiente curso o abandonará este. El problema real está en determinar una etapa temprana en la que la predicción sea lo suficientemente confiable.

Dada la necesidad de entender estos grandes volúmenes de datos que se presentan en las instituciones académicas, la minería de datos en el entorno educativo, ha adquirido gran importancia en el último tiempo, debido a su implementación como sistema de apoyo o alerta temprana. Ésto ha llevado a múltiples investigadores a abordar el problema de la deserción o abandono académico. A continuación, se presentan algunas investigaciones realizadas en los últimos años en este ámbito.

2.3. Minería de datos para la predicción temprana

Peña-Ayala (2014) realiza una revisión de los trabajos hechos en minería de datos educacional al año 2014. Encontró 240 investigaciones entre los años 2010 al 2013, identificando su enfoque y herramientas utilizadas en cada uno de ellos. Los resultados obtenidos muestran que existen diversas disciplinas que abordan estos estudios, entre ellos: la probabilidad, el aprendizaje automático y la Estadística, estos se encuentran presente en un 87 % de los trabajos relacionados a *EDM*. Por otra parte, en el 42 % de estas investigaciones, se tiene como tarea la clasificación de los alumnos. En cuanto a las técnicas utilizadas, existe una gran variedad y no se logra identificar un predominio en la utilización para datos educativos, ya que depende específicamente del objetivo y de los datos que se estén analizando. Lo recomendable a realizar, es utilizar distintas técnicas para entrenar, y posteriormente verificar cuál de estas tiene el mejor poder predictivo.

La predicción del rendimiento académico mediante la clasificación de los estudiantes, es una de las más populares implementadas en el área de la *EDM*, en la cual se han utilizado diferentes técnicas para mitigar este problema, tales como: redes neuronales, máquinas de vectores de soporte (*Support Vector Machine*, en Inglés), redes bayesianas, reglas de decisión y análisis de regresión, entre otras.

A continuación se describen algunos trabajos que han utilizado técnicas de minería de datos para la clasificación de estudiantes en riesgo:

Delen (2010) realiza un análisis comparativo de técnicas de aprendizaje automático para la retención de estudiantes en una institución de los Estados Unidos, la cual contaba con información de 16.066 estudiantes matriculados entre 2004 y 2008, esta era de tipo académico, financiera y demográfica del alumno. Las técnicas utilizadas fueron: árboles de decisión, redes neuronales, *SVM* y regresión logística. Los resultados alcanzaron una precisión cercana al 80 % en todas las técnicas, siendo *SVM* las

más alta con un 81,18 %. Además, se observó que el conjunto de datos balanceado (en comparación con el conjunto de datos desbalanceado) produce mejores modelos de predicción para la detección de estudiantes que están en riesgo de abandonar antes del segundo año. Entre los factores más importantes se encuentran aquellos relacionados con el éxito educativo en secundaria e institución actual, además de la presencia de ayuda financiera, como por ejemplo becas y créditos.

Jia and Mareboyana (2013) aplicaron algoritmos de aprendizaje automático y modelos predictivos para la retención de estudiantes universitarios en Estados Unidos, los cuales presentaban información de 771 alumnos y 12 atributos, entre los años 2006 y 2011; se emplearon técnicas como: árboles de decisión, redes neuronales y máquinas de vectores de soporte. Se observó en los resultados que las medidas de eficiencia aumentaron, al considerar un conjunto de datos con mayor información por año académico de los estudiantes, obteniendo el mejor modelo con la técnica de SVM y tan solo seleccionando dos atributos significativos (promedio acumulado de calificaciones y horas totales de crédito tomada), alcanzando una medida de precisión del 94,29 %.

En México, Márquez-Vera et al. (2016) implementaron técnicas de minería de datos para la predicción temprana de abandono, en donde se consideraron 419 estudiantes de secundaria matriculados en la Universidad Autónoma de Zacatecas, de los cuales se utilizó la información académica al momento de ingreso de estos y posteriormente, se aplicaron encuestas para obtener mayor información en el transcurso del primer semestre. Estas fueron aplicadas en tres instancias, al paso de la cuarta, sexta y décima semana de haber iniciado el periodo regular. De esta manera, con el uso de las encuestas se pasó de tener, 12 atributos a un total de 60. Los resultados evidenciaron que, utilizando regresión logística, es posible detectar tempranamente a un alumno en riesgo al cabo de la sexta semana, obteniendo un nivel de precisión del 73,7 % y al término de semestre esta medida aumenta a un 81 %.

A nivel nacional Celis et al. (2015) propusieron un modelo analítico para la predicción del rendimiento académico de estudiantes de primer año del plan común de Ingeniería y Ciencias de la Universidad de Chile, mediante la técnica de regresión logística. Para la construcción del modelo utilizaron datos de ingreso de 3.573 alumnos, comprendido entre 2010 y 2014, en la cual se contaba con información personal y académica. Entre las variables significativas en el modelo, se destacan: sexo, tipo de establecimiento de enseñanza media, créditos reprobados, promedio de notas del 1er semestre, entre otras. Los resultados obtenidos alcanzaron una sensibilidad del 86 % y una medida de precisión de 37,5 %. Además, estos resultados apoyaron una serie de intervenciones, desde comunicaciones personalizadas a los estudiantes, reforzamientos y tutorías para los alumnos en riesgo.

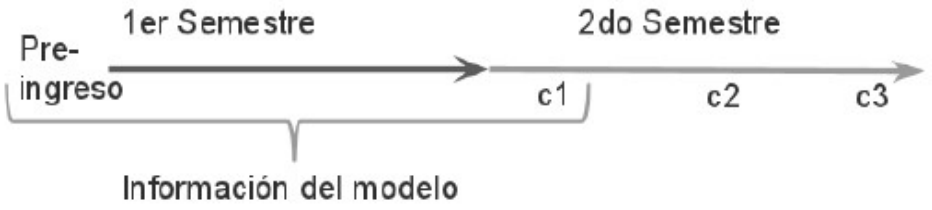

Figura 2.2: Tiempo de captura de información académica para el modelo propuesto en (Celis et al., 2015).

En la Figura 2.2, se puede observar el tiempo en que es intervenido el alumno para clasificarlo como en riesgo o no, y así brindar el apoyo necesario para evitar un posible abandono en la institución.

En un trabajo reciente en Chile, Miranda and Guzmán (2017) analizaron la deserción de estudiantes universitarios de la carreras de Ingeniería de la Universidad Católica del Norte en Antofagasta y Coquimbo. Para el estudio consideraron los alumnos ingresados entre 2000 y 2013, obteniendo un total de 9.195 registros. El objetivo era comprobar dos hipótesis: primero, la información que presenta el alumno al momento de ingresar a la universidad, es determinante en el momento de la deserción y segundo, la situación económica del alumno es influyente en el abandono académico. Para ello se utilizaron tres técnicas: redes bayesianas, redes neuronales y árboles de decisión. Los resultados fueron favorables y permitieron demostrar las dos hipótesis planteadas.

2.4. Factores influyentes en el fenómeno de la deserción

En las secciones anteriores se pudo observar que no existe un consenso en la literatura de cuáles son las variables significativas al momento de predecir la deserción académica y tampoco en cuanto a la técnica correcta a emplear. Esto se debe a que todos los conjuntos de datos difieren entre instituciones académicas, siendo distintos en cantidad, calidad y el tipo de información presente. Además, existen instituciones que solo poseen información al término de cada periodo académico, observandose en trabajos de: Delen (2010); Jia and Mareboyana (2013); Miranda and Guzmán (2017) y, por otro lado, otras que logran generar información durante el transcurso del periodo académico del estudiante, mediante de la implementación de cuestionarios (Márquez-Vera et al., 2016) o la obtención de las notas parciales referidas a una asignatura en particular (Celis et al., 2015).

El determinar cuáles son las variables determinantes en el momento de la deserción es algo complejo de estudiar e identificar, debido a que son muchos los factores

que influyen. Riobóo and Pedroza (2018) explican que la deserción es un fenómeno multicausal, el cual involucra distintos factores y espacios temporales que intervienen en el proceso de enseñanza y aprendizaje. Entre estas causas coexisten determinantes personales, sociales e institucionales, entre otras.

Capítulo 3

Minería de datos

El proceso para la detección de alumnos en riesgo académico puede ser comprendido mediante la metodología *KDD* (*Knowledge Discovery in Databases*, en Inglés). Se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información, según Han et al. (2011), con el propósito de convertir los datos originales en información consistente, que pueda ser utilizada para la toma de decisiones.

En la Figura 3.1 se puede observar las distintas etapas de la metodología *KDD* (Han et al., 2011).

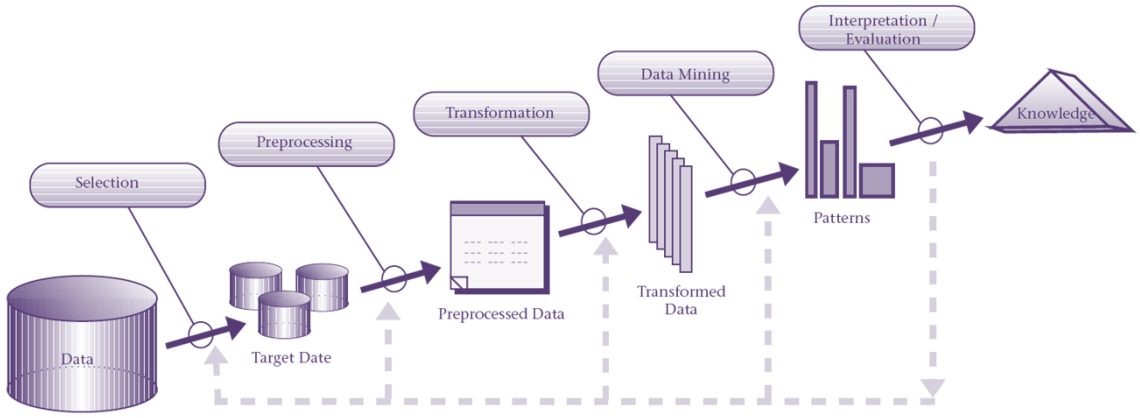

Figura 3.1: Proceso de extracción de conocimiento *KDD* (Han et al., 2011).

Etapas del proceso *KDD*

1. Selección de los datos: en esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

2. Pre-procesamiento de los datos: esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

3. Transformación de los datos: consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

4. Minería de datos: es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u ocultos en los datos.

5. Evaluación e interpretación: se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Técnicas de clasificación de minería de datos

En esta sección se explican las técnicas a emplear en la fase de minería de datos en el proceso *KDD* para la construcción de un modelo de clasificación que ayude a la detección temprana de estudiantes en riesgo académico. Para este trabajo se considerarán las siguientes técnicas: regresión logística, máquinas de vectores de soporte y redes neuronales.

El proceso de clasificación es la tarea de asignar una clase a los objetos de una entrada dada. Esta considera las características del individuo, generando un modelo que permite obtener la clase de salida para un individuo, y así poder detectarlo como posible éxito o alumno en riesgo académico.

3.1. Clasificación logística

Las funciones logísticas son adecuadas cuando se pretende hacer una clasificación binaria y es una generalización de la regresión lineal, ya que la esta última no permite modelar una variable discreta. Por lo tanto, en lugar de la predicción de una estimación puntual del evento en sí, se construye el modelo para predecir la probabilidades de su ocurrencia (Delen, 2010).

Delen (2010), Celis et al. (2015) y Lama et al. (2017) desarrollaron trabajos relacionados con la implementación de la clasificación logística. Este modelo de regresión implica la obtención de la probabilidad de que una observación pertenezca a un grupo determinado, en función del comportamiento de las variables independientes.

El modelo se formula como sigue:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{3.1}$$

donde, $(\frac{p}{1-p})$ es la posibilidad de ocurrencia (*odds ratio*, en Inglés) o la razón de oportunidad y el $\ln(\frac{p}{1-p})$ es la transformación logarítmica de la razón, la cual toma valores entre $[0, 1]$ (Lama et al., 2017). Por otra parte, p representa la variable respuesta, en este caso el posible riesgo o éxito académico; $X_1, \ldots, X_n$ las variables independientes seleccionadas mediante el proceso de selección de atributos; β_n los parámetros del modelo.

En resumen, el modelo de clasificación logística determina la probabilidad de que un individuo pertenezca a un grupo (o en términos prácticos, la probabilidad de que la variable Y tome el valor 1), dependiendo de los valores concretos que tomen las variables $X_1, \ldots, X_k$.

Es importante la selección de atributos para la construcción del modelo, su objetivo es encontrar un subconjunto de atributos con el mayor poder predictivo, evitando ruido en el fase de entrenamiento. Celis et al. (2015) describe y utiliza alguno de los enfoques más utilizados:

- Selección de atributo hacia adelante (*Forward Feature Selection*, en Inglés): se comienza sin atributos en el modelo, se agregan una a una las variables y se evalúan bajo cierta métrica el desempeño de agregar cada variable, eligiendo la que mejore más el rendimiento. Este proceso se repite hasta que ninguna variable mejora el modelo al ser agregada.

- Selección de atributo hacia atrás (*Backward Feature Selection*, en Inglés): en este enfoque se comienza con todos los atributos, luego se avalúa la eliminación de cada

variable, eliminando la que muestre un aumento en el desempeño del modelo. El proceso se repite hasta que no se observe ninguna mejora.

3.2. Máquinas de vectores de soporte

Las máquinas de vectores de soporte (*Support Vector Machines*, en inglés), son un modelo de clasificación propuesto por Vapnik (1998). Originalmente fueron pensadas para resolver problemas de clasificación binaria, pero en la actualidad son utilizadas para resolver multiclasificación (Suárez, 2014).

SVM busca minimizar el error cuadrático de la clasificación, construyendo un hiperplano que separe un conjunto de datos con una alta dimensionalidad, de la forma más precisa posible. La idea es seleccionar un hiperplano óptimo de separación que sea equidistante de las clases y, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir los vectores de soporte, sólo se consideran aquellos de cada clase que caen justo en la frontera de dichos márgenes, llamados vectores de soporte (Suárez, 2014).

Las *SVM* de clasificación binaria que discriminan puntos de datos de dos categorías posibles, están representados por un vector n-dimensional, cada uno de estos puntos de datos pertenece a solo una de las dos clases y un clasificador lineal los separa con un hiperplano, como se muestra en la Figura 3.2 (b) donde se observan muchos clasificadores lineales que logran separar de manera correcta estas dos clases para un caso ejemplo bidimensional o $\Re^2$ (Yu and Kim, 2012).

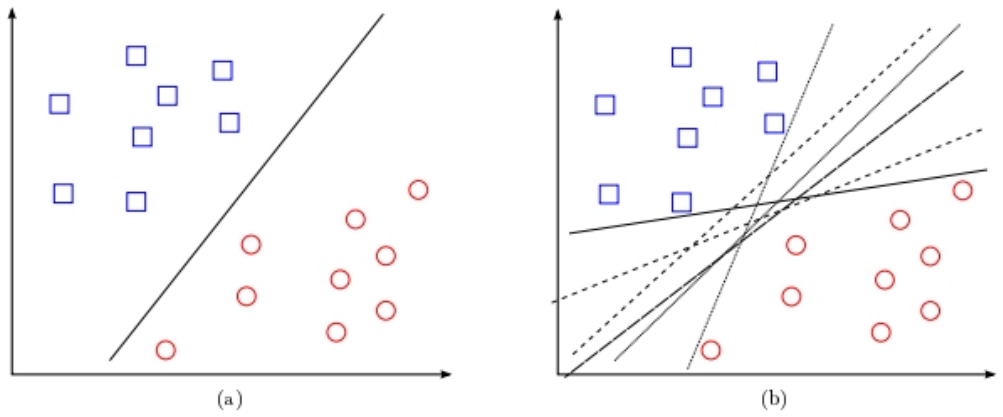

Figura 3.2: Hiperplanos de separación en un espacio bidimensional: (a) Un hiperplano de separación (b) Infinitos hiperplanos existentes de separación (Suárez, 2014).

