UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DESARROLLO DE UN MODELO PREDICTIVO DE DESERCIÓN DE ESTUDIANTES DE PRIMER AÑO EN INSTITUCIÓN DE EDUCACIÓN SUPERIOR

MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL ELÉCTRICO

MATÍAS GALLEGUILLOS AGUILAR

PROFESOR GUÍA:
CLAUDIO MARCELO HELD BARRANDEGUY

MIEMBROS COMISIÓN:
MARCOS EDUARDO ORCHARD CONCHA
ANDRÉS CABA RUTTE

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR POR
EL TÍTULO DE INGENIERO CIVIL ELÉCTRICO.
POR: MATÍAS GALLEGUILLOS A.
FECHA 10-09-2018
PROF. GUÍA: Dr CLAUDIO HELD B.

DESARROLLO DE UN MODELO PREDICTIVO DE DESERCIÓN DE ESTUDIANTES DE PRIMER AÑO EN INSTITUCIÓN DE EDUCACIÓN SUPERIOR

En Chile, durante los últimos 30 años ha habido un crecimiento significativo en el acceso de las personas a la educación superior. Acompañado de este crecimiento se ha visto un aumento en la deserción universitaria, siendo particularmente elevada la de alumnos de primer año. Este problema tiene grandes costos de distinta índole tanto para los alumnos como para las universidades, haciendo que se haya posicionado como una de las métricas más importantes que se utiliza para acreditar a las instituciones.

La Universidad de las Américas se ha visto enfrentada a una alta tasa de deserción, traduciéndose en que en el año 2013 haya contribuido de manera importante a la pérdida de su acreditación, por lo que se transformó en tema prioritario a resolver. Por esto se ideó un plan para ayudar a los alumnos con mayor probabilidad de desertar. Actualmente UDLA no posee un sistema automatizado que clasifique a los alumnos en base a análisis de datos de su comportamiento, sólo se cuenta con un sistema de reglas creado en base al conocimiento de deserción de miembros de la universidad, por lo que tiene una alta tasa de errores.

En el último estudio publicado por el Servicio de Información de Educación Superior sobre retención de alumnos de primer año, construido con datos de alumnos que ingresaron a estudiar el año 2016, la Universidad de las Américas se ubica en la posición 47 de 58 universidades. Por esto, desarrollar un sistema capaz de identificar a los alumnos que estén en riesgo de desertar sigue siendo un tema prioritario para la institución.

El objetivo del presente trabajo es desarrollar un sistema capaz de entregar un índice de riesgo de deserción de cada alumno de primer año. Para esto se propone plantear el proceso de asignar riesgo como un problema de clasificación y afrontarlo con herramientas de inteligencia computacional.

Para resolver el problema se dividió el semestre en tramos y se entrenó un modelo para cada uno de éstos. La precisión del primer modelo fue más baja que la de estudios similares que afrontaron el mismo problema en otras universidades del mundo, teniendo un 70,1% de aciertos. El modelo de cada tramo entregó mejores resultados que los del tramo anterior, siendo el del final del semestre el de mejores resultados llegando a un 82,5% de precisión, lo que se asemeja a otros trabajos.

A mi familia y amigos

Agradecimientos

Quiero agradecer a mis padres por haberme apoyado siempre, en las buenas y en las malas. Por haberse preocupado siempre de que sus hijos tuvieran la mejor educación posible y de que nunca nos faltara nada. Como adulto sé que la crianza no es sencilla y que no existe la forma perfecta de enfrentarla, pero creo que ustedes siempre tuvieron a sus hijos como prioridad, lo que creo que es lo más importante.

A mis hermanas, por estar siempre ahí y demostrarme su amor a diario, aunque a mi me cueste demostrarlo. Por las risas, las bromas y toda una vida compartidos. Porque sé que siempre voy a poder contar con ustedes y siempre van a poder contar en mí.

A Pablito, porque no sé dónde estaría hoy sin un amigo como tú, mi tercer hermano. Gracias por tener siempre tu puerta abierta, por ayudarme sin dudarlo cuando las cosas se han puesto difíciles, por estar siempre listo para levelear, por estar dispuesto a escuchar mis ideas más estúpidas y a llevarlas a cabo entre risas.

A los chiquillos, Cordero, Smith, Ale, Hernán, Narduth, Peyuco, Pipe por tantos buenos momentos y por entregarme felicidad.

A los profesores Andrés Caba y Claudio Held, por permitirme una segunda oportunidad.

A la Dani y Pancho, por siempre hacerme sentir siempre bienvenido, por compartir tantas tardes de juego y hasta de fútbol americano en ocasiones.

A Javier Causa por haberme dado la oportunidad de hacer este trabajo y haber sido un mentor en mis primeros pasos en el mundo laboral.

A Roberto Arce, por ser un excelente jefe, dejarme aprender y apoyarme en este proceso de titulación.

A la Sección 1 del 2008, que aunque ya casi no los vea siempre voy a recordar nuestros primeros años de compañeros.

A Mariet, por empujarme a enfrentar mis miedos y crecer como ser humano.

A mis primos, abuelos y tíos que siempre me han demostrado cariño y aunque me desaparezca a ratos.

Y finalmente a la Choli, que aunque no sea humana y por eso nunca vaya a leer estas palabras, ha afectado mi vida en formas que nunca imaginé y porque no importa si es un buen o un mal día, ella siempre esta a mi lado, siempre lista para jugar moviendo la cola y mirándome con sus ojitos de aceituna.

TABLA DE CONTENIDO

Índice de Tablas

Índice de Figuras

1 INTRODUCCIÓN

1.1 CONTENIDO DE LOS CAPÍTULOS

En esta memoria se muestran la fundamentación, marco teórico, metodología, resultados y conclusiones del desarrollo de un modelo predictivo sobre la deserción de estudiantes de primer año de la Universidad de las Américas.

En el presente capítulo se fundamente el problema describiendo el estado de la universidad y cómo ha llegado al punto en que se encuentra hoy.

En el capítulo 2 se expone la contextualización del problema, incluyendo detalles de la deserción en instituciones de educación superior, el funcionamiento de UDLA en el contexto de la deserción de estudiantes, el marco teórico de las herramientas que se utilizan para afrontar el problema y el estado del arte de otros trabajos que han atacado la deserción universitaria.

En el capítulo 3 se explica la metodología utilizada en este trabajo y los pasos seguidos para llegar al modelo utilizado para la predicción.

En el capítulo 4 se muestran los resultados obtenidos y se realiza un análisis sobre ellos indicando los parámetros encontrados para maximizar el poder de clasificación del modelo.

Finalmente, en el capítulo 5 se muestran las conclusiones obtenidas y las mejoras propuestas para el modelo.

1.2 FUNDAMENTACIÓN

La deserción universitaria de alumnos es un problema presente en todo el mundo, lo que se traduce en costos de distinta índole tanto para alumnos como para las instituciones de educación superior. En el caso particular de la Universidad de las Américas durante los últimos años ha habido una tasa de deserción que ha variado entre un 33% y un 45%, contribuyendo a que la universidad haya perdido su acreditación en el año 2013 por tres años.

A raíz de lo anterior ha habido un cambio en la política de esta institución, haciendo de la retención de alumnos un tema prioritario a enfrentar, motivo por el que en el año 2013 se ideó un plan para mejorar esta situación. Parte de este plan consistió en contactar a los alumnos en riesgo de desertar para ofrecerles apoyo y así evitar su deserción, sin embargo el sistema actual tiene una alta tasa de errores, puesto que está basado en reglas definidas en base a la experiencia de empleados de UDLA sin análisis de datos sobre el comportamiento de los alumnos, haciendo que sea poco fiable.

1.3 OBJETIVOS

1.3.1 Objetivo General

Desarrollar un modelo capaz de predecir qué alumnos de primer año de la Universidad de las Américas están en riesgo de desertar de la institución, utilizando herramientas de inteligencia computacional. El sistema debe ser capaz de recibir información de cada alumno y entregar un índice que refleje qué tan probable es que deserte de la institución durante su primer año de estudio.

1.3.2 Objetivos específicos

- Escoger una metodología para afrontar el problema mencionado
- Estudiar cuáles son las variables más importantes del problema y crear un set de datos que las contenga
- Escoger alguna técnica que permita modelar el problema de manera efectiva
- Obtener resultados comparables a otros estudios realizados

1.3.3 Propuesta

Se propone plantear el problema de detección temprana de desertores de primer año de la universidad como un problema de clasificación con el algoritmo de inteligencia computacional Random Forest [1], que utiliza un conjunto de árboles de decisión para clasificar, utilizando la metodología Knowledge Discovery in Databases [2] que entrega un procedimiento para modelar con inteligencia computacional. Este algoritmo fue escogido por los siguientes motivos:

- En otros estudios [3] [4] [5] [6] se ha establecido que la detección de desertores tiene muchas variables asociadas y esta técnica es capaz de manejar una gran cantidad de variables [1]
- Ha sido calificado por otros trabajos como el algoritmo que, en general, entrega mejores resultados
- Es rápido de entrenar
- Ha sido utilizado por otros trabajos similares
- El ajuste de sus parámetros no es tan importante como para otros algoritmos, haciendo que el sistema pueda ser reentrenado con facilidad en el futuro [1].

Para el desarrollo del modelo se propone utilizar la librería Scikit Learn de Python, puesto que cuenta con una gran cantidad de herramientas para afrontar problemas de Machine Learning, además de ser de código abierto y gratuito, por lo que no traería consigo costos adicionales para la Universidad de las Américas.

2 CONTEXTUALIZACIÓN, MARCO TEÓRICO Y ESTADO DEL ARTE

En esta sección del informe se explica el contexto en el que se desarrolló el trabajo, el problema a resolver, los aspectos teóricos de las diferentes técnicas utilizadas y el estado del arte de otros trabajos similares.

2.1 EDUCACIÓN SUPERIOR EN CHILE

La educación superior en Chile ha tenido cambios profundos a lo largo del tiempo. En los años 70 se contaba con un número reducido de instituciones, por lo que sólo un porcentaje menor de la población podía ingresar.

Con el transcurso del tiempo el acceso a la educación superior en Chile ha tenido un aumento significativo. En el año 1990 la cobertura alcanzaba un 14%, mientras que en el año 2013 ya alcanzaba el 55%, tal como se aprecia en la Figura 2-1, donde se muestra que la población entre 18 y 25 años se mantuvo relativamente constante, pero hubo un marcado incremento en el número de matrículas de pregrado.

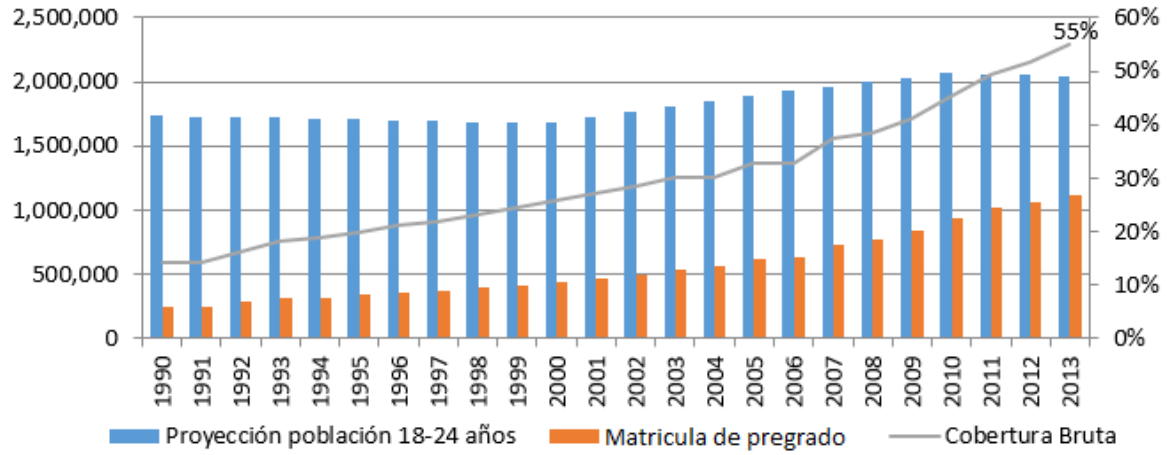

Figura 2-1 Evolución de la cobertura de la educación superior en Chile de la población entre 18 y 25 años desde 1990 a 2013 [7]

Este aumento en la cobertura ha traído consigo cambios en la composición del alumnado, puesto que se han integrado sectores socioeconómicos que antes quedaban al margen. En la Tabla 2-1 se aprecia esta evolución. Se muestran los datos de algunos años entre 1990 y 2011, identificados en la columna de la izquierda, mientras que en la parte superior están los deciles de la población, siendo el primero el de menor ingreso y el décimo el más rico. Se aprecia que en el año 1990 la tasa de ingreso del sector más rico de la población era 12 veces mayor a la del más pobre, mientras que en el año 2011 esta diferencia se había reducido a tres

veces. Esto se traduce en una mayor cantidad de estudiantes con una situación económica desventajada.

Tabla 2-1 Evolución de la Tasa Bruta de Cobertura de la Educación Superior (1990-2011) según decil de ingreso autónomo [8]

	I	II	III	IV	V	VI	VII	VIII	IX	X	Total
1990	4,1	3,5	5,0	7,9	10,2	11,4	14,5	27,0	29,3	47,9	14,3
1998	6,4	8,2	11,4	12,0	19,8	22,2	30,0	44,1	62,5	82,6	27,5
2003	11,0	13,6	15,1	22,8	29,7	34,5	41,2	56,7	84,5	107,2	37,8
2006	15,7	18,5	18,0	26,3	26,2	37,4	41,5	57,5	70,7	90,7	38,1
2009	19,1	20,4	25,1	28,5	31,7	33,7	40,3	55,2	72,6	93,3	39,7
2011	27,2	26,9	32,7	35,0	37,4	42,1	51,8	61,2	78,3	90,9	45,9

2.1.1 Deserción en la Educación Superior

La deserción según Himmel [9] corresponde al abandono de un programa de estudios antes de obtener el título o grado correspondiente, considerando un tiempo suficientemente largo como para descartar la posibilidad de reincorporación. A pesar de esto, Stratton, Ottole & Wetzel [5] diferencian entre la deserción y la suspensión, siendo esta última definida como los casos en que los estudiantes hacen una pausa en sus estudios para luego reincorporarse al sistema. Además, existen casos en que un estudiante deserta de una institución para integrarse a otra, cuyo caso se denomina transferencia. En la Figura 2-2 se muestra un diagrama con la diferenciación recién descrita.