En la práctica, poder encontrar casos en que se pueda utilizar un clasificador lineal en el espacio de entrada, como se muestra en la Figura 3.2, es muy complejo debido a la alta dimensionalidad que tienen los conjuntos de datos; en los cuales se emplean las *SVM*.

Dos grandes ventajas de las *SVM*, es que puede generar un modelo con una amplia generalización a través de la maximización del margen y el aprendizaje eficiente de problemas no lineales a través de la función kernel (Yu and Kim, 2012).

A continuación, se describe el enfoque para funciones mediante un clasificador lineal, para luego extender el método a funciones no lineales de clasificación.

Clasificación lineal para problemas linealmente separables

Dado un conjunto de ejemplos $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, donde $x_i \in \Re^d$ e $y_i \in \{+1, -1\}$, se puede definir un hiperplano de separación (Figura 3.2) como una función lineal:

$$D(x) \quad = \quad (w_1 x_1 + \ldots + w_d x_d) + b \quad = \quad < w, x > + b \quad ; \tag{3.2}$$

donde, w y b son coeficientes reales. El hiperplano de separación cumplirá la siguiente restricción para todo x_i del conjunto de ejemplos:

$$y_i D(x_i) \geqslant 0 \quad , i = 1, \ldots, n \quad . \tag{3.3}$$

Como se puede observar en la Figura 3.2 (b), existen infinitos hiperplanos que permiten separar las clases, cumpliendo todos con la restricción en (3.4), entonces para poder identificar el hiperplano óptimo de clasificación, se debe definir antes el concepto de margen de un hiperplano, denotado por τ, la cuál es la mínima distancia entre el hiperplano y el ejemplo más cercano de cualquiera de las dos clases, como se observa en la Figura 3.3 (a). En donde el objetivo es encontrar un hiperplano de separación que equidiste de los ejemplos más cercanos para cada clase, de esta forma, conseguir lo que se denomina un margen máximo, como se puede ver en la Figura 3.3 (b) (Suárez, 2014).

La distancia entre un hiperplano $D(x)$ y un punto x_i cualquiera viene dado por:

$$\frac{\mid D(x_i) \mid}{\parallel w \parallel} \quad . \tag{3.4}$$

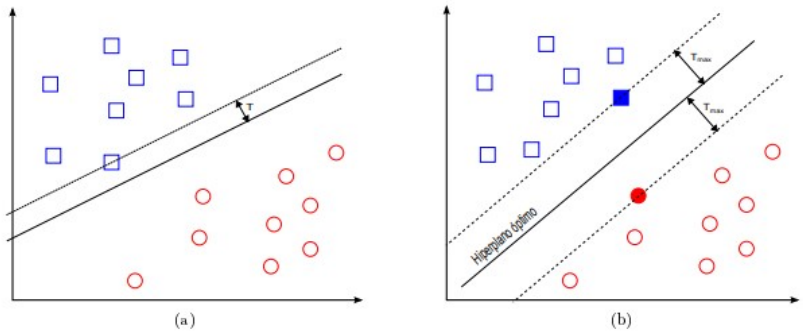

Figura 3.3: Margen de un hiperplano de separación: (a) no-óptimo y (b) óptimo (Suárez, 2014).

Utilizando las expresiones (3.4) y (3.5), todo los ejemplos de entrenamiento cumplirán con lo siguiente:

$$\frac{y_i D(x_i)}{\| w \|} \geqslant \tau \quad , i = 1, \ldots, n \quad ; \tag{3.5}$$

o también,

$$y_i D(x_i) \geqslant \tau \parallel w \parallel \quad , i = 1, \ldots, n \quad . \tag{3.6}$$

la escala del producto de τ y la norma de w se fija, de forma arbitraria, a la unidad, es decir:

$$\tau \parallel w \parallel = 1 \quad . \tag{3.7}$$

Se observa en (3.8), que el aumentar el margen τ es equivalente a disminuir la norma w, obteniendo finalemente que el margen es igual a:

$$\tau = \frac{1}{\| w \|} \quad . \tag{3.8}$$

Por tanto, de acuerdo a su definición, un hiperplano de separación óptimo (Figura 3.4) será aquel que posee un margen máximo y, por tanto, un valor mínimo de $\| w \|$. Además, está sujeto a la restricción en las ecuaciones (3.7) y (3.8), es decir:

$$y_i D(x_i) \geqslant 1 \quad , i = 1, \ldots, n \quad . \tag{3.9}$$

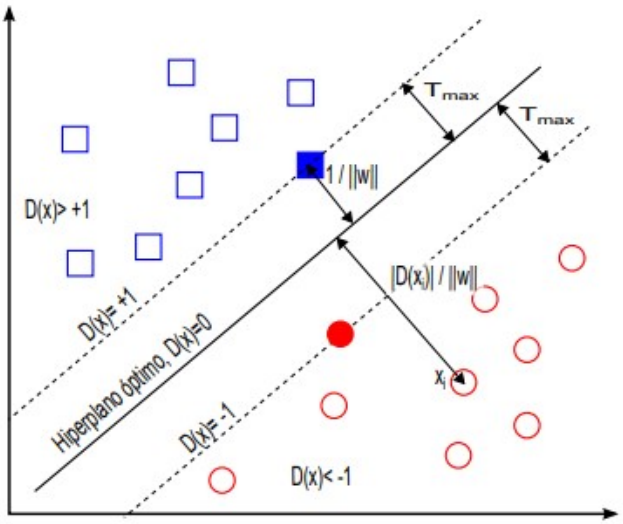

Figura 3.4: Hiperplano óptimo para un caso linealmente separable con margen máximo (Suárez, 2014).

El concepto de margen máximo está relacionado directamente con la capacidad de generalización del hiperplano de separación, de tal forma que, a mayor margen, existirá entre las dos clases mayor distancia de separación. Los ejemplos que están situados a ambos lados del hiperplano óptimo, en la Figura 3.4, azul y rojo, respectivamente, son aquellos que para (3.10) es una igualdad y toman el valor de $\tau = 1$, estos reciben el nombre de vectores soporte, debido a que son los puntos más cercanos al hiperplano de separación, siendo los más difíciles de clasificar, y además son los puntos que se consideran al momento de construir el hiperplano óptimo de separación.

Finalmente, el problema de maximización del margen se puede formular como un problema de optimización, en donde se busca minimizar con respecto a w y b. Como se expresa en (3.11), sujeta a la restricción de (3.10).

$$\acute{min} \quad f(w) = \frac{1}{2} \parallel w \parallel^2 \quad . \tag{3.10}$$

La teoría establece que un problema de optimización, denominado primal, tiene una forma dual si la función a optimizar y las restricciones son funciones estrictamente convexas y, en estas circunstancias, resolver el problema dual permitirá obtener la solución del problema primal. Así, puede demostrarse que el problema de optimización dado en (3.11), satisface el criterio de convexidad y, por tanto, tiene un dual (Suárez, 2014).

Para resolver el problema de optimización, se introducen la técnica del multiplicador de Lagrange, utilizando la siguiente función langrangiana:

$$L(w, b, \alpha) = \frac{1}{2} \parallel w \parallel^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(w^t x_i + b) - 1 \right] \quad , \tag{3.11}$$

donde, los $\alpha_i \geqslant 0$ son los denominados multiplicadores de Lagrange.

Flores et al. (2015), definen el problema dual como:

$$máx \;\; L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \quad , \tag{3.12}$$

donde, $\alpha_i \geqslant 0$ y $\sum_{i=1}^{n} \alpha_i y_i = 0$

Los multiplicadores que cumplen con $\alpha_i > 0$, son llamados vectores de soporte; ya que, son los que participan en la construcción del hiperplano de clasificación. Además, se tiene que $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$ y $b^* = y_i - w^* x_i$ para cada vector de soporte x_i. La función de clasificación para un problema linealmente separable, queda expresada como:

$$f(x) = \; y_i(w^* \times x + b^*) \; = \; y_i(\sum_{i=1}^{n} y_i \alpha_i^* (x \times x_i) + b^*) \, , \tag{3.13}$$

donde, (x_i, y_i) representa la dupla de cualquier vector de soporte (Flores et al., 2015).

Clasificación lineal para problemas linealmente no separables

El problema planteado en la sección anterior, para el caso linealmente separable, no suele suceder habitualmente en problemas reales; ya que, estos se caracterizan por tener ejemplos ruidosos y con alta dimensionalidad. Se entiende como no separable, cuando un punto o ejemplo, se encuentra dentro de la frontera de decisión o margen, y es clasificado correctamente (círculo rojo), y en otro caso, cuando el punto cae al otro lado de la frontera (cuadrado azul) y es clasificado incorrectamente, como se muestra en la Figura 3.5.

Desde el punto de vista de lo expresado en (3.10), el que no se cumpla dicha restricción, es porque se presenta un caso linealmente no separable. La solución a este problema, es explicado a continuación, incluyendo un nuevo parámetro a la formulación hecha anteriormente, llamada variable de holgura ξ_i, que permitirá cuantificar el número de ejemplos no separables que se esta dispuesto a admitir. Se obtiene una nueva fórmula, a partir de (3.10) con la inclusión de este nuevo parámetro ξ_i, quedando de la siguiente manera:

$$y_i D(x_i) \geqslant 1 - \xi_i \quad , \xi_i \geqslant 0; \;\; i = 1, \ldots, n \quad . \tag{3.14}$$

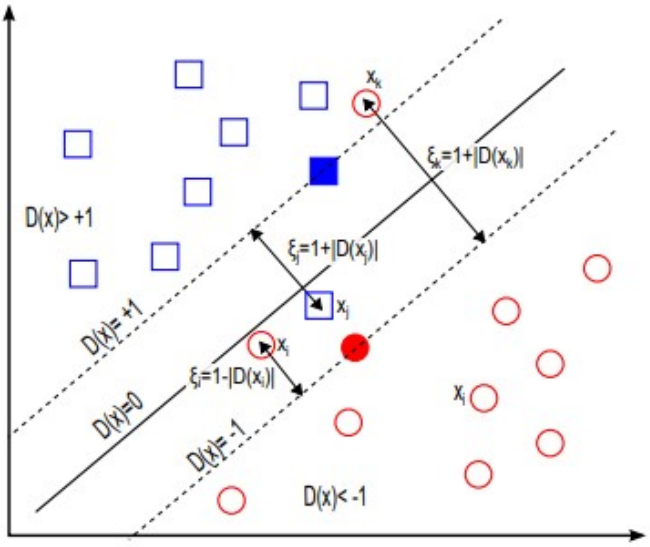

Figura 3.5: Caso de ejemplo linealmente no separables (Suárez, 2014).

Por lo tanto, el nuevo problema de optimización, queda dado por la siguiente expresión:

$$m\acute{i}n \quad f(w,\xi) = \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} \xi_i \quad , \tag{3.15}$$

donde, C es una constante, que permite controlar en qué grado influye el término del coste de ejemplos no separables en la minimización del problema, es decir, permitirá regular el compromiso entre el grado de sobreajuste del clasificador final y la proporción del número de ejemplos no separables (Suárez, 2014). Así, un valor de C muy grande permitiría valores de ξ_i muy pequeños. Caso contrario, valores de ξ_i muy grandes, permitiría un número elevado de ejemplos mal clasificados.

En consecuencia, el nuevo problema de optimización consiste en encontrar el hiperplano, definido por w y b, que minimiza a (3.16) con respecto a w y b, considerando las restricciones planteadadas en (3.15).

Finalmente, la función de clasificación se mantiene igual que para el caso anterior:

$$f(x) = \ y_i(w^* \times x + b^*) \ = \ y_i(\sum_{i=1}^{n} y_i \alpha_i^* (x \times x_i) + b^*) \, , \tag{3.16}$$

donde, $0 < \alpha_i < C$.

Clasificación no Lineal

Anteriormente, se mostró que los hiperplanos de separación son capaces de poder discriminar puntos en dos grupos de clasificación, solo cuando los ejemplos son perfectamente separables o casi separables, mediante una función lineal. A continuación, se describirá cómo clasificar un conjunto de datos más complejo, cuando no es suficiente una función lineal para poder discriminar ejemplos de dos clases en el espacio original.

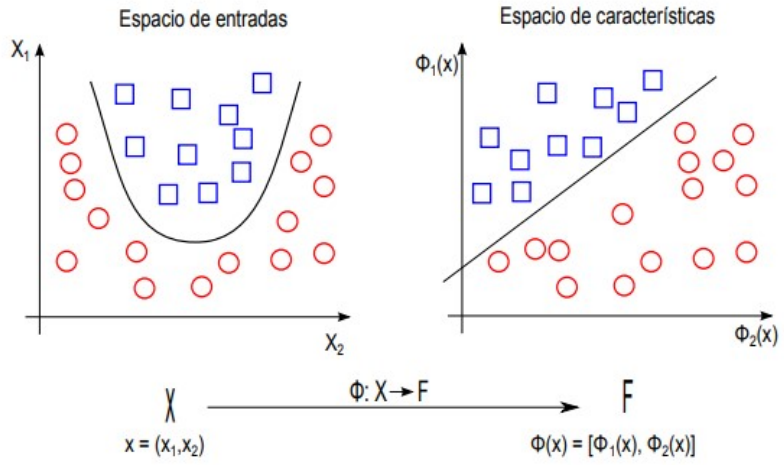

Figura 3.6: Problema de clasificación no lineal de separación, mediante la función Kernel (Suárez, 2014)

Para el caso de una clasificación mediante una función no lineal (véase Figura 3.6), *SVM* proyecta el conjunto de datos de entrada a un espacio de mayor dimensión $\Re^+$, por ejemplo un espacio en $\Re^2$ llevado a $\Re^3$, utilizando una función que transforma el espacio de entrada $X \to \phi(X) \in \Re^+$.

A cada uno de estos espacios se le denomina espacio de características, para diferenciarlo del espacio de ejemplos de entrada, de manera formal se puede escribir como:

Sea $\Phi : X \to F$ la función de transformación que hace corresponder cada vector de entrada x, con un nuevo punto en el espacio de características F. De manera formal:

$$\Phi(x) = [\phi_1(x), \ldots, \phi_m(x)] \quad , \tag{3.17}$$

donde, $\Phi_i(x)$ es una función no lineal.

Como se explicó anteriormente, la idea es poder construir un hiperplano de separación lineal en este nuevo espacio de características y transformándolo en una función no lineal en el espacio original de entrada (Figura 3.6).

En este contexto, la función de decisión planteada en (3.3), en el nuevo espacio de características queda definida por:

$$D(x) \quad = \quad (w_1\phi_1(x) + \ldots + w_d\phi_d(x)) \quad = \quad < w, \phi(x) > \quad , \tag{3.18}$$

$$K(x, x') \; = < \Phi(x), \Phi(x') > = < \phi_1(x)\phi_1(x') + \ldots + \phi_n(x)\phi_n(x') > \quad . \tag{3.19}$$

La selección del mejor kernel para una aplicación es todavía un tema de investigación (Ali and Smith-Miles, 2006). El procedimiento más común se realiza bajo un proceso de metaaprendizaje, en el cual se evalúan los resultados de clasificacion al utilizar distintos Kernel, en otras palabras se emplea el ensayo y error, como se hace en el caso de la selección del mejor modelo.