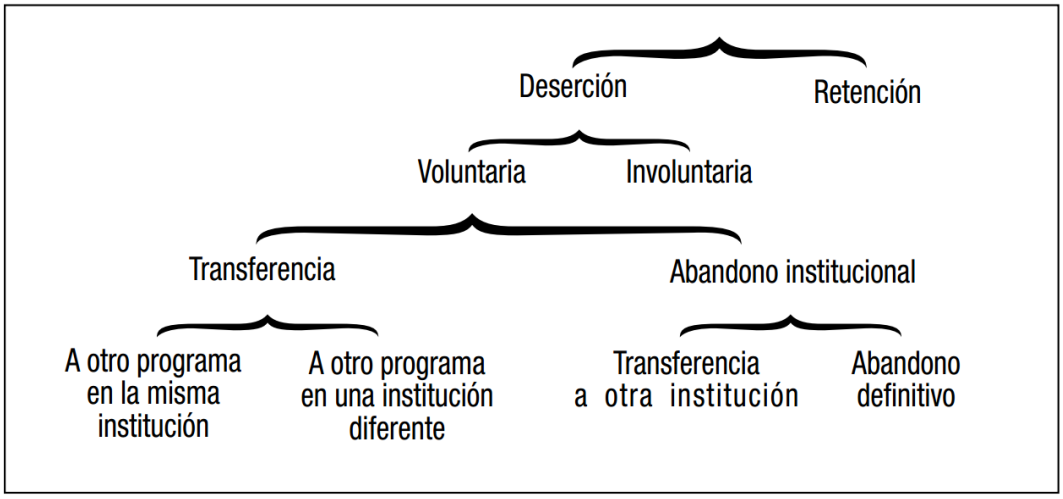

Figura 2-2 Diagrama de los tipos de deserción identificados por Himmel, Stratton, Ottole & Wetzel [10]

Si se define como tasa de deserción al porcentaje de alumnos que deserta del total de alumnos matriculados; de forma análoga puede introducirse el concepto de

retención, que indica el porcentaje que se mantiene en la institución por la duración de su programa de estudios.

En las Figura 2-3 se pueden observar estadísticas generales de la retención de primer año en Chile. En promedio las universidades poseen una tasa de retención que varía entre 75% y un 80% dependiendo del año, mientras que los institutos profesionales y los centros de formación técnica están entre un 55% y un 68% dependiendo del año. Para los tres tipos de instituciones se observa que a partir del año 2011 la tendencia es a aumentar la retención. En la Tabla 2-2 se muestran porcentajes de retención de alumnos de primer año, para universidades chilenas.

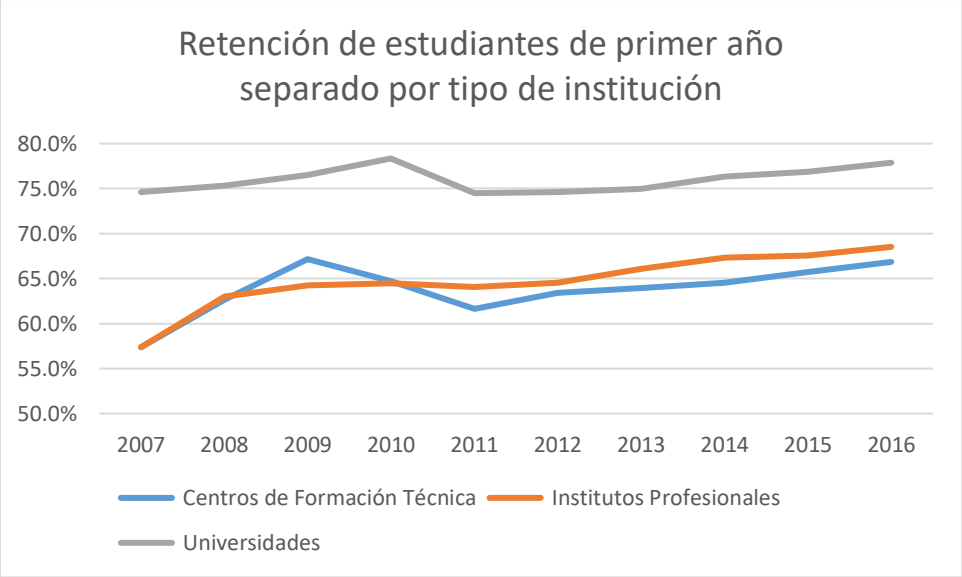

Figura 2-3 Evolución del porcentaje de retención de alumnos de primer año diferenciada por tipo de institución de educación superior construido a partir de [11] y [7]

Los porcentajes de deserción mostrados evidencian que existe un problema que trae consecuencias negativas tanto a alumnos como a universidades. Por el lado de los estudiantes se tiene que el tiempo y dinero del grupo familiar (y en ocasiones dinero estatal) invertido en los estudios se pierde, sin recibir retorno alguno al desertar, junto a una disminución de oportunidades de trabajo y posibles ingresos a futuro del estudiante que deserta. Por otro lado, las universidades también se ven afectadas por la deserción. En primer lugar, si ésta es muy elevada, el prestigio de la institución puede verse dañado, lo que a su vez puede traer consigo otra serie de consecuencias como dificultades para captar buenos alumnos o docentes. En segundo lugar, está la disminución de ingresos que, por ejemplo, puede mermar las posibilidades de realizar inversiones o actividades para enriquecer la formación universitaria. En tercer lugar, asociado a lo anterior, se presenta el problema la determinación de presupuestos al haber incertidumbre sobre el pago de matrículas. Finalmente, la acreditación de carreras con una deserción alta puede complicarse. La pérdida de acreditación trae consigo que tanto los alumnos como la universidad

pierdan beneficios como el acceso al Crédito con Aval del Estado o a la Tarjeta de Nacional Estudiantil para utilizar el transporte público.

Las pérdidas anuales que produce la deserción no pueden calcularse fácilmente, pero González y Uribe [12] estimaron su costo en $47 mil millones de pesos anuales (pesos chilenos del año 1999).

Tabla 2-2 Porcentajes de retención de alumnos de primer año de universidades para el 2016 [11]

Ranking	Universidades	Retención 1er año 2016
1	UNIVERSIDAD DE LOS ANDES	88.7%
2	UNIVERSIDAD AUTONOMA DE CHILE	87.4%
3	UNIVERSIDAD CATOLICA DEL MAULE	87.1%
4	UNIVERSIDAD CATOLICA SILVA HENRIQUEZ	87.0%
5	UNIVERSIDAD ADOLFO IBAÑEZ	86.5%
⋮	⋮	⋮
9	PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE	85.7%
10	UNIVERSIDAD DE TALCA	84.4%
11	UNIVERSIDAD DE TARAPACA	83.5%
12	**UNIVERSIDAD DE CHILE**	**83.2%**
13	UNIVERSIDAD SAN SEBASTIAN	82.6%
⋮	⋮	⋮
25	UNIVERSIDAD TECNICA FEDERICO SANTA MARIA	80.1%
⋮	⋮	⋮
29	UNIVERSIDAD DE SANTIAGO DE CHILE	79.3%
⋮	⋮	⋮
45	UNIVERSIDAD TECNOLOGICA DE CHILE INACAP	70.3%
46	UNIVERSIDAD ACADEMIA DE HUMANISMO CRISTIANO	69.1%
47	**UNIVERSIDAD DE LAS AMERICAS**	**67.4%**
48	UNIVERSIDAD MIGUEL DE CERVANTES	65.6%
⋮	⋮	⋮
56	UNIVERSIDAD SEK	44.6%
57	UNIVERSIDAD DE ARTE Y CIENCIAS SOCIALES ARCIS	38.2%
58	UNIVERSIDAD UCINF	6.1%

2.1.2 Posibles Causas de la deserción según Modelos Teóricos

En los últimos años se han realizado una gran cantidad de estudios para intentar explicar las causas y consecuencias de la deserción universitaria. Estos se han hecho desde diversas disciplinas como la psicología, economía y sociología. A

continuación se presentan a grandes rasgos las posibles causas que podrían llevar a un alumno a desertar según los modelos presentes en la revisión de Himmel [9]:

- Las intenciones de titularse que tenía el alumno al ingresar a su carrera se van debilitando [13].
- La percepción que tienen los alumnos de los cambios en su vida luego de ingresar a la universidad [14]
- Como el alumno percibe su rendimiento académico universitario comparado con el que tenía en el colegio [15]
- La forma en que el alumno percibe que puede sostener los costos asociados a sus estudios [16]
- Los incentivos que tiene para trabajar, comparados a seguir estudiando [16]
- Falta de integración con el entorno de la educación superior [17]
- Influencia de la familia [17]
- Calidad de los profesores y experiencia en la sala de clases [18]
- Disponibilidad de salud y actividades deportivas, culturales, etc. [18]
- Disponibilidad de recursos bibliográficos, laboratorios e indicadores como el número de alumnos por profesor [19]
- Relación del alumno con su institución [20]

Como la gran mayoría de estas variables no son medibles, es necesario encontrar variables relacionadas y medibles que puedan reflejar algunas de estas motivaciones, como la relación entre el rendimiento académico escolar y el universitario, reclamos hacia la universidad, morosidad de pago, entre otras.