En un estudio empírico, Ali and Smith-Miles (2006) recopilaron 112 diferentes problemas de clasificación, utilizando distintas funciones Kernel. Lograron identificar que la función RBF o gaussiana es la que logra los mejores resultados de clasificación.

Dentro de las distintas funciones de Kernel existentes, los más utilizados son:

1. Kernel Lineal: $K(x, x') = < x, x' >$.
2. Kernel Polynomial: $K_p(x, x') = [< x, x' > +1]^p$.
3. Kernel Gaussiano: $K(x, x') = exp(-\|x - x'\|^2)$.

A continuación se explicará una forma de evaluación de los atributos relevantes dentro de las *SVM*, a partir de la ganancia de la información:

Ganancia de información: Este es un método de clasificación de atributos que introduce el concepto de la entropía, propuesta por Quinlan (1986) la cual es una medida de incertidumbre o de desorden, y es usado para ayudar a decidir entre distintos atributos, cuál es el más relevante respecto a la clase. En general, un atributo que puede ayudar a discriminar más objetos, tiende a reducir más la entropía. En otras palabras, mide cuanto ayuda el conocer el valor de una variable X para conocer el verdadero valor de una variable Y. En el caso estudiado, X es un atributo, mientras que Y es la clase. Una alta ganancia de información del atributo X permite reducir la incertidumbre de la clasificación.

Para calcular la entropía de la clases se utiliza la fórmula:

$$Entropía(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad , \tag{3.20}$$

donde:
S: es el conjunto de la clase.
c: es el número de clases.
p_i: es la proporción de cada clase dentro del conjunto de datos.

En el caso particular de una clasificación binaria (ejemplo: positivo/negativo), la fórmula anterior queda como:

$$Entropía(S) = \sum_{i=1}^{2} -P \log_2 P - N \log_2 N \quad , \tag{3.21}$$

donde:
S: es el conjunto de la clase.
P: es la proporción de positivos dentro del conjunto de la clase.
N: es la proporción de negativos dentro del conjunto de la clase.

Para determinar los atributos relevantes, se introduce el concepto de ganancia de información, esta es una medida de discriminación de los atributos, que se describe en la siguiente ecuación:

$$Entropía(C) = -\sum_{c \in \mathbb{C}} p(c) log_2 p(c) \quad , \tag{3.22}$$

$$Entropía(C|A) = -\sum_{a \in \mathbb{A}} p(a) \sum_{c \in \mathbb{C}} p(c|a) log_2 p(c|a) \quad , \tag{3.23}$$

$$GanInf(C, A) = Entropía(C) - Entropía(C|A) \quad , \tag{3.24}$$

donde:
C: es el conjunto de la clase.
c: es el número de clases.
A: son el conjunto del atributo.
a: son los posibles valores que puede tomar el atributo.
$(c|a)$: proporción de el valor a en la clase c.

Mientras más alta sea la ganancia de la información de un atributo respecto a la clase, este será considerado más relevante dentro del conjunto de atributos.

Por ejemplo, se desea hacer una selección de atributos entre A y B. Para ver cuál es más relevante, *Gan Inf*(C;A,B); *Gan Inf*(C,A) > *Gan Inf*(C,B), entonces el atributo A es considerado más relevante que el atributo B, dado que reduce la incertidumbre de la clasificación.

3.3. Redes neuronales artificiales

El cerebro humano está formado por millones de neuronas que se conectan entre sí, transmitiendo información entre ellas. Luego de que ésta procesa, se genera una respuesta en función del estímulo recibido por otra neurona. Las redes neuronales

artificiales (*Artificial neural networks*, en Inglés), son el resultado de los intentos por reproducir el funcionamiento del cerebro humano mediante computadoras, obteniendo un modelo abstracto y simple de una neurona artificial, este es el elemento básico del procesamiento en una red neuronal artificial (McCulloch and Pitts, 1943).

Como parte de este trabajo, se considera importante comenzar describiendo cómo funcionan las neuronas biológicas, para luego describir el comportamiento de las *ANN*. El cerebro humano continuamente recibe estímulos de entrada de muchas fuentes y las procesa a manera de crear una apropiada respuesta de salida. Las neuronas son las células que forman la corteza cerebral de los seres vivos, cada una está formada por elementos llamados cuerpo, axón y dendritas, como se muestra en la siguiente figura:

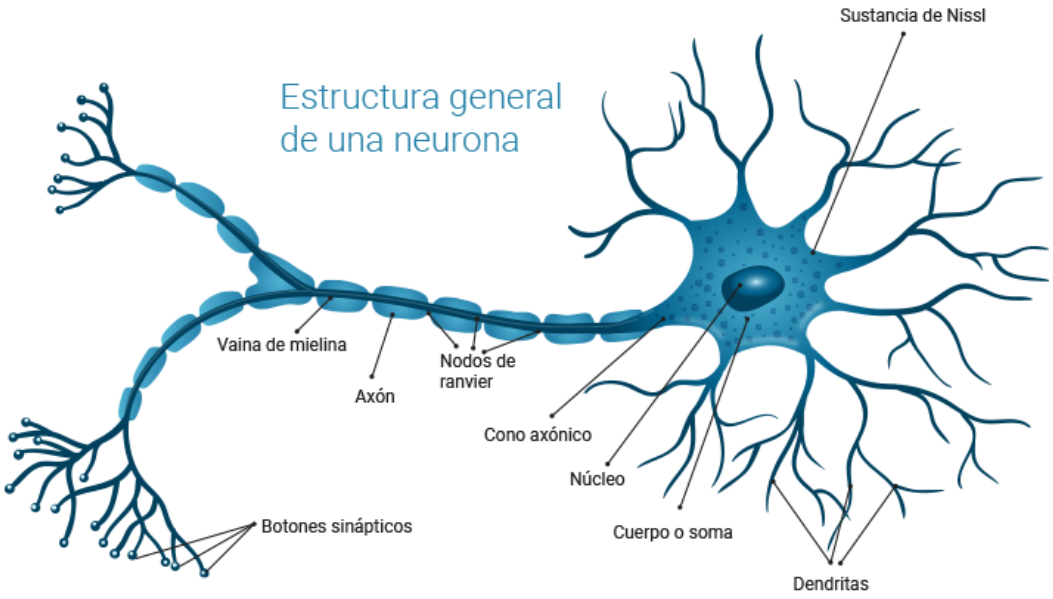

Figura 3.7: Estructura de una neurona biológica (Soberanis, 2013).

Cada neurona recibe impulsos eléctricos por medio de las dendritas, que se transmiten a lo largo de una parte de la célula muy alargada llamada axón, al final del axón este se ramifica hasta conectarse con otras neuronas por medio de sus dendritas.

Estructura de una Red Neuronal Artificial

Las redes neuronales están formadas por una serie de capas de neuronas que están unidas entre si mediante sinapsis. Las neuronas artificiales como unidades independientes no son muy eficaces para el tratamiento de la información y se agrupan en estructuras más grandes. A continuación se observan las estructuras más comunes que forman las neuronas para la formación de una *ANN*. Para esto entenderemos el número y la forma de interconexión de las capas.

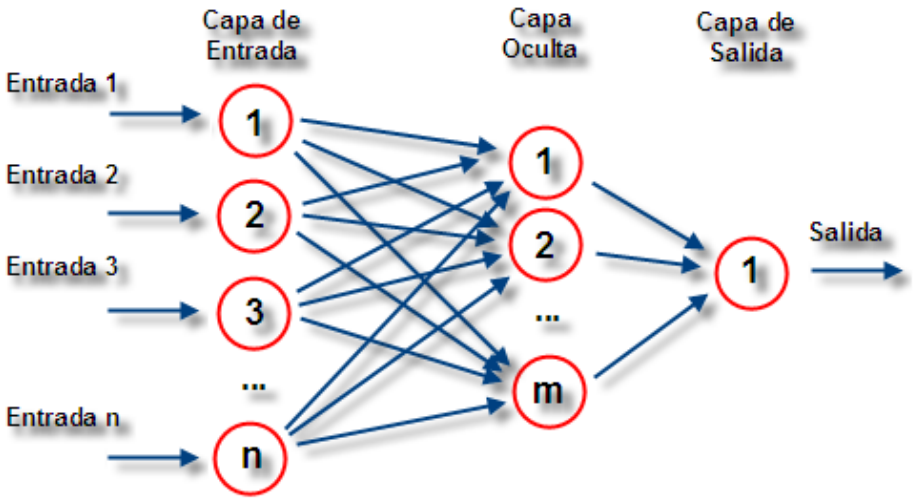

Figura 3.8: Red neuronal artificial perceptrón multicapa (Anónimo, 2015).

En la Figura 3.8, se puede observar una *ANN* perceptrón multicapa, con n neuronas de entrada, m neuronas en su capa oculta y una neurona de salida.

Niveles de neuronas:

La distribución de neuronas dentro de la red se realiza formando niveles o capas de un número determinado de neuronas cada una. A partir de su situación dentro de la red, se pueden distinguir tres tipos de capas:

- **Capa de entrada:** recibe la información bruta entregada desde el exterior por un vector $X = (x_1, x_2, \cdots, x_n)^\top$.
- **Capa oculta:** ésta se encarga de recibir, procesar y memorizar la información entregada por la capa anterior, en una función Y.
- **Capa de salida:** esta contiene la información de respuesta de la red, dando un resultado final a partir de un umbral de activación ϕ.

Forma de conexión de las capas:

Las neuronas se conectan unas a las otras usando sinapsis, esto se puede observar en las uniones a nivel de capas, que forman distintas estructuras, entre las cuales se puede distinguir:

- **Unión todos con todos:** Consiste en unir cada neurona de una capa con todas las neuronas de la otra capa. Este tipo de conexión es el más usado en las redes neuronales.
- **Unión lineal:** Consiste en unir cada neurona con otra neurona de la otra capa. Este tipo de unión se usa menos que el anterior y suele usarse para unir la capa de entrada con la capa de procesamiento.
- **Predeterminado:** Este tipo de conexión aparece en redes que tienen la propiedad de agregar o eliminar neuronas de sus capas y de eliminar también conexiones.

Teniendo un orden en las capas se establece el tipo de conexión entre éstas, esto sirve para clasificar las *ANN* en redes de propagación hacia delante (*feedforward*, en Inglés) o redes recurrentes hacia atrás (*feedback*, en Inglés). Las conexiones laterales son entre neuronas dentro de la misma capa, son llamadas redes monocapa y si la red admite que las neuronas estén unidas así mismas, se dice que son redes autorecurrente.

Funciones de una red neuronal artificial

Las *ANN* consideran a las neuronas como una serie de funciones que se componen entre ellas, siendo los resultados de una de ellas, los parámetros de otra. De este modo, la función de propagación considera los valores que le llegan desde las entradas y los pesos de las sinapsis, para posteriormente transformarlos dentro de la función de activación y, por medio de la interacción de estas funciones, se procesa la información final.

A continuación, se explicará con más detalle las diferentes funciones que influyen en la construcción del modelo.

Función de propagación o ponderación

Ésta calcula el valor de base o entrada total de información a la neurona, mediante la suma ponderada de sus n señales de entradas recibidas en X_j, con $j = 1, 2, \cdots, n$; es decir, de las entradas (x_{i1}, x_{i2}, ..., x_{in}) multiplicadas por el peso (w_{i1}, w_{i2}, ..., w_{in}). La función de propagación proporciona el valor del potencial postsináptico de la neurona i-ésima en función de sus pesos w_i y entradas x_i. Esto se ilustra en Figura 3.9:

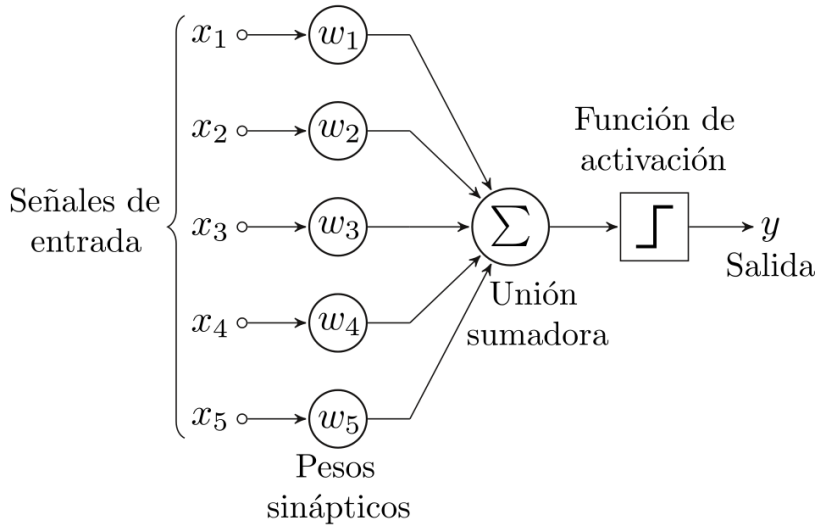

Figura 3.9: Función de propagación hacia una unión sumadora (Ruiz, 2015)

Función de activación

Es la característica principal o definitoria de las neuronas; es decir, la que mejor define el comportamiento de la misma. Se usan diferentes tipos de funciones, tanto lineales como no lineales. Con éstas se pretende calcular el nivel o estado de activación de la neurona en función de la entrada global.

La función de activación calcula el estado de actividad de una neurona, transformando la entrada global (menos un umbral ϕ_i) en un valor o estado de activación, cuyo rango está comprendido entre [0, 1] o de [-1, 1]. Una neurona puede estar totalmente inactiva $\{0, -1\}$ o activa cuando toma el valor $\{1\}$.

Las funciones de activación más utilizadas se detallan a continuación:

- **Función lineal:** esta función es útil en aquellos casos en que el valor de salida es una variable continua o en aquellos en que se desea que la red aprenda los eventos menos frecuentes. A diferencia de la función sigmoide, la lineal no se hace menos sensible al alejarse de cero.

$$f(y) = \begin{cases} -1 & y \leqslant \frac{-1}{a} \\ a \times y & \frac{-1}{a} < y < \frac{1}{a} \\ 1 & y \geq \frac{1}{a} \end{cases} \quad , y = \sum x_{ij} \times w_{ij} - \phi \quad , a > 0 \quad . \tag{3.25}$$

- **Función logística o sigmoide:** es una de las funciones más comunes, los valores de salida que proporciona esta función están comprendidos dentro de un rango

que va de 0 a 1. Al modificar el valor de g se ve afectada la pendiente de la función de activación.

$$f(y) = \frac{1}{1 + \exp^{-g \times y}} \quad , y = \sum x_{ij} \times w_{ij} - \phi \quad . \tag{3.26}$$

- **Función tangente hiperbólica:** esta función es la más usada para redes neuronales binarias; ya que, no es lineal y es muy simple. Tiene las mismas propiedades que la logística, sin embargo, el rango de salida de esta función permite respuestas simétricas (-1, 1); manteniendo una intermedia en cero. Esta función suele converger antes que la función logística.

$$f(y) = \frac{exp^{g \times y} - exp^{-g \times y}}{exp^{g \times y} + \exp^{-g \times y}} \quad , y = \sum x_{ij} \times w_{ij} - \phi \quad . \tag{3.27}$$

Función de salida

Si la función de activación está por debajo de un umbral determinado ξ_i, ninguna salida se pasa a la neurona subsiguiente. Normalmente, no cualquier valor es permitido como una entrada para una neurona, por lo tanto, los valores de salida están comprendidos en el rango [0, 1] o [-1, 1] en el caso de los reales. También pueden ser binarios $\{0, 1\}$ o $\{-1, 1\}$

Capítulo 4

Medidas de rendimiento de los clasificadores

Las medidas de evaluación son muy importantes, ya que permiten evaluar el rendimiento de los clasificadores y guiar los algoritmos de aprendizaje en el proceso *KDD*. Phung et al. (2009) explican que las métricas no valoran la clase minoritaria, entonces los algoritmos de aprendizaje no pueden manejar de manera adecuada el problema de desequilibrio entre clases.

Para evaluar el rendimiento, la métrica comúnmente utilizada es la tasa de clasificación general, es decir la exactitud (*Accuracy*, en Inglés). Sin embargo, en un conjunto de datos desequilibrado, la tasa de clasificación general deja de ser una medida adecuada; ya que, la clase pequeña tiene menos efecto sobre la precisión en comparación con la clase mayoritaria. Phung et al. (2009) muestran el desarrollado de otras métricas para evaluar el rendimiento de los algoritmos de clasificación para conjuntos de datos desequilibrados.

Finalmente, para poder entender como se evalúa el rendimiento de los modelos, se debe entender algunos conceptos como: matriz de confusión, medidas de eficiencia y validación cruzada, que son explicadas a continuación.

4.1. Matriz de confusión

Esta contiene información de las instancias de la clase verdadera y de la predicción realizada, el clasificador predice la clase para cada instancia, si está correcto es contada como éxito y en caso contrario como error. A partir de esta matriz de tamaño $n \times n$, donde n corresponde al número de clases, se calculan diversas métricas que sirven para evaluar el rendimiento de los modelos empleados.

Matriz de confusión	**Valor real +**	**Valor real -**
Predicción +	*TP*	*FP*
Predicción -	*FN*	*TN*

Tabla 4.1: Matriz de confusión.

La Tabla 4.1 representa las posiciones que muestra una matriz de confusión de dos clases, donde:

- TP (*True Positive*, en Inglés): verdaderos positivos son los casos que pertenecen a la clase positiva y que el modelo los clasificó correctamente.
- TN (*True Negative*, en Inglés): verdaderos negativos son los casos que pertenecen a la clase negativa y el modelo los clasificó correctamente.
- FP (*False Positive*, en Inglés): falsos positivos son los casos que no pertenecen a la clase positiva y que el modelo los clasificó incorrectamente.
- FN (*False Negative*, en Inglés): falsos negativos son los casos que no pertenecen a la clase negativa y que el modelo los clasificó incorrectamente.

4.2. Medidas de eficiencia

En esta sección se definen las fórmulas principales utilizadas para medir el rendimiento de los clasificadores, calculadas a partir de la matriz de confusión (Phung et al., 2009):

- Exactitud: es el porcentaje de predicciones correcta respecto del total, se determina con la siguiente expresión:

$$Exactitud = \frac{TP + TN}{TP + FP + FN + TN} \, . \tag{4.1}$$

- Precisión: es el porcentaje de predicciones positivas realizadas por el clasificador que son correctas.

$$Precisión = \frac{TP}{TP + FP} \, . \tag{4.2}$$

- Sensibilidad: también llamada sensibilidad del modelo, es el porcentaje de la clase verdadero, que el algoritmo clasifica correctamente como verdadero.

$$Sensibilidad = \frac{TP}{TP + FN} \, . \tag{4.3}$$

- Especificidad: es el porcentaje de la clase negativo, que el algoritmo clasifica correctamente como negativo.

$$Especificidad = \frac{TN}{TN + FP} . \tag{4.4}$$

- F-1: se define como la media armónica de *sensibilidad* y *precisión*. Un alto valor de *F-1* significa un alto valor para estas dos métricas.

$$F - 1 = \frac{2 * Sensibilidad * Precisión}{Sensibilidad + Precisión} . \tag{4.5}$$

4.3. Validación cruzada

Permite evaluar los resultados de un análisis estadístico y garantizar que son independientes la partición entre datos de entrenamiento y prueba en un conjunto de datos y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones (Devijver and Kittler, 1982). Consiste en la partición del conjunto de datos en k partes del mismo tamaño. Durante cada ejecución, una de las particiones se elige para la prueba, mientras que el resto de ellos se utiliza para el entrenamiento, esto se repite "k"veces hasta que cada partición se utilice como prueba (Tan et al., 2006).

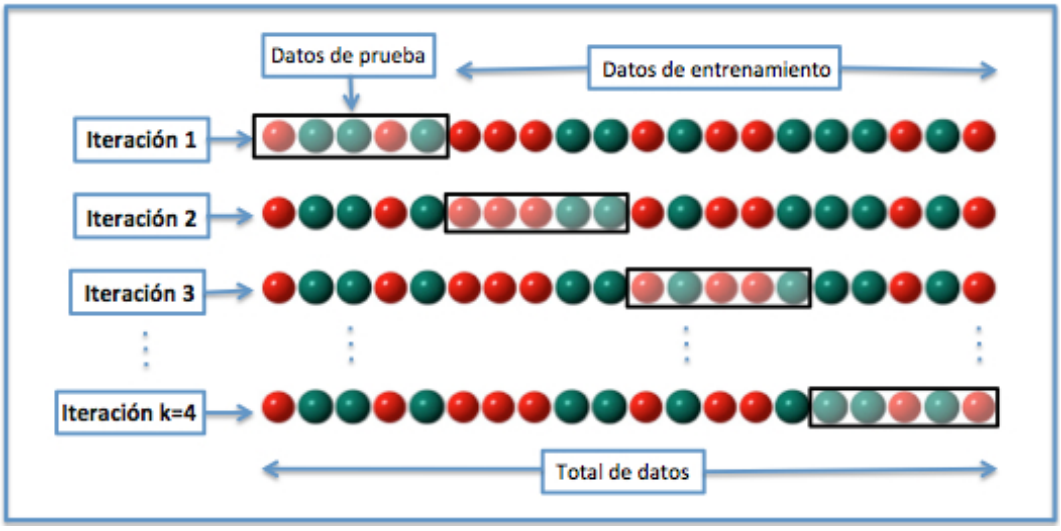

Figura 4.1: Validación cruzada para k=4 (Tan et al., 2006).

De lo anterior, se desprende la siguiente fórmula para el cálculo final de rendimiento de un clasificador:

$$VD_k = \frac{1}{k} \sum_{i=1}^{k} VD_i \quad , \tag{4.6}$$

donde k, corresponde a la cantidad de iteraciones y subconjuntos, una de estas particiones se utiliza como prueba y el resto $(k-1)$ como datos de entrenamiento. VD_k es el resultado del rendimiento para cualquier métrica empleada y VD_i el rendimiento para cada iteración. En la Figura 5.2, se presenta un ejemplo de esta situación para $k = 4$.

Capítulo 5

Aplicación y resultados

En esta sección se presenta la aplicación y resultados obtenidos de los distintos modelos de clasificación propuestos, sobre un conjunto de datos reales, el cual se lleva a cabo bajo un proceso *KDD*, explicado anteriormente en la Sección 3. Se propuso construir diferentes modelos de clasificación, con el objetivo de poder detectar tempranamente aquellos alumnos en riesgo académico, para cada periodo de tiempo en el cual se presenta nueva información de los estudiantes. De esta forma, pueden obtenerse varios modelos de predicción para distintos periodos de tiempo, logrando identificar en qué etapa se puede establecer una predicción suficientemente confiable y que además, logre una alerta temprana sobre los estudiantes.

Conjunto de datos

La primera etapa del proceso de *KDD* consiste en la selección de los datos, en donde se determina la fuente y el tipo de información a utilizar. Segundo, se realiza una etapa de pre-procesamiento de los datos, la cual conlleva a una limpieza y preparación de la información proveniente de distintas fuentes de datos; y como tercera etapa, una transformación de las variables de ser necesario. Así, finalmente los datos logran tener una estructura adecuada para su posterior análisis. Las tres etapas mencionadas anteriormente, se pueden considerar como la administración de datos en estadística, siendo una labor muy relevante, ya que, la calidad y confiabilidad de la información disponible influye directamente en los resultados que posteriormente se obtengan al aplicar los modelos de clasificación. Lo mencionado anteriormente se llevó a cabo mediante el programa estadístico Stata 14.

En este trabajo de titulación se utilizó información de los alumnos de la Universidad de Valparaíso, de las cohortes de ingreso 2010 a 2016; las cuales provienen de dos fuentes de datos: (1) información de ingreso del estudiante y (2) historial de notas de cada asignatura inscrita. La fuente de información (1), se compone de 7 conjuntos de datos, para cada año respectivamente; y la fuente (2), está compuesta por distinto número de conjuntos de datos para cada año de ingreso, por ejemplo: para el 2010 un total de 7 conjuntos de datos, correspondiente a cada periodo académico desde el 2010 hasta el 2016; para el 2011, se tienen 6 conjuntos, y así sucesivamente, hasta obtener para el 2016 solo un conjunto de datos con el historial de notas de dicha cohorte.

Para la limpieza de los datos, se trabajó con un total de 36 conjuntos de datos, provenientes de 22.692 estudiantes matriculados entre los años 2010 a 2016. Luego de haber realizado una integración de todas estas bases de datos de manera minuciosa, se procedió a una revisión de cada variable identificando diversos problemas, como por ejemplo:

- Datos duplicados: se logró evidenciar estudiantes que se encontraban con registros duplicados, ya que presentaban información en dos carreras distintas para el mismo o distinto periodo académico y de manera similar en las calificaciones de los alumnos, en el cuál no se logra entender a que carrera están asociadas ciertas calificaciones, perdiendo información como: el alumno realizó un cambio interno de carrera o se volvió a matricular en la misma. Por lo tanto, se procedió a excluir a todos los estudiantes que tengan al menos dos registros dentro de la institución.

- Datos fuera de rango: este problema se observó en distintas variables, perdiendo la mayor información en aquellas que tienen que ver con la familia; tales como: ingreso familiar, jefe de familia, quien financia los estudios, familiares vivos, entre otras; finalmente se excluyeron, ya que en algunos casos alcanzaba un 80 % de información que se desconocía.

- Datos faltantes: en algunos casos se emplearon técnicas de imputación y para otros casos en que la falta de información de la variable era muy alta, se procedió a no considerarlas, como: puntaje ranking y unidad del colegio.

- Variables relacionadas: se eliminaron variables que eran explicadas por otra variable, es decir una correlación muy alta; como era el caso de: la comuna del estudiante, siendo explicada en la región, promedio del puntaje lenguaje-matemáticas, contenido en la PSU de Lenguaje y PSU de Matemática de manera separada, y por último el promedio de enseñanza media, el cual era explicado por el puntaje NEM.

- Reducción de dimensión: este proceso consiste en disminuir el rango de una variable, ya que las opciones que puede alcanzar un atributo es muy amplio en algunos casos, y en consecuencia, esta variable deja de ser tan significativa para las predicciones. Por lo tanto al reducir el rango, esta logra aportar mayor información sobre la variables respuesta. Por ejemplo, en los datos originales se tiene la región de procedencia del estudiante, en la cual existen 15 posibles valores, por lo que se identificaron los grupos con mayor frecuencia, quedando finalmente segmentada en tres categorías: (1) Región de Valparaíso, (2) Región Metropolitana y (3) Otra región; de esta manera la variable región del estudiante aporta gran información al momento de construir los modelos de clasificación.

Construcción de variables

Por otro lado, se realizó la creación de nuevas variables para poder obtener información sobre el historial académico de los estudiantes de manera generalizada; ya que, todos poseen distintas asignaturas, logrando obtener: promedio de notas de cada año académico y número de asignaturas aprobadas y reprobadas. Estas variables se consideraron para la construcción de los modelos de clasificación, al transcurrir un año académico el estudiante. Estas variables se muestran a continuación en la Tabla 5.1.

Variables (t=1)	**Tipos de datos**	**Rango**
Promedio de notas del año	Cuantitativa-continua	[1,7]
Asignaturas aprobadas	Cuantitativa-discreta	[0,14]
Asignaturas reprobadas	Cuantitativa-discreta	[0,14]

Tabla 5.1: Nuevas variables para la construcción del modelo de clasificación al término del primer año académico.

Finalmente, la información obtenida luego de una administración de los datos, se resume a un total de 20.586 estudiantes, de los cuales tan solo 645 pertenecen a la Facultad de Ciencias de la Universidad de Valparaíso. A continuación en la Tabla 5.2, se listan las 20 variables con sus rangos respectivos, con las cuales se llevó a cabo este trabajo, siendo la información que se tiene de los estudiantes al momento de ingreso a la institución.

Variables (t=0)	Tipos de datos	Rango
vía de ingreso	Cualitativa-ordinal	{0,1}
sexo	Cualitativa-ordinal	{0,1}
carrera de ingreso	Cualitativa-ordinal	{1,2,3,4,5,6}
rama del colegio	Cualitativa-ordinal	{1,2,3}
promoción de ingreso	Cualitativa-ordinal	{1,2}
número de dependencias	Cualitativa-nominal	{1,2,3}
región de procedencia	Cualitativa-ordinal	{1,2,3}
ingreso familiar	Cualitativa-nominal	{1,2,3,4}
cobertura de salud	Cualitativa-ordinal	{1,2}
proseguir estudios	Cualitativa-ordinal	{1,2,3,4,5,6}
grupo familiar	Cualitativa-nominal	[1,11]
jefe familia	Cualitativa-ordinal	{1,2,3,4}
financiamineto	Cualitativa-ordinal	{1,2}
familia viva	Cualitativa-ordinal	{1,2,3,4}
puntaje NEM	Cuantitativa-continua	[300,850]
PSU de Lenguaje	Cuantitativa-continua	[300,850]
PSU de Matemáticas	Cuantitativa-continua	[300,850]
PSU de Ciencias	Cuantitativa-continua	[300,850]
PSU de Historia	Cuantitativa-continua	[300,850]
respuesta	Cualitativa-ordinal	{1,2}

Tabla 5.2: Tipo y rango de las variables en el conjunto de datos.

Variable respuesta

Para especificar la variable respuesta, se procedió al análisis del estado académico de los estudiantes en cada periodo, encontrando un gran problema, ya que existen un total de 27 estados que pueden ser asignado a cada estudiante en un periodo académico, no dejando claro si es un caso de deserción o si continua los estudios, por nombrar algunos ejemplos: moroso, en espera de inscripción de asignaturas, pasivo, de equivalencia, entre otros. Por lo anterior, no se puede hacer un juicio si el alumno continua o deserta de la institución respecto a los estados de cada estudiante.

Para los alumnos con estado de espera de título, egresado o ya titulados, estos son identificados de manera inmediata como éxito, siendo la mayoría estudiantes que ingresaron en los años 2010 y 2011; ya que, se tiene un periodo suficientemente largo para identificarlos. Por otro lado, se presenta el caso de alumnos que aún siguen estudiando y que no han logrado culminar su carrera, como lo son los estudiantes que ingresaron

desde el 2012 en adelante, ya que en muchos casos, estos aún permanecen como alumno en la institución y de los cuales se desconoce su estado académico. Aunque existe la posibilidad de identificar como desertores ciertos estudiantes, considerando un tiempo suficientemente largo (2 años), el cual se consideró en este trabajo para descartar la opción de que el estudiante se reincorpore según lo planteado por Himmel (2018) y clasificarlo como alumno desertor, esto sigue siendo un problema por la gran cantidad de alumnos en los cuales el criterio anterior no se puede considerar.