2.1.3 Deserción en la Universidad de las Américas (UDLA)

Actualmente UDLA no posee un sistema de detección temprana de alumnos con riesgo de desertar que esté basado en estadísticas sobre sus alumnos. Solo se cuenta con un sistema de reglas desarrollado en el año 2013. Este sistema de reglas se ideó de acuerdo a la experiencia de distintos empleados de la universidad que habían tratado con alumnos desertores en el pasado. Con esta experiencia se idearon reglas para clasificar a los alumnos en del tipo:

- Si el alumno tiene promedio de notas entre 4 y 5 en las cátedras 1 que ha rendido y su asistencia es menor al 30% entonces es un alumno medianamente riesgoso
- Si el alumno tiene cuotas impagas hace más de 90 días entonces es altamente riesgoso
- Si el alumno tiene asistencia mayor al 80%, no tiene cuotas morosas y su promedio en sus cátedras 1 es mayor a 5 entonces es un alumno sin riesgo de desertar.

Como estas reglas suponen un comportamiento muy simple del alumno y no se construyeron en base a un análisis de datos de alumnos desertores de la universidad, este sistema no entrega resultados satisfactorios para UDLA.

Luego de la pérdida de acreditación en el año 2013 se definió como proyecto prioritario de la universidad reducir la deserción de sus alumnos. Esto se ve reflejado en la Figura 2-4, construida a partir de los datos reportados por el Servicio de Información de Educación Superior (SIES), donde se muestra la evolución del porcentaje de alumnos retenidos entre el 2012 y 2016. Se puede apreciar que a partir del año 2013 el porcentaje comienza a subir, pero aún sigue siendo bajo, llegando a un 67,4% de alumnos retenidos en 2016 (equivalente a un 32,6% de alumnos desertores).

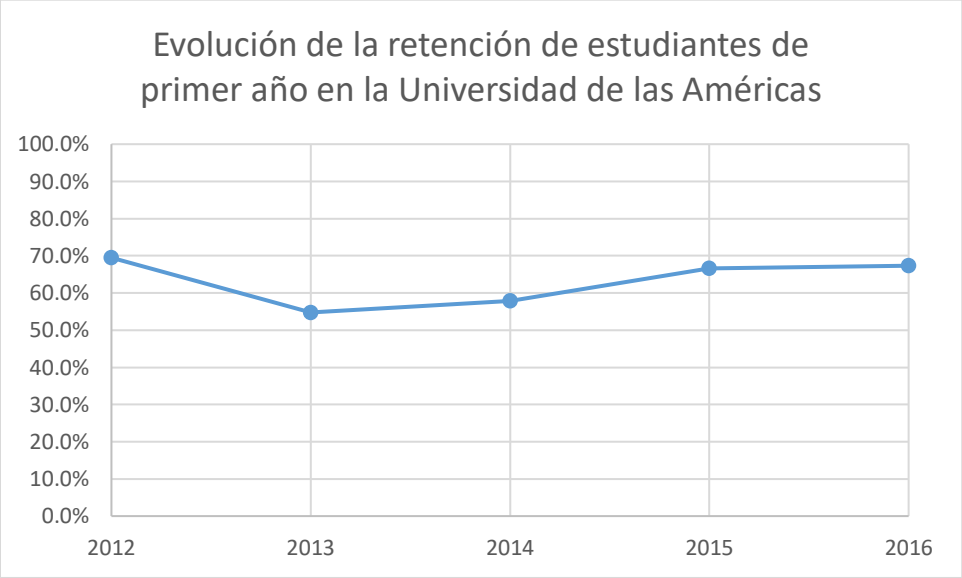

Figura 2-4 Evolución en la retención de alumnos de primer año en la Universidad de las Américas entre 2012 y 2016 [11]

2.1.4 Acreditación en la Universidad de las Américas

La acreditación es un proceso voluntario al que se someten las instituciones de educación superior autónomas del país, así como las carreras de pregrado, programas de postgrado y especialidades del área de la salud que ellas imparten, para contar con una certificación de la calidad de sus procesos internos y sus resultados.

Cuando una institución, programa o carrera están acreditados, cuentan con la certificación otorgada por Comisión Nacional de Acreditación (CNA) respecto de la operación de sus mecanismos de aseguramiento de la calidad y sus resultados. Además, los alumnos nuevos que se incorporan a instituciones acreditadas pueden acceder a financiamiento estatal o recursos que cuentan con garantía fiscal para el

financiamiento de sus estudios. En el caso de los postgrados acreditados, sus alumnos pueden optar a fondos concursables de becas con financiamiento estatal.

La Ley 20.129 [21] establece que los alumnos nuevos de instituciones que no cuenten con la acreditación no podrán acceder a ningún tipo de recursos otorgados por el Estado o que cuenten con su garantía. Los alumnos antiguos que ya contaban anteriormente con este crédito pueden mantenerlo.

En el año 2010 UDLA renovó su acreditación hasta el año 2013. Una vez cumplido este período se le notificó a la institución que perdería esta certificación, situación que luego de una serie de apelaciones por parte de la universidad, fue confirmada a principios del 2014. Posteriormente, la universidad recuperó su acreditación en marzo de 2016 por los tres años siguientes.

2.2 CLASIFICACIÓN

Los seres humanos son capaces de reconocer objetos, leer a partir de caracteres escritos, diferenciar sonidos y realizar otras actividades similares en que se clasifican elementos del entorno. A pesar de que este tipo de tareas se llevan a cabo con facilidad, tienen asociadas una serie de procesos complejos que no se realizan conscientemente, pero que pueden afrontarse debido a que el cerebro ha desarrollado estructuras cognitivas suficientemente sofisticadas para ello.

Con el avance de la tecnología se ha generado un interés marcado en automatizar estos procesos, de modo que puedan ser ejecutados por un computador. En la mayoría de los casos donde se quiere resolver un problema de clasificación, existe un número muy elevado de variables asociadas, con relaciones complejas entre ellas. Por esto, muchas veces una modelación explícita del problema no es posible, se suele utilizar una rama del conocimiento denominada inteligencia computacional [22]. Ésta área se han tratado ampliamente éste tipo de problemas desde reconocimiento de objetos en imágenes a clasificación de clientes en de una empresa.

A grandes rasgos, en un problema de clasificación se tienen objetos con ciertas propiedades medibles, ordenadas en un vector de características $\vec{x}$ de dimensión N y C clases a las que el objeto puede pertenecer. El objetivo de la clasificación es establecer a cuál de las C clases pertenece cada objeto a partir de su vector de características $\vec{x}$. Por tanto, se define una función de clasificación F que va desde un espacio de dimensión N a otro que posee C valores posibles [23]. El dominio y recorrido de F está dado por:

$$F: \mathbb{R}^N \rightarrow \{1, \dots, C\}. \quad (1)$$

Existen diversas metodologías propuestas para afrontar un problema de clasificación, y para este trabajo se escogió Knowledge Discovery in Databases (KDD) [2].