Con el fin de poder detectar posibles alumnos en riesgo académico y según Tinto (1989), se procedió a la construcción de una variable de avance curricular de cada alumno, como se muestra en la fórmula 5.1, de esta manera se logra penalizar los casos de estudiantes que se mantienen por varios periodos académicos, no logrando aprobar las asignaturas, y que finalmente en muchos casos son desertores de la institución.

$$Avance_i = \frac{f_i}{F_k} \quad , i = 1, 2, ..., 8 \quad , \tag{5.1}$$

donde, f_i es el total de asignaturas aprobadas al i-ésimo año, F_k es la cantidad total de materias de la carrera y $Avance_i$ es un valor entre 0 y 1 que representa el progreso del alumno que tiene en su respectiva carrera al finalizar el i-ésimo año.

Se consideró el avance hasta los 8 años, ya que es el máximo de años que puede permanecer un estudiante, según el reglamento de la institución. Para poder clasificar cada estudiante como alumno en riesgo o éxito académico, se procedió a considerar un criterio de avance curricular mínimo para no considerarlo en riesgo (véase la Tabla 5.3).

Año académico	Éxito = 1	Riesgo = 0
1	$\geq$ 12,5 %	<12,5 %
2	$\geq$ 25,0 %	<25,0 %
3	$\geq$ 37,5 %	<37,5 %
4	$\geq$ 50,0 %	<50,0 %
5	$\geq$ 62,5 %	<62,5 %
6	$\geq$75,0 %	<75,0 %
7	$\geq$ 87,5 %	<87,5 %
8	100,0 %	<100,0 %

Tabla 5.3: Criterio de clasificación para la variable respuesta.

Se consideró que al primer año el estudiante debe tener al menos un 12,5 % de asignaturas aprobadas del total de la malla para clasificarlo como éxito. En caso de ser menor a este porcentaje, este es identificado como posible alumno en riesgo de desertar la institución. De esta manera, se logra obtener a todos los alumnos clasificados como alumnos en riesgo (y=0) o éxito (y=1), para todos los periodos académicos en estudio. De esta manera, se soluciona el problema presentado por la variable estado académico.

Descripción de los datos

A continuación, se observa un análisis descriptivo sobre las tasas de deserción académica para el año 2010 de las diferentes facultades de la Universidad de Valparaíso. Considerando un tiempo suficientemente largo como para descartar la posibilidad que el estudiante se reincorpore, según lo expuesto en(Himmel, 2018).

Facultad	**n**	**Deserción(%)**	
Arquitectura	307	161	52,4 %
Derecho y Ciencias Sociales	200	83	41,5 %
Medicina	590	105	17,8 %
Odontología	82	20	24,4 %
Economía y Administración	731	280	38,3 %
Ciencias	**168**	**116**	**69,0 %**
Ciencias del Mar	33	14	42,4 %
Farmacia	97	23	23,7 %
Humanidades	145	61	42,1 %
Ingeniería	602	370	61,5 %
Total	2.955	1233	41,7 %

Tabla 5.4: Tasa de deserción para el año 2010 de la Universidad de Valparaíso, por facultad.

En la Tabla 5.4 se observa que, la mayor tasa de deserción en la institución, se presenta en la Facultad de Ciencias con un 69 % de los estudiantes que se matricularon para el año 2010; y respecto a la menor tasa de deserción, es la Facultad de Medicina con tan solo un 17,8 % de estudiantes que abandonó los estudios.

Se comparó el 41,7 % mostrado en la Tabla 5.4 de abandono total en la Universidad de Valparaíso para el año 2010, respecto a los países de la OCDE (2007), los cuales estiman que existe un promedio de deserción en las instituciones universitarias de un 31 %, en el cual la Universidad de Valparaíso está por encima de los países de la

OCDE, alcanzando tasas de deserción similares a las de países como Italia, Estados Unidos, Nueva Zelanda y Hungría, las cuales superan el 40 % de abandono en instituciones académicas.

En la siguiente tabla se detallan los porcentajes de deserción al primer, segundo y tercer año académico, para los alumnos que ingresaron en el año 2010 a la Facultad de Ciencias de la Universidad de Valparaíso.

Fac. de Ciencias	n	Primer año	Segundo año	Tercer año	Total
Ing. en Estadística	43	14 (32,6 %)	18 (41,9 %)	23 (53,5 %)	27 (62,8 %)
Lic. en Ciencias	25	7 (28,0 %)	9 (36,0 %)	13 (52,0 %)	14 (56 %)
Lic. en Física	33	12 (36,4 %)	16 (48,5 %)	21 (63,6 %)	29 (87,9 %)
Matemáticas	**54**	**23 (42,6 %)**	**32 (59,3 %)**	**33 (61,1 %)**	**36 (66,7 %)**
Meteorología	13	5 (38,5 %)	6 (46,2 %)	7 (53,8 %)	10 (76,9 %)
Total facultad	168	61 (36,3 %)	81 (48,2 %)	97 (57,7 %)	116 (69,0 %)

Tabla 5.5: Tasa de deserción de las carreras de la Facultad de Ciencias, para los alumnos que ingresaron en el año 2010.

Como se puede observar en la Tabla 5.5, la mayor tasa de deserción al primer año académico se produce en la carrera de Matemáticas (42,6 %), seguido de Meteorología (38,5 %); y en cuanto a la deserción total por carrera para el año 2010, Licenciatura en Física alcanza un abandono del 87,9 % de los alumnos ingresados en dicho año.

Se compararon los resultados de la Facultad de Ciencias de la Universidad de Valparaíso, con los mostrados a nivel nacional, en donde se logra evidenciar que para el año 2010, las tasas de abandono al primer año académico en la facultad son 36,3 %, superando el 22 % expuesto en (SIES, 2014), y también el 25 % reportado en Santelices et al. (2013).

Selección del mejor modelo de clasificación

Para la construcción de los modelos de clasificación, se utilizaron dos conjuntos de datos en distintos periodos de tiempo, conjunto 1 para la información de pre-ingreso de los estudiantes (modelo 1), y un conjunto 2 para la información del estudiante al cabo del primer año académico (modelo 2), en el que se consideraron las nuevas variables (véase Tabla 5.2). En cuanto a la selección del mejor modelo, se determinó por las medidas de eficiencia propuestas en el Sección 4.2, las cuales fueron calculadas mediante un proceso de validación cruzada, explicado en el Sección 4.3.

Clasificación logística

La construcción de este modelo se realizó en combinación con la selección de variables. Ésta se llevó a cabo mediante los métodos *Forward*, *Backward* y una combinación de estas dos *Stepwise*, explicado en la Sección 3.1, de esta forma se seleccionaron las que producen mayor poder predictivo reflejado en las medidas de eficiencia y, así poder obtener resultados óptimos y evitar posibles variables que produzcan ruido en el modelo. Este desarrollo se realizó a través de un proceso de validación cruzada.

Para el modelo de clasificación al momento del pre-ingreso de los estudiantes, la selección de atributos permitió identificar cuatro variables relevantes, siendo:

- Puntaje NEM.
- PSU de Matemáticas.
- Cobertura de salud.
- De proseguir estudios superiores, ¿con quien vivirá?

Por otra parte, para el modelo de clasificación al cabo del primer año, esta permitió obtener un total de cinco variables. Respecto al modelo anterior, se mantienen las variables de PSU de Matemáticas y la cobertura de salud:

- PSU de Matemáticas.
- PSU de Lenguaje.
- Cobertura de salud.
- Carrera de ingreso.
- Número de ramos aprobados en el primer año.

Considerando las variables seleccionadas anteriormente, a continuación se muestra en la Tabla 5.6 y 5.7 los coeficientes para la determinación de cada modelo, los cuales son los que lograron los mejores resultados de predicción para los distintos periodos de tiempo.

Variable	Coeficiente	Error estándar	Valor p
PSU de Matemáticas	0.0107	0.0019	**<0.01**
Puntaje NEM	0.0037	0.0011	**<0.01**
Cobertura de salud	-0.0739	0.0952	0.4374
Proseguir estudios	-0.0782	0.0518	0.1312
Constante	-8.5576	1.2074	<0.01

Tabla 5.6: Resultados de la clasificación logística al momento del pre-ingreso del estudiante.

Se puede observar en la Tabla 5.6, que las variables PSU Matemáticas y puntaje NEM, son estadísticamente significativas en el modelo y las variables cobertura de salud y proseguir estudio sin serlo, presentan igual de importancia al momento de una mejor clasificación.

Variable	Coeficiente	Error estándar	Valor p
PSU de Matemáticas	0.0158	0.0027	**0.0337**
PSU de Lenguaje	0.0029	0.0020	0.1477
Carrera de ingreso	0.2271	0.0728	**<0.01**
Cobertura de salud	-0.2560	0.2606	0.3259
N° de ramos aprobados	0.6067	0.0520	**<0.01**
Constante	-2.1109	1.5429	0.1713

Tabla 5.7: Resultados de la clasificación logística al cabo del primer año académico del estudiante.

En cuanto a la Tabla 5.7, se puede destacar las variables carrera y número de ramos aprobados como las estadísticamente significativas en el modelo, sin dejar de lado las demás variables, las cuales tienen gran importancia a la hora de una mejor predicción.

Razón de verosimilitud	**gl**	**Chi-cuadrado**	**valor-p**
Modelo de clasificación 1	5	54,54	<0.01
Modelo de clasificación 2	6	293,96	<0.01

Tabla 5.8: Test de razón de verosimilitud para significancia de los modelos.

En cuanto a los modelos mostrados anteriormente, se logra observar en la Tabla 5.8, que estos se ajustan de buena manera a los datos, es decir, que las variables regresoras tienen poder explicativo sobre el evento de estar de que los estudiantes estén en riesgo académico, siendo altamente significativos.

Medida de eficiencia	Modelo 1	Modelo 2
Exactitud	0,66	**0,81**
Precisión	0,67	**0,84**
Sensibilidad	**0,89**	0,86
Especificidad	0,30	**0,73**
F-1	0,76	**0,85**

Tabla 5.9: Promedios de las medidas de eficiencia, a partir de la validación cruzada utilizando clasificación logística.

En la Tabla 5.9, se puede observar los promedios de las medidas de eficiencia para las predicciones realizadas por los modelos de clasificación. Se logra evidenciar que ambos modelos la mayor eficiencia en la medida de sensibilidad 0,89 y 0,86; respectivamente; esto quiere decir que los modelos logran clasifican a un 89 % y 86 % de los alumnos que se encuentran verdaderamente en situación de riesgo académico. En cuanto a la precisión de los modelos, es más baja en el modelo que predice a un alumno al ingresar a la universidad. Esto nos dice que del total de clasificaciones de estudiantes en riesgo que clasifica el modelo, este acierta un 67 % y 84 %, obteniendo una mayor tasa de error en el modelo de pre-ingreso.

De manera gráfica se puede observar en la Figura 5.1, que en general el modelo al momento de tener información del estudiante al cabo del primer año, estos muestran mejores resultados de predicción, a excepción en la sensibilidad de este; ya que, es mejor en el modelo de pre-ingreso.

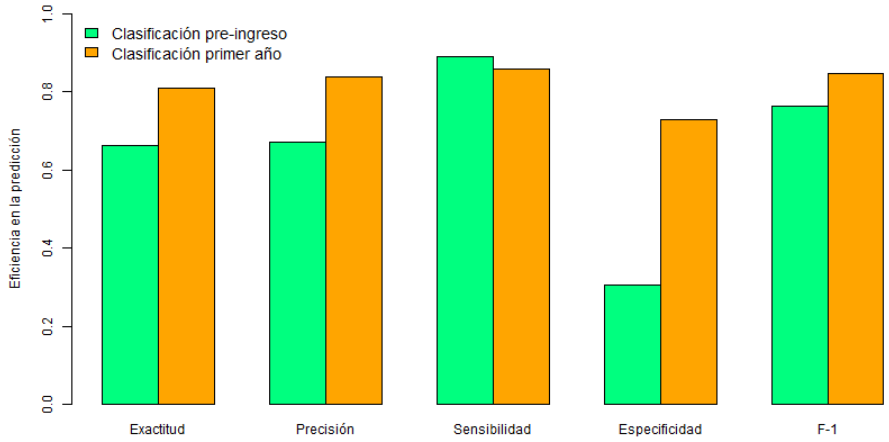

Figura 5.1: Comparación de las medidas de eficiencia utilizando clasificación logística.

Como se explicó anteriormente, los resultados de las medidas de eficiencia fueron calculados a partir de un proceso de validación cruzada ($k = 10$). A continuación, se pueden observar de manera gráfica los resultados de las medidas para cada iteración en el proceso de validación del modelo de clasificación logística.

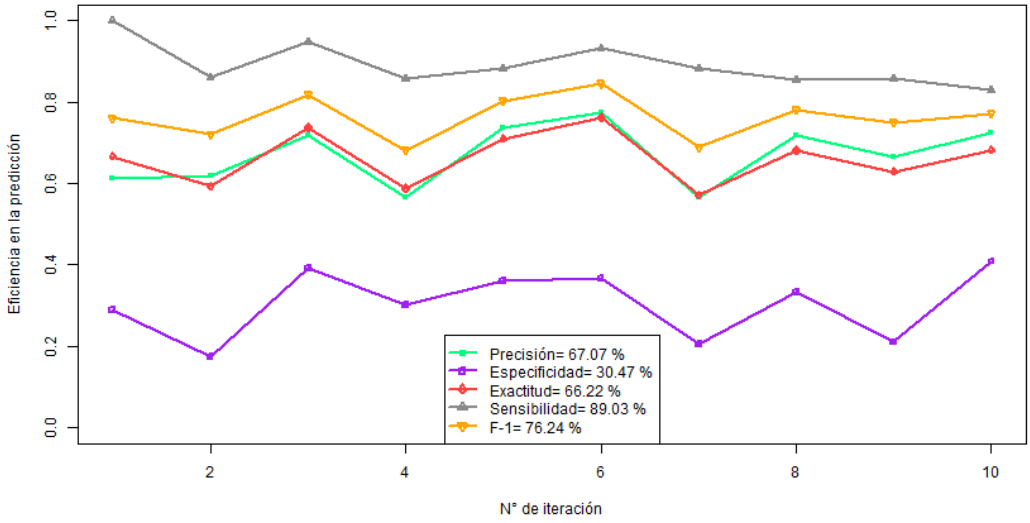

Figura 5.2: Medidas de eficiencia del modelo de clasificación logística, para los alumnos al momento del ingreso a la institución.

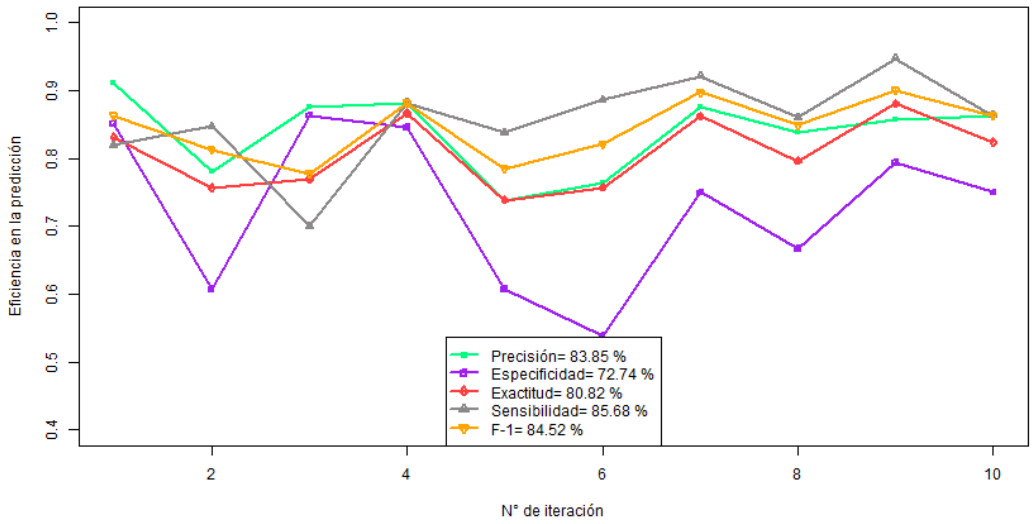

Figura 5.3: Medidas de eficiencia del modelo de clasificación logística, para los alumnos al cabo del primer año académico.

Se puede observar en la Figura 5.2 y 5.3, que para cada iteración en las medidas de eficiencia, estas se comportan de manera similar, ya que aumentan o disminuyen todas en conjunto en la mayoría de las iteraciones, para ver en detalle cada valor que toman estas medidas en cada etapa del proceso de validación, estas se pueden encontrar en el Anexo del presente trabajo, específicamente en la Tabla 7.2 y 7.3.

Se comparó los resultados obtenidos con un estudio reciente hecho en estudiantes de la Universidad de Chile (Celis et al., 2015), en el cual se obtuvo una sensibilidad de 0,86 y una precisión de 0,38. Obteniendo para el caso del modelo de clasificación al momento del pre-ingreso una sensibilidad de 0,89 y precisión de 0,66, siendo en ambos casos mejores. Ahora, para el modelo de clasificación al cabo del primer año académico, estos mejoran ya que, la sensibilidad es de un 0,86 y la precisión del modelo aumenta considerablemente a 0,84.

Máquinas de vectores de soporte

La construcción de este modelo se llevó a cabo, luego de la selección de atributos mediante la ganancia de información (véase Sección 3.2), ésta permitió identificar cuatro variables para la construcción del modelo de predicción al momento del pre-ingreso de los estudiantes:

- PSU de Matemáticas.
- PSU de Ciencias.
- PSU de Lenguaje.
- Puntaje NEM.

Para el caso de la selección de variables para el modelo 2, se encontraron seis atributos relevantes para la clasificación, manteniendo tres variables respecto al modelo anterior:

- Cantidad de ramos aprobados.
- Cantidad de ramos reprobados.
- Promedio de asignaturas inscritas.
- PSU de Matemáticas.
- PSU de Ciencias.
- Puntaje NEM.

Para la obtención del mejor modelo de *SVM*, se empleó el llamado truco del Kernel debido a que se tiene gran cantidad de variables en el conjunto de datos para ambos casos, por lo que su separación linealmente es imposible, explicado en la Sección 3.2. Para el caso del modelo 1, se identificó que el mejor Kernel de separación es con una función sigmoidal en un espacio de mayor dimensión; y para el conjunto 2, se identificó que la función que separa de mejor es una función radial. Para la selección entre un Kernel y otro, esto se llevó a cabo mediante una validación cruzada del conjunto de datos, en donde se escogió la que producía menor tasa de error en las predicciones, esto se logra evidenciar en la Tabla 5.10 que se muestra a continuación.

Kernel	**Lineal**	**Radial**	**Sigmoidal**	**Polynomial**
Modelo 1	0,3443	0,3409	**0,3378**	0,3473
Modelo 2	0,1957	**0,1878**	0,1920	0,2248

Tabla 5.10: Selección del mejor kernel por medio del error de clasificación.

A continuación, se muestran los resultados de las predicciones realizadas por estos dos modelos de clasificación para los dos periodos de tiempo.

En la Tabla 5.11 se observa que para ambos modelos la medida de sensibilidad es la más alta para los dos periodos de tiempo, alcanzando un 0,94 y 0,87, respectivamente. Esto quiere decir que, por ejemplo, de 100 alumnos que se encuentren realmente en situación de riesgo académico, los modelos detectan al 94 % y 87 % de estos alumnos. En cuanto a la precisión en la clasificación, se observa para modelo 1 una eficiencia del 0,65 en la clasificación; en cuanto al modelo 2, éste alcanza un 0,84. En otras palabras, de las clasificaciones que el modelo detecta como riesgo, estos tienen un porcentaje de acierto de 65 % y 84 % para cada modelo. Cuando se necesita clasificar un alumno como éxito académico; el modelo 1 obtiene un porcentaje de acierto del 19 % observado en la medida de especificidad; en cuanto al modelo al modelo 2, esta aumenta considerablemente, logrando clasificar a un 75 % de los alumnos como éxito, cuando realmente el estudiante es un éxito.

Medida de eficiencia	**Modelo 1**	**Modelo 2**
Exactitud	0,65	0,82
Precisión	0,65	0,84
Sensibilidad	**0,94**	**0,87**
Especificidad	0,19	0,75
F-1	0,77	0,86

Tabla 5.11: Promedios de las medidas de eficiencia, a partir de la validación cruzada utilizando las máquinas de vectores de soporte.

De manera gráfica se puede observar en la Figura 5.4, que en general el modelo al momento de tener información del estudiante al cabo del primer año, estos muestran mejores resultados de predicción, a excepción en la sensibilidad de este, ya que es mejor en el modelo de pre-ingreso. La comparación de estas medidas mostradas son calculadas a partir del promedio de la validación cruzada.

La Figura 5.5 y 5.6 muestra los resultados de la validación cruzada, realizada por las *SVM* en cada iteración dentro del proceso. Se logra visualizar que las líneas en cada gráfico tienen similares comportamientos de tendencia y por otro lado, se logra ver en la Figura 5.5 que existe una gran diferencia entre la sensibilidad (color gris) y la especificidad (color rojo). En cambio, en la Figura 5.6 se logra ver comportamientos similares en los resultados de cada iteración, esto debido a que las medidas de eficiencia son mejores en el modelo 2. Los valores de las medidas de eficiencia para cada iteración, se pueden observar en la Tabla 7.4 y 7.5 del Anexo.

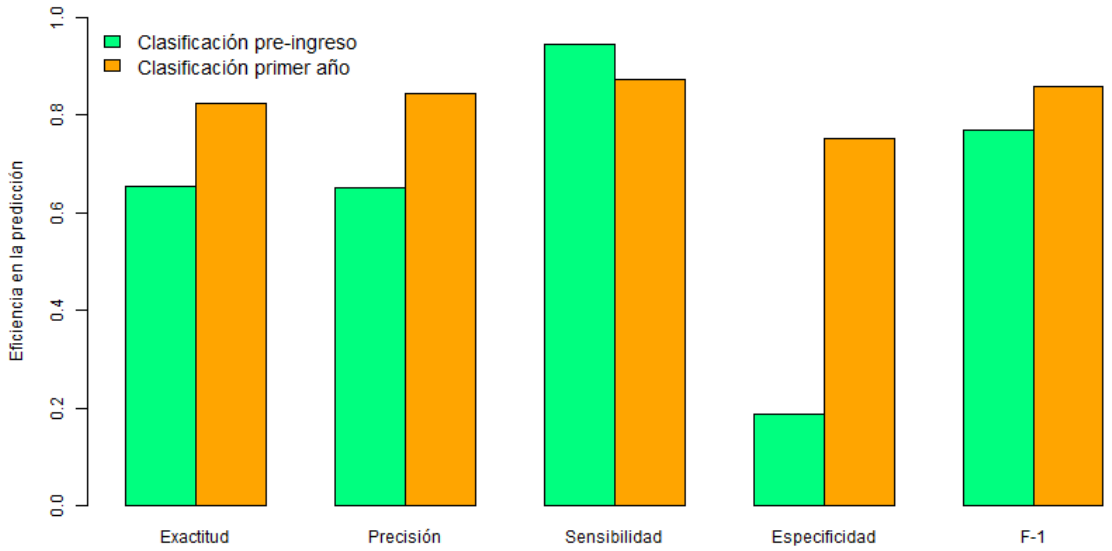

Figura 5.4: Comparación de las medidas de eficiencia utilizando máquinas de soporte vectorial.

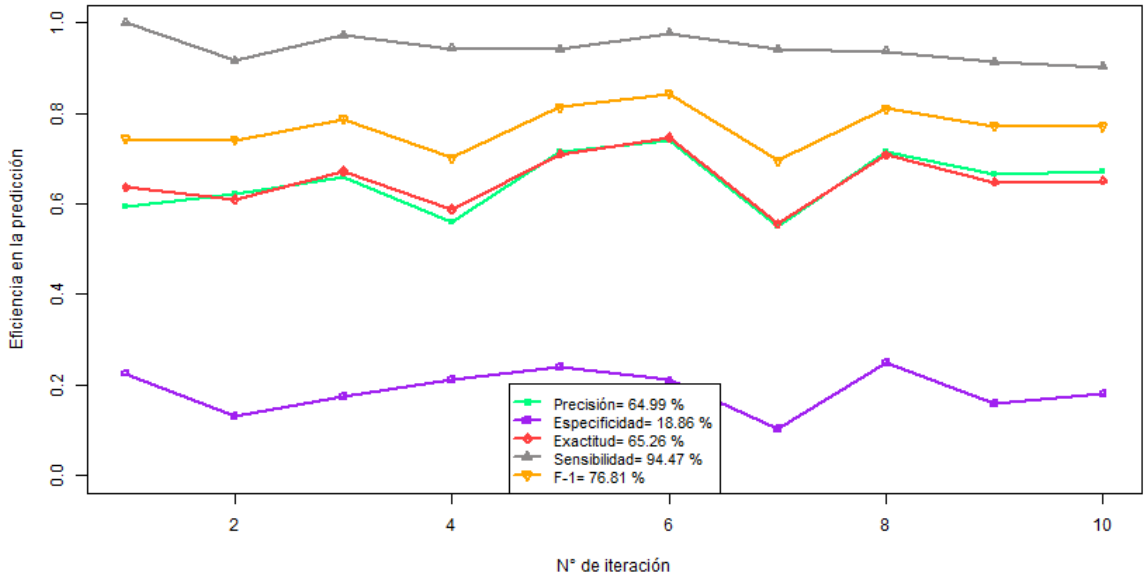

Figura 5.5: Medidas de eficiencia del modelo *SVM*, para los alumnos al momento del ingreso a la institución.

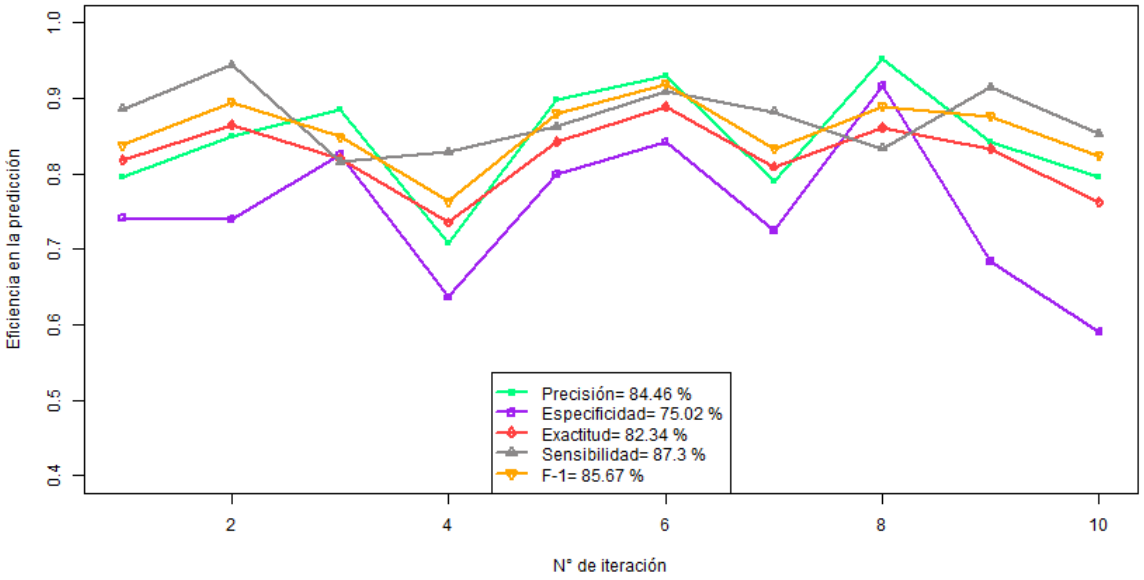

Figura 5.6: Medidas de eficiencia del modelo de *SVM*, para los alumnos al cabo del primer año académico.

Redes Neuronales

Para la obtención de este modelo de clasificación por medio de las redes neuronales, se utilizó un perceptrón multicapa (*MultiLayer perceptron*, en Inglés) con conexión hacia adelante *feedforward* (véase Figura 3.8), la cual tiene la particularidad de poder resolver problemas que no son linealmente separables. Este modelo se compone de tres capas: entrada, oculta y salida; la primera se encarga de introducir los patrones a la red; la capa oculta es la encargada de hacer el procesamiento no lineal de los datos de entrada y entregar la información a la siguiente capa por medio de una función de activación; y la última capa de salida, proporciona al exterior la respuesta de la red, que para el caso particular de este trabajo es binaria y su función de activación es sigmoide.

Para la determinación del número de neuronas en la capa de entrada, ésta se llevó a cabo mediante los métodos de selección de variables: *Forward*, *Backward* y *Stepwise*; obteniendo igual número de neuronas como variables seleccionadas. Permitiendo identificar para el caso del modelo 1:

- PSU de matemáticas.
- Puntaje NEM.
- Cobertura de salud.
- De proseguir estudios superiores, ¿con quien vivirá?.

Para la selección de variables en el modelo 2, ésta permitió obtener 5 atributos relevantes para la clasificación:

- PSU de Matemáticas.
- PSU de Lenguaje.
- Cobertura de salud.
- Carrera de ingreso.
- Número de ramos aprobados.

De acuerdo a lo anterior, se visualiza a continuación la estructura de las redes, tanto para el modelo 1 (Figura 5.7) y el modelo 2 (Figura 5.8).

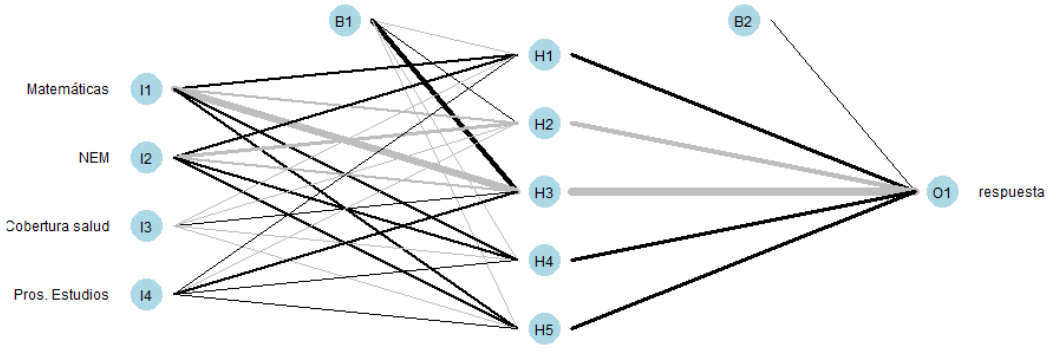

Figura 5.7: Red neuronal perceptrón multicapa para el modelo 1.