2.3 KNOWLEDGE DISCOVERY IN DATABASES (KDD)

En la publicación original [2] KDD se define como el proceso de identificar patrones válidos, novedosos, potencialmente útiles y entendibles en datos, para luego adaptar un modelo a ellos. La metodología involucra una serie de pasos, que incluyen la preparación de la información, búsqueda de patrones, evaluación del conocimiento, y refinamiento; todos repetidos en múltiples iteraciones. En la literatura se menciona que el proceso en sí contiene 5 fases esenciales y dos adicionales: una previa, de estudio del problema, y una final de integración al sistema donde se utilizará el clasificador. A continuación, se describen estos pasos según se explican en [2]:

1. Estudio del contexto: entender el dominio en el cual se desarrolla la aplicación y definir los objetivos desde el punto de vista del usuario del sistema. Para esto se debe determinar con claridad cuál es el problema a resolver, ya sea fuga de clientes, clasificación de objetos u otro.
2. Selección: definir la base de datos y las etiquetas a utilizar para realizar el estudio.
3. Preprocesamiento: limpiado y preparación de los datos para su uso. Las operaciones que se realizan en esta parte incluyen quitar el ruido, reparar o descartar datos erróneos y definir qué hacer con los datos faltantes.
4. Transformación: reducción y proyección de los datos. Utilizando métodos de reducción de características se puede disminuir el número de variables, mientras que con transformaciones se pueden generar otras que enriquezcan la base de datos haciendo ciertos patrones más evidentes.
5. Data Mining: determinar qué tipo de técnica se utilizará en base a los objetivos que se tengan para la aplicación, por ejemplo establecer si es necesario utilizar un algoritmo de clasificación, clustering o regresión y aplicarlo sobre los datos. Posteriormente se debe escoger el algoritmo particular a utilizar.
6. Interpretación y exploración: análisis de los resultados obtenidos y visualización de ellos, estudio del comportamiento del modelo utilizado e identificación de patrones de interés.
7. Integración: dado que el objetivo de este proceso es resolver un problema para un usuario, es necesario que éste sea capaz de entender, interpretar y utilizar los resultados encontrados, por lo que se debe preparar una plataforma adecuada para él.

En la Figura 2-5 se muestran los cinco pasos básicos del proceso KDD, se puede observar la transición de una fase a la siguiente y lo que se obtiene luego de cada

fase. Además, se representa con las flechas segmentadas la naturaleza iterativa de las etapas del proceso.

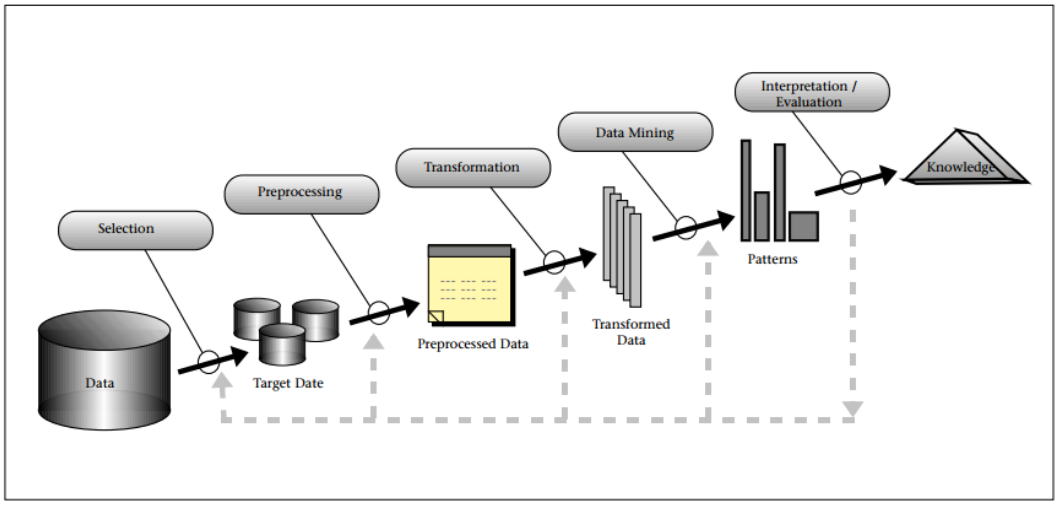

Figura 2-5 Diagrama del proceso Knowledge Discovery in Databases mostrando las fases esenciales del procedimiento. Luego de cada fase se muestra la salida obtenida y en líneas segmentadas la naturaleza iterativa de la metodología [2].

2.4 VALIDACIÓN CRUZADA

Como se explica en [24], al crear un modelo con herramientas de inteligencia computacional que utilice aprendizaje supervisado, es necesario contar con, al menos, dos conjuntos: uno de entrenamiento y otro de prueba, siendo recomendable contar con un tercer set de datos: el de validación. Para explicar el rol de estos conjuntos, es necesario tener claro que al crear un clasificador se deben tener dos objetivos en mente:

1. Selección del modelo: estimación del desempeño de diferentes modelos para escoger el mejor.
2. Valoración del modelo: una vez escogido el modelo final, estimación de su error en la predicción de datos nuevos (error de generalización), simulando la operación de este en el mundo real.

Típicamente los modelos de inteligencia computacional poseen estructura debe ajustarse a los datos del problema a resolver, lo que se logra entrenando el modelo con un subconjunto de los datos llamado conjunto de entrenamiento. Este ajuste se realiza para distintos clasificadores candidatos, siendo común probar un algoritmo particular con distintos valores de sus parámetros, los que corresponden a variables del modelo que se fijan antes de comenzar el entrenamiento.

El conjunto de validación se utiliza para estimar el desempeño de los modelos entrenados y permite escoger el que tenga los candidatos que entreguen el mejor

poder de clasificación. Un casi particular de esto en algoritmos que se entrenan de manera iterativa es detener el proceso de entrenamiento de manera temprana.

El conjunto de prueba se emplea en la valoración del modelo ya escogido, para estimar su desempeño al operar en el mundo real. Este último conjunto no se debe intervenir durante el proceso de selección y sólo se para evaluar el modelo. Es necesario determinar en qué proporción divide el set de datos, siendo común destinar un 50% para entrenamiento, 20% para validación y 30% para prueba.

La utilización de los conjuntos descrita en los párrafos anteriores se conoce como validación cruzada y tiene la característica de evitar que los modelos se sobre-ajusten al conjunto de entrenamiento, evitando que pierdan su capacidad de generalizar para otros datos, problema conocido como sobre-entrenamiento [24].

La separación de conjuntos se suele realizar al inicio de la cuarta fase del proceso KDD, de modo que la selección de variables sólo se base en la información del conjunto de entrenamiento. En [25] se muestra evidencia de que realizarlo de esta forma no produce un sesgo significativo en problemas de clasificación.

Se han realizado estudios sobre el impacto que tiene la separación de conjuntos sobre el desempeño de un modelo, llegándose a la conclusión de que puede introducir una varianza muy elevada en la capacidad de predicción si no se realiza correctamente. Esta varianza puede llegar a ser incluso mayor que la variabilidad asociada a la estructura del modelo [26]. Si los conjuntos no se seleccionan de manera apropiada, podrían no ser estadísticamente representativos del total de datos.

En vista del efecto que puede llegar a tener la separación de conjuntos al momento de seleccionar o valorizar un modelo, existe el riesgo de que patrones importantes no estén presentes en los conjuntos. Si esto ocurre en el conjunto de entrenamiento, el modelo no estará preparado para clasificar correctamente una porción de los datos, mientras que si se pasa en el conjunto de validación o prueba no se estaría evaluando a cabalidad su comportamiento, pudiéndose elegir un modelo inferior a otro.