Como se puede observar en la Figura 5.7, existen distintos colores y grosores en las líneas que unen las capas dentro de la red. Las líneas negras significan pesos positivos y grises negativos, cuanto más ancho es la línea más peso tiene la neurona. Por lo tanto, las variables PSU de Matemáticas y puntaje NEM son las que aportan mayor información en el modelo de clasificación.

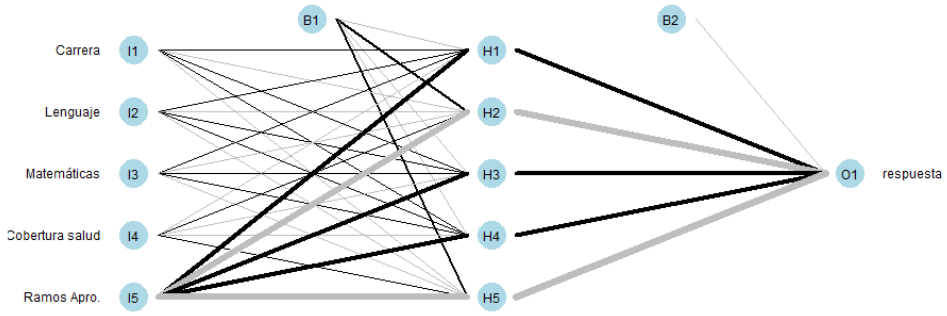

Figura 5.8: Red neuronal perceptrón multicapa para el modelo 2.

Se puede observar de manera clara en la Figura 5.8 que, la variable número de ramos aprobados, es la que aporta mayor información, tanto para las neuronas en la capa oculta, como también en la de salida de la red, esto porque la variable es la que entrega mayor información al momento de hacer las predicciones.

Los resultados de las medidas de eficiencia de los clasificadores, se muestran en la Tabla 5.12 a continuación:

Medida de eficiencia	**Modelo 1**	**Modelo 2**
Exactitud	0,67	0,80
Precisión	0,67	0,81
Sensibilidad	**0,91**	**0,88**
Especificidad	0,29	0,68
F-1	0,77	0,84

Tabla 5.12: Promedios de las medidas de eficiencia, a partir de la validación cruzada utilizando redes neuronales.

Como se observa en la Tabla 5.12, la medida de sensibilidad es la más alta con valores de 0,91 y 0,88 respectivamente para cada modelo, ésto explica que del total de alumnos que se encuentran en verdadera situación de riesgo, los modelos detectan aproximadamente un 90 % de estos estudiantes. Por otro lado, analizando la medida de precisión, la cual indica el total de clasificaciones en riesgo, que realmente son riesgo académico, estas son 0,67 para el modelo 1 y 0,80 en modelo 2, siendo superior en este último. No obstante, cuando se desea clasificar a un estudiante como éxito académico el error en el modelo 1 aumenta considerablemente, ya que solo tiene una especificidad de 0,29.

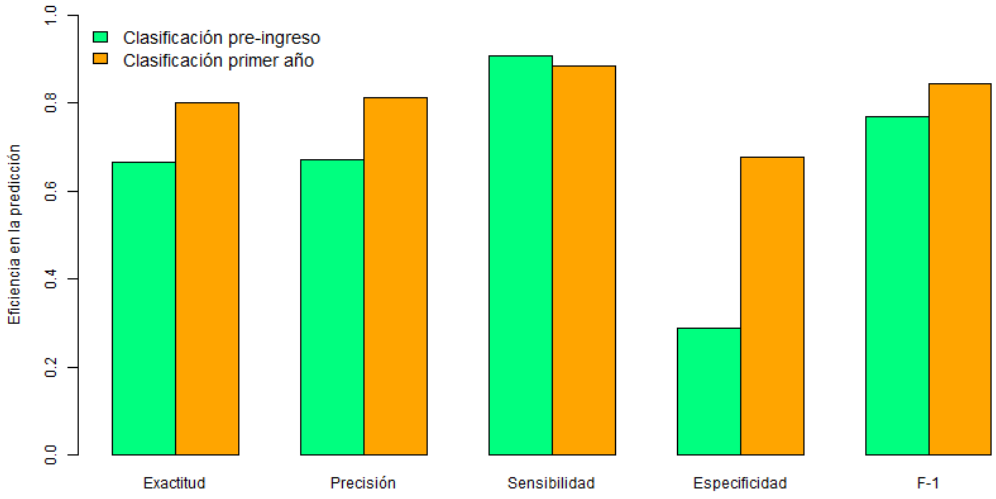

Figura 5.9: Comparación de las medidas de eficiencia utilizando máquinas de soporte vectorial.

La Figura 5.9 muestra la comparación de los promedios de estas medidas de eficiencia de manera gráfica, donde se puede observar que, para cada una de las medidas de eficiencia los resultados mejoran en el modelo 2, como se explicó anteriormente, a excepción de la sensibilidad que tienen resultados muy similares.

En la Figura 5.10 y 5.11, se muestran los resultados en cada iteración del proceso de validación cruzada para cada modelo, tanto para el modelo de clasificación al momento del pre-ingreso del estudiante, como para el término del primer año académico. Para ver los valores obtenidos por cada medida de rendimiento en el proceso de validación, este puede ser obervado en el Anexo del trabajo, en la Tabla 7.6 y Tabla 7.7, respectivamente.

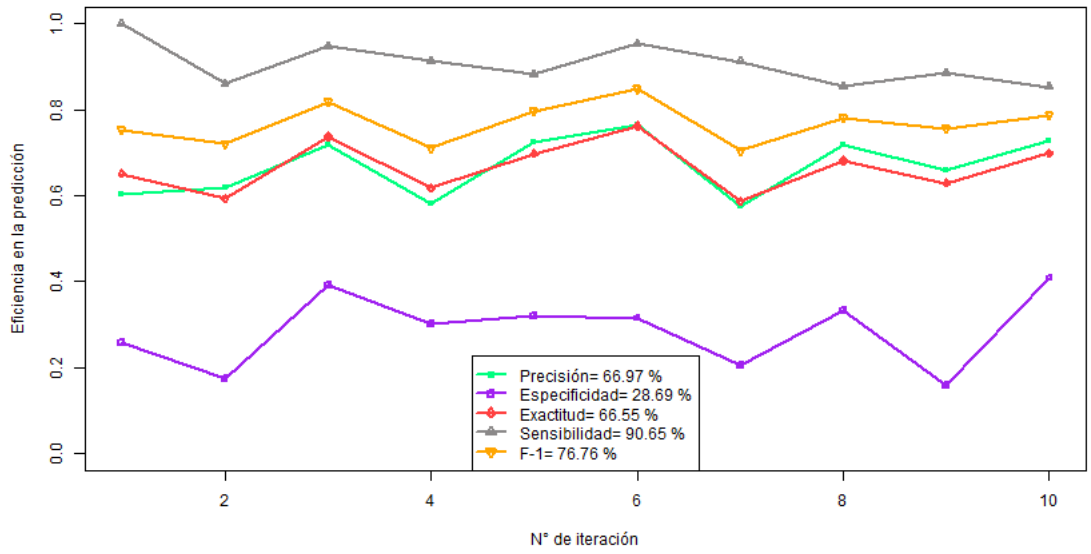

Figura 5.10: Medidas de eficiencia del modelo *ANN*, para los alumnos al momento del ingreso a la institución.

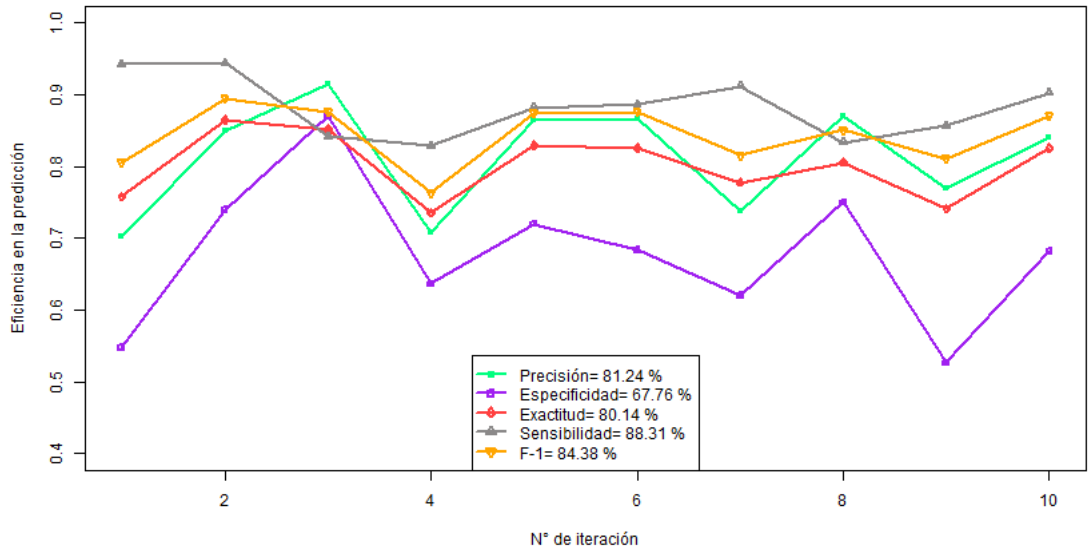

Figura 5.11: Medidas de eficiencia del modelo de *ANN*, para los alumnos al cabo del primer año académico.

Comparación de las medidas de eficiencia de las técnicas de clasificación

Por último, el objetivo es comparar estos distintos modelos de clasificación, logrando identificar el que tenga los mejores resultados de predicción, y así poder encontrar el que mejor se adapte a los objetivos. Para esto, se comparó las medidas de sensibilidad, precisión y F-1, las cuales son las que satisfacen nuestros objetivo de poder detectar tempranamente alumnos en riesgo académico.

Medidas de eficiencia	**Clasificación logística**	***SVM***	**Redes neuronales**
Precisión	**0,67**	0,65	**0,67**
Sensibilidad	0,89	**0,94**	0,91
F-1	0,76	**0,77**	**0,77**

Tabla 5.13: Resultados de las técnicas de clasificación de minería de datos, para la clasificación de los alumnos al pre-ingreso.

Considerando los resultados de las distintas técnicas de clasificación empleadas, para poder identificar posibles alumnos en riesgo al momento del pre-ingreso a la institución, se logra visualizar en la tabla 5.13 que, si bien los resultados fueron similares en ambas técnicas, las máquinas de soporte vectorial y redes neuronales fueron las que obtuvieron los valores más altos en la medida F-1, la cual indica un promedio entre la presición y sensibilidad de las clasificaciones. Sin embargo, analizando el número y tipo de variables que se utilizaron en cada técnica; *SVM* considera tan solo 4 variables, entre las cuales coincide con las ya estudiadas en la literatura, como determinantes a la hora de la deserción académica. Estas variables son: PSU de Matemáticas, PSU de Ciencias, PSU de Lenguaje y por último el puntaje NEM.

Medidas de eficiencia	**Clasificación logística**	***SVM***	**Redes neuronales**
Precision	**0,84**	**0,84**	0,81
Sensibilidad	0,86	0,87	**0,88**
F-1	0,85	**0,86**	0,84

Tabla 5.14: Resultados de las técnicas de clasificación de minería de datos, para la clasificación de los alumnos al término del primer año académico.

Comparando los resultados de la Tabla 5.13, se puede observar que los resultados tienen valores similares, pero considerando la medida F-1, *SVM* tiene una eficiencia mayor sobre las otras técnicas. Además, al evaluar las variables que utiliza esta técnica, tan solo considera tres variables para lograr predecir y obtener los mejores resultados; siendo las variables número de ramos aprobados, reprobados y el promedio de notas obtenido en el primer año de universidad.

Entre los resultados obtenidos por *SVM*, se destaca la sensibilidad del modelo al momento de realizar las predicciones de estudiantes en riesgo, logrando identificar al 87 % del total de alumnos en riesgo académico, como posibles estudiantes desertores de la institución. En cuanto a la precisión del modelo, este logra un 84 % de acierto en las predicciones que el modelo clasifica como en riesgo.

En cuanto a los resultados obtenidos en (Delen, 2010) y (Jia and Mareboyana, 2013), utilizando las mismas técnicas de clasificación. Estos coinciden con los presentados en el presente trabajo, ya que ambas técnicas arrojaban resultados similares a la hora de clasificar, obteniendo mejores resultados con las máquinas de vectores de soporte por una mínima diferencia. Además, en ambos trabajos solo se dispone de la información de los estudiantes al término del año académico.

Capítulo 6

Conclusiones

En la literatura se pudo observar que existe un notorio aumento de investigaciones que buscan detectar a estudiantes en situación de riesgo académico, esto se debe al gran incremento de las tasas de deserción en las distintas instituciones a nivel mundial. Por lo que, diferentes sistemas de alerta temprana ayudan a mitigar esta problemática, en donde algunos países lo han considerado como un desafío de políticas públicas. Chile alcanza una tasa de deserción del 40 % según estudios al año 2016, estando en el promedio a nivel mundial.

Se logró identificar que para el caso particular de la Universidad de Valparaíso, la tasa de deserción alcanzó un 41,7 % en aquellos alumnos que ingresaron en el año 2010; sin embargo, se destaca particularmente la Facultad de Ciencias que obtiene cifras cercanas al 69 %.

El objetivo principal del presente trabajo de titulación, fue proponer un modelo eficiente de clasificación para la detección temprana de alumnos en riesgo académico en la Universidad de Valparaíso.

Para tales efectos se utilizaron distintos datos personales y académicos históricos de siete cohortes de ingreso de estudiantes a la Universidad de Valparaíso. Centrando los objetivos en la Facultad de Ciencias se logró obtener un alto poder predictivo utilizando distintas técnicas de minería de datos, identificando *SVM* como la mejor técnica a utilizar para futuras predicciones. El modelo logra identificar a un 87 % de los alumnos que se encuentran en riesgo académico, los cuales son posibles desertores de la institución.

En cuanto a los otros métodos de clasificación que se utilizaron en este trabajo, se logra concluir que las técnicas de minería de datos, son una herramienta válida para poder generar políticas internas en la institución, y de tal forma de poder disminuir las tasas de deserción en la Universidad de Valparaíso.

Una de las limitaciones identificadas en este trabajo de título, se debe a la calidad de los datos obtenidos; ya que, provocó gran pérdida de información relevante sobre los estudiantes, debido a que existe gran cantidad de información que hoy en día se recopila, pero esta no se obtiene de manera adecuada, puesto que, existen variables desactualizadas, como también en algunas se desconoce la interpretación de los valores obtenidos, también existen errores a lo hora de ingresar los datos, entre otros.

Por lo tanto, como trabajo a futuro se propone implementar este modelo en las futuras generaciones de estudiantes, así de esta manera la institución logrará tener un juicio no tan solo subjetivo de un estudiante a la hora de eliminarlo o considerar su permanencia en la institución. Además, se propone mejorar los sistemas de recopilación de la información, como son la asistencia del estudiante, notas parciales actualizadas y lo más importante, el estado académico al cierre de cada semestre. De esta manera se podrá obtener datos consistentes y con mayor cantidad de variables explicativas que les permita identificar aquellos patrones de comportamiento, para así lograr realizar un modelo de detección temprana en estudiantes con el fin de poder disminuir la tasa de deserción estudiantil de la Universidad de Valparaíso.

Anexos

Año de ingreso	n	Primer año	Segundo año	Tercer año
2010	2.925	299 (10,1 %)	704 (23,8 %)	920 (31,1 %)
2011	2.752	364 (13,2 %)	653 (23,7 %)	914 (33,2 %)
2012	2.878	336 (11,7 %)	788 (27,4 %)	1.012 (35,2 %)
2013	2.891	449 (15,5 %)	925 (32,0 %)	1.094 (37,8 %)
2014	2.713	364 (13,4 %)	715 (26,4 %)	-
2015	2.999	411 (13,7 %)	-	-
Promedio	**2.860**	**370 (12,9 %)**	**757 (26,7 %)**	**985 (34,3 %)**

Tabla 6.1: Tasas de abandono al primer, segundo y tercer año académico de los estudiantes de la Universidad de Valparaíso.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,667	0,614	1,000	0,290	0,760
2	0,593	0,620	0,861	0,174	0,721
3	0,738	0,720	0,947	0,391	0,818
4	0,588	0,566	0,857	0,303	0,682
5	0,711	0,738	0,882	0,360	0,804
6	0,762	0,774	0,932	0,368	0,845
7	0,571	0,566	0,882	0,207	0,690
8	0,681	0,719	0,854	0,333	0,781
9	0,630	0,667	0,857	0,211	0,750
10	0,683	0,723	0,829	0,409	0,773
Promedio	**0,662**	**0,671**	**0,890**	**0,305**	**0,762**

Tabla 6.2: Validación cruzada para clasificación logística del modelo 1.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,831	0,911	0,820	0,852	0,863
2	0,757	0,780	0,848	0,607	0,813
3	0,769	0,875	0,700	0,864	0,778
4	0,867	0,882	0,882	0,846	0,882
5	0,738	0,738	0,838	0,607	0,785
6	0,757	0,765	0,886	0,538	0,821
7	0,862	0,875	0,921	0,750	0,897
8	0,796	0,838	0,861	0,667	0,849
9	0,881	0,857	0,947	0,793	0,900
10	0,824	0,864	0,864	0,750	0,864
Promedio	**0,808**	**0,838**	**0,857**	**0,727**	**0,845**

Tabla 6.3: Validación cruzada para clasificación logística del modelo 2.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,636	0,593	1,000	0,226	0,745
2	0,610	0,623	0,917	0,130	0,742
3	0,672	0,661	0,974	0,174	0,787
4	0,588	0,559	0,943	0,212	0,702
5	0,711	0,716	0,941	0,240	0,814
6	0,746	0,741	0,977	0,211	0,843
7	0,556	0,552	0,941	0,103	0,696
8	0,708	0,714	0,938	0,250	0,811
9	0,648	0,667	0,914	0,158	0,771
10	0,651	0,673	0,902	0,182	0,771
Promedio	**0,653**	**0,650**	**0,945**	**0,189**	**0,768**

Tabla 6.4: Validación cruzada para *SVM* del modelo 1.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,818	0,795	0,886	0,742	0,838
2	0,864	0,850	0,944	0,739	0,895
3	0,820	0,886	0,816	0,826	0,849
4	0,735	0,707	0,829	0,636	0,763
5	0,842	0,898	0,863	0,800	0,880
6	0,889	0,930	0,909	0,842	0,920
7	0,810	0,790	0,882	0,724	0,833
8	0,861	0,952	0,833	0,917	0,889
9	0,833	0,842	0,914	0,684	0,877
10	0,762	0,796	0,854	0,591	0,824
Promedio	**0,823**	**0,845**	**0,873**	**0,750**	**0,857**

Tabla 6.5: Validación cruzada para *SVM* del modelo 2.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,652	0,603	1,000	0,258	0,753
2	0,593	0,620	0,861	0,174	0,721
3	0,738	0,720	0,947	0,391	0,818
4	0,618	0,582	0,914	0,303	0,711
5	0,697	0,726	0,882	0,320	0,796
6	0,762	0,764	0,955	0,316	0,848
7	0,587	0,574	0,912	0,207	0,705
8	0,681	0,719	0,854	0,333	0,781
9	0,630	0,660	0,886	0,158	0,756
10	0,698	0,729	0,854	0,409	0,787
Promedio	**0,666**	**0,670**	**0,906**	**0,287**	**0,768**

Tabla 6.6: Validación cruzada para RN del modelo 1.

Iteración	Exactitud	Precisión	Sensibilidad	Especificidad	F-1
1	0,758	0,702	0,943	0,548	0,805
2	0,864	0,850	0,944	0,739	0,895
3	0,852	0,914	0,842	0,870	0,877
4	0,735	0,707	0,829	0,636	0,763
5	0,829	0,865	0,882	0,720	0,874
6	0,825	0,867	0,886	0,684	0,876
7	0,778	0,738	0,912	0,621	0,816
8	0,806	0,870	0,833	0,750	0,851
9	0,741	0,769	0,857	0,526	0,811
10	0,825	0,841	0,902	0,682	0,871
Promedio	**0,801**	**0,812**	**0,883**	**0,678**	**0,844**

Tabla 6.7: Validación cruzada para RN del modelo 2.

Código de programación en R-project

Para encontrar el código completo de programación que se utilizó en el presente trabajo de titulación, se puede visitar el siguiente sitio web, en el cuál está disponible gratuitamente en: `https://github.com/felipequezada1994/EDM`.

A continuación, se muestra de manera generaliza el código utilizado para los resultados:

Validación cruzada para clasificación logítica

```
set.seed(1994)
Folds     <- 10
datos$kfold    <- sample(1:Folds, nrow(datos), replace = T)
Iter  <- data.frame(iteracion = NULL, Accurracy = NULL, Precision        =
    NULL, Recall = NULL, specificidad = NULL, Score = NULL)
for (i in 1:Folds)
{
Test          <- subset(datos, kfold  == i)
Entrenamiento <- subset(datos, !kfold == i)
modelo = glm(respuesta~psu_matematica+psu_nem+familia_salud+
             familia_proseguir_estudios,data =Entrenamiento,
                                    family = "binomial")
  glm.prob <- predict (modelo , Test, type ="response")
  glm.pred <- rep (0 ,nrow(Test))
  glm.pred [ glm.prob >.5] <- 1
  MC            <- table(glm.pred,Test$respuesta)
  accurracy     <- sum(diag(MC)) / sum(MC)
  precision     <- MC[1,1] / (MC[1,1] + MC[1,2])
  recall        <- MC[1,1] / (MC[1,1] + MC[2,1])
  specificidad  <- MC[2,2] / (MC[2,2] + MC[1,2])
  score         <- (2*precision*recall)/(recall+precision)
  Iter          <- rbind(Iter, data.frame(Iter =i,
                                    Accurracy = accurracy, Precision =
     precision, Recall =                 recall, Specificidad =
     specificidad,Score = score))
}

mean(Iter$Accurracy)
mean(Iter$Precision)
mean(Iter$Recall)
mean(Iter$Specificidad)
mean(Iter$Score)
```

Validación cruzada para *SVM*

```
library(e1071)
set.seed(1994)
Folds         <- 10
datos$kfold    <- sample(1:Folds, nrow(datos), replace = T)
Iter  <- data.frame(iteracion = NULL, Accurracy = NULL, Precision        =
    NULL, Recall = NULL, specificidad = NULL, Score = NULL)
for (i in 1:Folds)
{
```

```
  Test          <- subset(datos, kfold  == i)
  Entrenamiento <- subset(datos, !kfold == i)
  Modelo <- svm(respuesta ~ psu_matematica + psu_ciencia+
                              psu_lenguaje+psu_nem, data = Entrenamiento,
                                        kernel="sigmoid", gamma=0.01,
      cost=10)
  Prediccion <- predict(Modelo, new= Test, type="class")
  MC         <- table(Prediccion, Test$respuesta)
  accurracy  <- sum(diag(MC)) / sum(MC)
  precision  <- MC[1,1] / (MC[1,1] + MC[1,2])
  recall     <- MC[1,1] / (MC[1,1] + MC[2,1])
  specificidad <- MC[2,2] / (MC[2,2] + MC[1,2])
  score      <- (2*precision*recall)/(recall+precision)
  Iter                <- rbind(Iter, data.frame(Iter =i,
                                          Accurracy = accurracy,
      Precision = precision, Recall =                recall,
      Specificidad = specificidad,Score = score))
}

mean(Iter$Accurracy)
mean(Iter$Precision)
mean(Iter$Recall)
mean(Iter$Specificidad)
mean(Iter$Score)
```

Validación cruzada para redes neuronales

```
library(nnet)
library(caret)
library(NeuralNetTools)
set.seed(1994)
Folds      <- 10
datos$kfold  <- sample(1:Folds, nrow(datos), replace = T)
Iter  <- data.frame(iteracion = NULL, Accurracy = NULL, Precision      =
    NULL, Recall = NULL, specificidad = NULL, Score = NULL)
for (i in 1:Folds)
{
  Test          <- subset(datos_norm, kfold  == i)
  Entrenamiento <- subset(datos_norm, !kfold == i)
  Modelo <- nnet(respuesta ~ psu_matematica+psu_nem
                            +familia_salud + familia_proseguir_estudios,
      data=                              Entrenamiento, size=size, decay=
     decay, trace=F)
  Prediccion <- predict(Modelo, new= Test, type="class")
  MC         <- table(Prediccion, Test$respuesta)
  accurracy  <- sum(diag(MC)) / sum(MC)
```

```
  precision     <- MC[1,1] / (MC[1,1] + MC[1,2])
  recall        <- MC[1,1] / (MC[1,1] + MC[2,1])
  specificidad  <- MC[2,2] / (MC[2,2] + MC[1,2])
  score         <- (2*precision*recall)/(recall+precision)
  Iter          <- rbind(Iter, data.frame(Iter =i,
                            Accurracy = accurracy, Precision = precision
    , Recall =                  recall, Specificidad = specificidad,
    Score = score))
}
mean(Iter$Accurracy)
mean(Iter$Precision)
mean(Iter$Recall)
mean(Iter$Specificidad)
mean(Iter$Score)
```

Gráfico de líneas

```
library(png)
metodo <- #matriz de resultados validación cruzada por método
png("SVM_0.jpg", width=840)
plot_color<-c("springgreen","purple","brown1","snow4","orange")
plot(metodo$Precision, axes=T , type = "o", lwd= 2,lty=1,
          pch=21, cex=.8, xlab ="N° de iteración",
                                ylab="Eficiencia en la predicción",
    xlim = c(1,10), ylim= c(0,1), col = plot_colores[1])
lines(metodo$Specificidad, lty=1, cex=.8, pch=22, type = "o",
   lwd= 2, col = plot_color[2])
lines(metodo$Accurracy, lty=1, cex=.8, pch=23, type = "o", lwd=
   2, col = plot_color[3])
lines(metodo$Recall, lty=1, cex=.8, pch=24, type = "o", lwd= 2,
   col = plot_colores[4])
lines(metodo$Score, lty=1, cex=.8, pch=25, type = "o", lwd= 2,
   col = plot_color[5])
legend("bottom", horiz = F, lwd=2, pch= 21:25, legend = c(
        paste("Precisión=", format(mean(metodo
                                $Precision)*100,digits=4), "%"),
        paste("Especificidad=", format(mean(metodo
                                $Specificidad)*100,digits=4), "%"),
        paste("Exactitud=", format(mean(metodo
                                $Accurracy)*100,digits=4), "%"),
        paste("Sensibilidad=", format(mean(metodo
                                $Recall)*100,digits=4), "%"),
        paste("F-1=", format(mean(metodo$Score)*100,digits=4), "%")
        ),cex = 1, col = plot_color, lty = c(1,1,1,1,1))
box()
dev.off()
```

Referencias

Ali, S. and Smith-Miles, K. A. (2006). A meta-learning approach to automatic kernel selection for support vector machines. *Neurocomputing*, 70(1-3):173–186.

Anónimo (2015). Multilayer perceptron. [online] disponible en `http://informaticaaplicadaprimerobcsj.blogspot.com/2015/04/multilayer-perceptron.html` [revisado 20 de octubre de 2018].

Antunes, C. (2010). *Anticipating students failure as soon as possible*. Chapman & Hall/CRC Press, New York, EEUU.

Cabero Almenara, J. (2015). Reflexiones educativas sobre las tecnologías de la información y la comunicación (TIC). *Tecnología, Ciencia y Educación, 1, 19-27.*

Celis, S., Moreno, L., Poblete, P., Villanueva, J., and Weber, R. (2015). Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería. *Revista Ingenieía de Sistemas Volumen XXIX.*

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506.

Devijver, P. A. and Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice hall.

Flores, A., Maldonado, S., and Weber, R. (2015). Selección de atributos y support vector machines adaptado al problema de fuga de clientes. *Revista Ingeniería de Sistemas Volumen XXIX.*

Gutiérrez, J. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. *Pdfs. Semanticscholar. Org*, 1:1–17.

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

Himmel, E. (2018). Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación*, (17):91–108.

Jia, J.-W. and Mareboyana, M. (2013). Machine learning algorithms and predictive models for undergraduate student retention. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.

Lama, M. D. V., Sánchez, H. M., Sobrino, J. N. R., and Jiménez, M. T. (2017). Elección contable para la valoración de las inversiones inmobiliarias. contribución de las técnicas de minería de datos para determinar patrones de decisión. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 23:234–256.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124.

Martelo, R. J., Ponce, A. L., and Acuña, F. (2016). Guía metodológica para el diseño de un plan estratégico informático en instituciones de educación superior. *Formación universitaria*, 9(1):91–98.

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

Michavila, F. and Martínez, J. (2016). Educacion superior en iberoamerica informe 2016, informe nacional: España.

Miranda, M. A. and Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formación universitaria*, 10(3):61–68.

OCDE, I. (2007). Education at a glance 2007.

Orozco, L. (2016). Educacion superior en iberoamerica informe 2016, informe nacional: Colombia.

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.

Phung, S. L., Bouzerdoum, A., and Nguyen, G. H. (2009). Learning pattern classification tasks with imbalanced data sets.

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

Riobóo, L. M. D. and Pedroza, M. E. (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico universitario. *Revista Científica de FAREM-Estelí*, (24):143–152.

Romero, C. and Ventura, S. (2013). Data mining in education. wires data mining and knowledge discover.

Ruiz, S. (2015). La predicción del dato: Redes neuronales artificiales. [online] disponible en `https://www.analiticaweb.es/la-prediccion-del-dato-redes-neuronales-artificiales` [revisado 20 de octubre de 2018].

Santelices, V., Catalán, X., Horn, C., and Kruger, D. (2013). Determinantes de deserción en la educación superior chilena, con énfasis en efecto de becas y créditos. *Santiago de Chile: Fondo de Investigación y Desarrollo en Educación: Mineduc*.

Seidman, A. (1996). Retention revisited: R= e, id+ e & in, iv. *College and University*, 71(4):18–20.

Siemens, G. and Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 252–254.

SIES (2014). Panorama de la educación superior en chile.

Soberanis, M. (2013). Aprendizaje automatico. [online] disponible en `https://medium.com/soldai/tagged/aprendizaje-automatico` [revisado 20 de octubre de 2018].

Suárez, E. J. C. (2014). Tutorial sobre máquinas de vectores soporte (svm). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). Classification: basic concepts, decision trees, and model evaluation. *Introduction to data mining*, 1:145–205.

Tinto, V. (1989). Definir la deserción: una cuestión de perspectiva. *Revista de educación superior*, 71(18):1–9.

Vapnik, V. (1998). *Statistical learning theory. 1998*, volume 3. Wiley, New York.

Yu, H. and Kim, S. (2012). Svm tutorial-classification, regression and ranking. In *Handbook of Natural computing*, pages 479–506. Springer.