Simple Random Sampling (SRS) corresponde al método más usado para la segmentación. Esta técnica se basa en una asignación aleatoria de los datos a cada uno de los conjuntos previamente mencionados, mediante una función de densidad de probabilidad uniforme. Este método es adecuado si se tiene un gran número de datos no sesgados.

A partir de la validación cruzada se definió la validación cruzada de k iteraciones donde, en lugar de dividir el set de datos en tres conjuntos, se realizan k divisiones del set de datos. Posteriormente, se escoge una de estas divisiones como conjunto de prueba, se entrena el modelo con las k-1 restantes y se obtiene una evaluación

del modelo. Luego se escoge otra división como conjunto de prueba y se entrena con las k-1 restantes y así sucesivamente hasta completar k iteraciones.

En la Figura 2-6 se muestra un diagrama del funcionamiento de la validación cruzada de k iteraciones para un set de datos que pertenecen a la clase roja o verde. En este ejemplo $k = 4$, por lo que se divide la base de datos en cuatro partes. En la primera iteración se usa el primer cuarto como conjunto de prueba y el resto de entrenamiento, en la segunda se utiliza el segundo cuarto para probar y los otros para entrenar, y así sucesivamente.

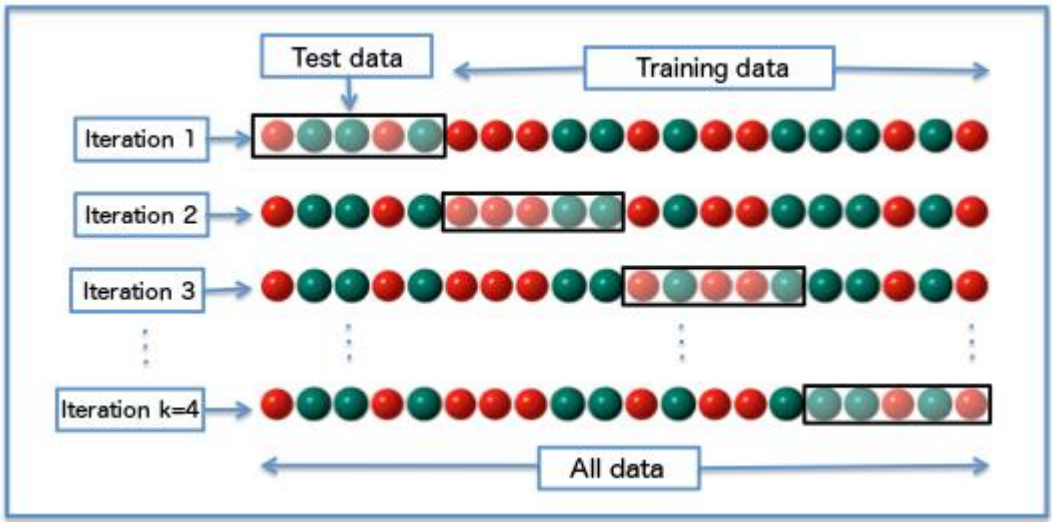

Figura 2-6 Diagrama del funcionamiento de la validación cruzada con k iteraciones con k=4 para un conjunto de datos que pertenecen a la clase roja o verde

2.5 SELECCIÓN DE VARIABLES

La selección de variables es el proceso de escoger un subconjunto de las características disponibles para utilizar en la modelación. Esto se realiza con tres objetivos: mejorar el desempeño del predictor al ser más facilitar su entrenamiento [27] [28]; reducir su costo y tiempo de ejecución; y simplificar el modelo para facilitar su entendimiento.

En las bases de datos suelen haber características redundantes o que no entregan información relevante para la clasificación, por lo que es importante utilizar alguna metodología para eliminar esas características. Una revisión más completa de los diferentes métodos puede encontrarse en [29].

A continuación, se explican brevemente las tres familias de métodos de selección de variables que existen.

2.5.1 Filtro

Esta familia de métodos evalúa cada variable con una métrica independiente del modelo escogido, siendo correlación con la etiqueta, información mutua con la etiqueta o pruebas de significancia las más comunes. La gran ventaja de este tipo de métodos es que son simples, rápidos y seleccionan variables que pueden ser utilizadas con cualquier algoritmo de inteligencia computacional. En la Figura 2-7 se muestra un diagrama de esta metodología, donde se ve que es un proceso sin iteraciones [29]. Aquí se muestra que se parte del conjunto de todas las variables candidatas, luego se realiza la selección y con estas variables se entrena el algoritmo de aprendizaje.

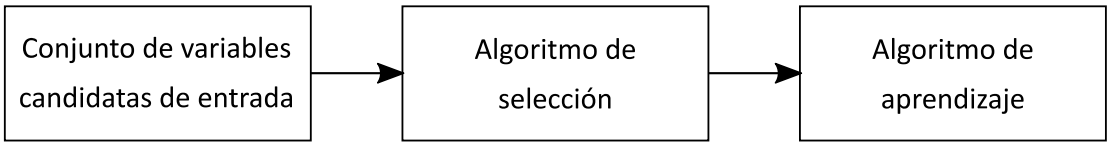

Figura 2-7 Diagrama del funcionamiento de los métodos de selección de variables tipo filtro, los que aplican sobre todo el conjunto de variables una vez, sin iterar

Si bien los métodos de tipo filtro son generales para cualquier proceso y simples de implementar, tienen tres grandes desventajas:

1. Al evaluar las variables aisladas no toma en cuenta las dependencias que podría haber entre ellas
2. Ignora la interacción de las variables con el modelo
3. No es claro el umbral de corte a utilizar para seleccionar las variables [30]. Al aplicar el método se obtendrá un valor que indique qué tan importante es cada variable, pero no hay una regla establecida del valor de corte a partir del que se debe considerar una variable candidata como seleccionada o eliminada.

2.5.2 Iterativa

Esta familia de métodos busca el subconjunto de variables que al introducirse en el modelo escogido entregue la mejor tasa de aciertos, lo que se logra probando todas las combinaciones de variables posibles utilizando los datos del conjunto de entrenamiento. En la Figura 2-8 se muestra un diagrama de esta forma de trabajo: a partir de un conjunto de variables candidato se escoge un subconjunto, se evalúa en el algoritmo de aprendizaje y se evalúa su desempeño, todo de manera iterativa hasta seleccionar un conjunto adecuado.

La cantidad de evaluaciones de estos métodos clásicos crece cuadráticamente con el número de atributos. Si se tienen N características para evaluar, en la primera

iteración se probarán cada uno de los atributos individualmente, es decir N. Luego se probarán $N-1$ combinaciones de pares de variables, luego $N-2$ grupos de 3 variables y así sucesivamente hasta agotar todas las posibilidades. El total de iteraciones que se debe calcular con esta forma de selección de variables se puede calcular con la siguiente expresión

$$\sum_{\{i=0\}}^{N}(N-i)=\frac{1}{2}\times N(N+1). \tag{2}$$

Por lo anterior, uno de los grandes problemas de este tipo de selección de variables es su naturaleza tipo fuerza bruta, volviéndose computacionalmente inviable aplicarlo sobre una cantidad elevada de variables: no solo se deben seleccionar las variables, sino que además se debe entrenar y evaluar el modelo para cada conjunto [29].

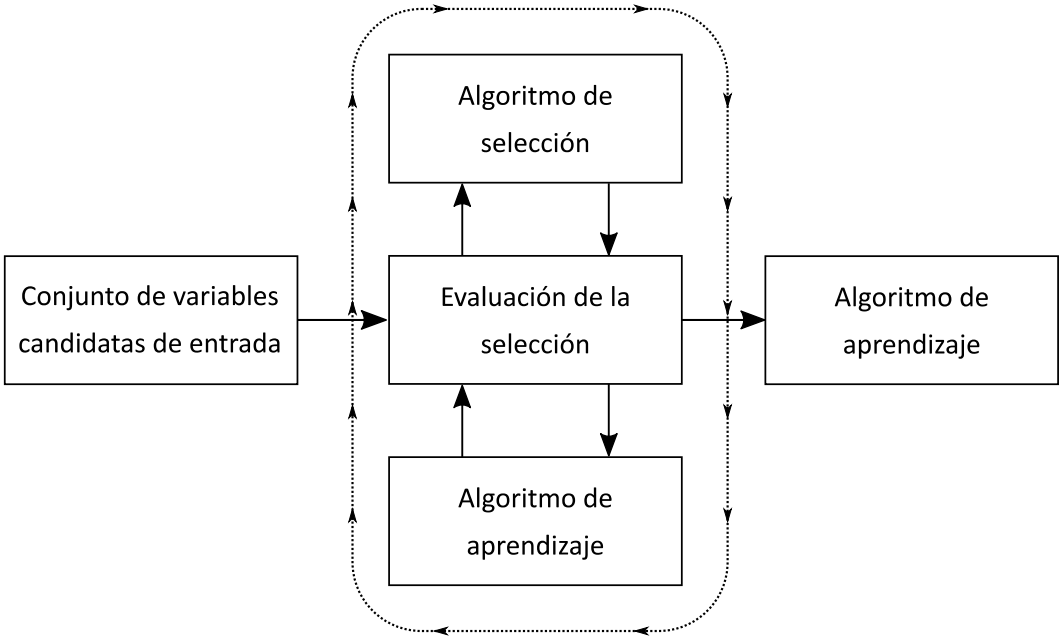

Figura 2-8 Diagrama del funcionamiento de los métodos de selección de variables iterativos donde se muestra que a partir del conjunto de variables candidatas se aplica el algoritmo de selección, el que luego es evaluado entrenando el algoritmo de aprendizaje con estas variables. Este proceso es iterativo lo que se representa con las líneas punteadas para finalmente escoger el conjunto de variables mejor evaluado.

2.5.3 Métodos embebidos

Este tipo de métodos de selección de características están incluidos en algunas técnicas de modelamiento, por lo que no es necesario utilizar ninguna de las técnicas descritas en las secciones anteriores [29]. Debido a que son específicos del algoritmo de clasificación no tienen una forma estándar. Algunos ejemplos de clasificadores que incluyen selección de variables son random forest y árboles de

decisión, que durante el entrenamiento van seleccionado las variables que de acuerdo al modelo aportan más a la clasificación.

2.6 USO DE VARIABLES CATEGÓRICAS

Una variable categórica contiene etiquetas en lugar de números, por lo que no hay una relación ordinal entre los distintos valores que posee. Un ejemplo para una variable que incluya en color de un objeto se muestra en la Tabla 2-3.

Tabla 2-3 Ejemplo de una variable categórica

Nº ejemplo	Color
1	Azul
2	Azul
3	Verde
4	Rojo
5	Verde

Muchos algoritmos (o implementaciones de algoritmos) no permiten trabajar con variables categóricas directamente, necesitan que todas sean numéricas. Una opción reemplazar los valores por números designados arbitrariamente, por ejemplo, reemplazando 'Azul' por 1, 'Verde' por 2 y 'Rojo' por 3, pero muchas veces esto puede producir que el algoritmo utilizado para modelar el problema llegue a encontrar relaciones de ordinalidad que no debiesen existir, por lo que en Machine Learning y estadística se suele utilizar one-hot encoding, donde si una variable categórica contiene n categorías se transforma en n variables binarias como lo muestra la Tabla 2-4, la que se construyó a partir de la información de la Tabla 2-3. Esta y otras técnicas de codificación de variables se pueden encontrar en [31]

Tabla 2-4 Ejemplo de one-hot encoding utilizando los valores mostrados en la Tabla 2-3

Nº ejemplo	Azul	Verde	Rojo
1	1	0	0
2	1	0	0
3	0	0	0
4	0	0	1
5	0	1	0

2.7 VALORES FALTANTES

Muchas veces en las bases de datos hay atributos sin valor asignado para algunas muestras, por diferentes razones. En la literatura existe una taxonomía para tratar estos casos de tres formas diferentes [32]:

- Missing Completely at Random (MCAR): la causa de que el valor no esté presente es independiente tanto de las demás variables que sí tienen un valor como del valor que no está presente: la probabilidad de que un valor falte es completamente aleatoria. Un ejemplo de esto sería que algunos alumnos no tuvieran ingreso familiar registrado porque se perdieron algunos formularios donde estaba esa información.
- Missing at Random (MAR): el motivo por el que el valor no se encuentra no depende de los datos faltantes, pero está relacionado a los valores de otra variable: la probabilidad de que falte un valor para una variable dada está condicionada por otra. Un ejemplo de esto sería que algunos alumnos de ciertas comunas no tuvieran su ingreso familiar porque se rehusaron a informarlo con lo que se puede establecer una relación entre la comuna y la probabilidad de que la información no esté disponible.
- Not Missing at Random (NMAR): los valores no están presentes por una razón específica, por lo que el hecho de que el valor no esté entrega información por sí misma. Un ejemplo de esto sería que a todos los alumnos con promedio sobre 5 en pruebas parciales de una asignatura les faltara la nota de examen, porque se les ofreció eximición de esta evaluación.

Dado que hay modelos que no permiten incluir valores faltantes y el hecho de que un valor no esté presente podría entregar información, existen dos familias de métodos para tratar con valores faltantes. La primera familia corresponde a la eliminación ya sea de las columnas (variables) con valores faltantes o de las filas (muestras) con la misma condición. Esto puede ser de utilidad cuando la variable tiene un porcentaje muy elevado de información no disponible o cuando hay muestras que tienen muy poca información, respectivamente. El segundo grupo de técnicas para tratar con valores faltantes se denomina imputación y consiste en rellenar los datos faltantes por otros a mediante algún método como la media, mediana o algún valor que refleje la razón por la que el valor falta, en casa de que esta razón se pueda deducir [32].

2.8 MÉTODOS DE CLASIFICACIÓN

Para clasificar las muestras en la base de datos de manera automática es necesario contar con un método de clasificación, al que se le deben encontrar los parámetros óptimos para adecuarse a los datos disponibles. En este trabajo se utilizó el método random forest. Para explicarlo primero es necesario entender cómo funciona un árbol de decisión.

2.8.1 Árboles de decisión

Es uno de los métodos de clasificación no métricos, es decir, permite clasificar incluso con variables que no tienen una medida de similitud, por lo que se pueden clasificar datos nominales. Los árboles de decisión se construyen con una serie de preguntas sucesivas sobre las diferentes variables, ya sea con respuestas tipo 'verdadero o falso' o $valor(variable) \in [rango\ de\ valores]$. En la Figura 2-9 se muestra un ejemplo de un árbol de decisión, donde se busca clasificar una fruta según distintas propiedades. Cada nodo de este árbol contiene una pregunta sobre alguna variable, que puede ser respondida de manera afirmativa o negativa [23].

Las preguntas se realizan a medida que se va siguiendo un camino comenzando por el nodo raíz, que por convención se muestra en la parte superior de la representación de su estructura. Este nodo se conecta a otros mediante ramas, los que a su vez se conectan a otros y así sucesivamente hasta llegar a las hojas o nodos terminales, mostrados en la parte inferior de la representación del árbol.

En el nodo inicial se realiza una pregunta sobre una variable particular y los diferentes caminos que conectan a este nodo con los demás corresponden a las diferentes respuestas que puede tener esta pregunta. Al ingresar una muestra en el árbol, dependiendo del valor que tenga esta variable se elige uno de los caminos, para luego realizar otra evaluación con otra variable y así sucesivamente hasta que ya no haya más preguntas que responder, llegándose a un nodo hoja. Toda muestra que sea clasificada por el árbol llegará a un nodo hoja y se el asignara la clase que tenga mayor presencia de ese nodo. En el caso de los árboles utilizados en este trabajo, cada nodo tiene caminos que son excluyentes entre sí y que en conjunto consideran todas las respuestas posibles.

Para el ejemplo de la Figura 2-9, si se tiene una sandía se seguiría el siguiente trayecto partiendo del nodo raíz:

1. ¿Color = verde? Sí
2. ¿Tamaño = grande? Sí
3. Sandía

Existen ciertos criterios para entrenar el árbol, de modo que se determinen las mejores variables para separar cada categoría, cuantos nodos se utilizarán, etc [23]. Al entrenar el árbol se va evaluando la capacidad de separación de cada variable con alguna métrica, escogiéndose la que separe mejor en cada nodo. Existen dos alternativas comúnmente usadas para realizar esta evaluación: la impureza de Gini y entropía. En este trabajo se escogió la impureza de Gini, que es una medida de cuán probable es que un elemento de un nodo sea clasificado de manera incorrecta. Dado un conjunto que posee J clases, con $i \in \{1,2,\dots J\}$ y f_i la fracción de elementos pertenecientes a la clase i, la impureza de Gini se puede calcular como:

$$I_G(f) = \sum_{i=1}^{J} f_i(1 - f_i) \tag{3}$$

$$= \sum_{i=1}^{J} f_i - f_i^2 \tag{4}$$

$$= 1 - \sum_{i=1}^{J} f_i^2. \tag{5}$$

Cabe notar que, si el conjunto sólo posee una clase, la impureza de Gini será 0. Por lo tanto, mientras mejor separe un nodo, menor será su impureza.

Las ventajas más grandes de los árboles de decisión en comparación con otros algoritmos de clasificación son su simplicidad, robustez frente a la inclusión de variables irrelevantes, ser invariante al escalar el problema, rapidez de construcción y posibilidad de interpretación. Respecto a esto último, al tenerse una muestra ya clasificada por el árbol es sencillo identificar las razones por las que se llegó a ese resultado, debido a que sólo se debe buscar el camino seguido para llegar al nodo hoja. Siguiendo el camino para llegar a cada nodo hoja es posible crear descripciones de los diferentes motivos por los que se llega a cada clasificación, aunque si el árbol es muy complejo su interpretabilidad disminuye.

La interpretabilidad de los modelos creados con esta técnica es muy valiosa al analizar los resultados y para justificar las clasificaciones, lo que podría ser utilizado por el usuario del sistema. Por ejemplo, en el caso de la deserción de estudiantes se podría determinar el motivo por el que el alumno tiene mayor probabilidad de desertar y así intentar ayudarlo para evitar que deje la institución (ofrecer becas si tiene problemas económicos, reforzamiento si es por rendimiento, etc.).

Por otro lado, la simplicidad del modelo muchas veces genera resultados inferiores con respecto a otras técnicas y si se construye un árbol muy complejo tenderá a sobreajustarse al conjunto de entrenamiento [23].

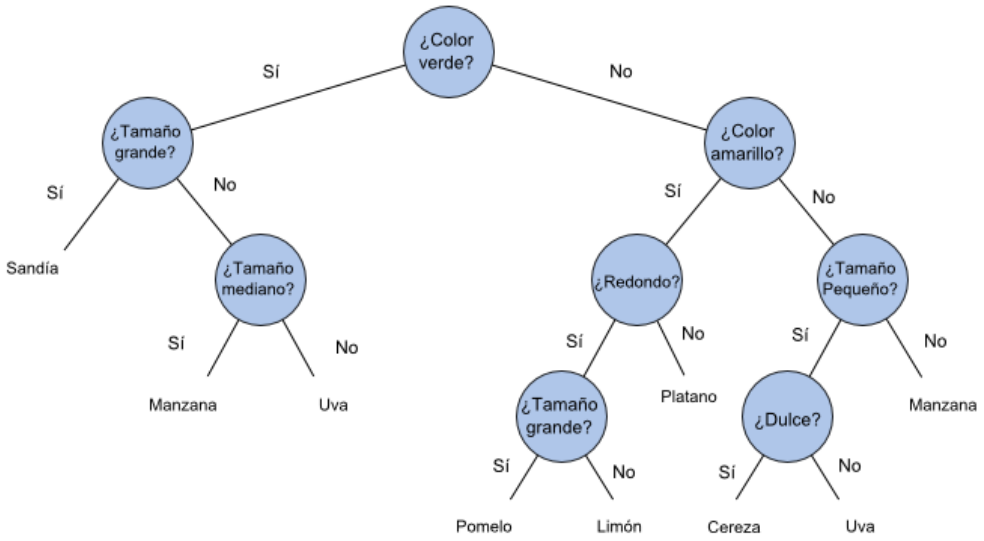

Figura 2-9 Ejemplo de árbol de decisión donde se clasifican frutas según algunas características [23]

2.8.2 Random Forest

Random Forest es una técnica que fue propuesta por Leo Breiman en 2001 [1], que busca solucionar la baja precisión y tendencia a sobreajuste al conjunto de entrenamiento de los árboles de decisión. Este es un método de clasificación en el que se construyen múltiples árboles de decisión complejos, sobre distintas partes del conjunto de entrenamiento, cuyas salidas se combinan con algún método (moda o predicción media) para la predicción final, lo que tiende a mejorar los resultados, pero pierde la facilidad de interpretar los resultados que tiene un árbol de decisión convencional.

El parámetro principal a escoger de un random forest es el número B de árboles que se utilizarán. Si se tiene un conjunto de entrenamiento que tiene N observaciones y P variables cada una, se realiza lo siguiente para cada árbol b del random forest:

1. Se toma un subconjunto aleatorio $\boldsymbol{Z}^*$ de tamaño L del conjunto de entrenamiento, donde cada elemento de $\boldsymbol{Z}^*$ es elegido al azar con reemplazo utilizando cierta distribución de probabilidad, usualmente uniforme. Por esto hay observaciones que pueden repetirse más de una vez y otras que pueden nunca ser elegidas.
2. Construir un árbol aleatorio T_b utilizando $\boldsymbol{Z}^*$ como entrenamiento repitiendo recursivamente los siguientes pasos para cada nodo del árbol, hasta que se llegue a algún criterio para detenerse:
    a. Seleccionar m de las P variables de manera aleatoria, donde m es un parámetro del modelo.

b. Elegir la mejor variable/punto de separación de entre las m
c. Separar el nodo en dos nodos hijos

Con esto se obtiene el conjunto de árboles $\{T\}_1^B$. Para garantizar que se utilice todo el conjunto de entrenamiento es necesario escoge un valor de B suficientemente grande.

Cabe destacar que el algoritmo introduce aleatoriedad en la construcción de cada T_b, puesto será entrenado con un subconjunto aleatorio $\boldsymbol{Z}^*$ de observaciones dentro de las N posibles, y el punto de separación cada nodo n_{ib} de T_b será escogido a partir de un subconjunto m de las variables. Esto genera una alta varianza en los árboles del random forest, cada árbol probablemente será diferente de los demás dada la naturaleza inestable de un árbol de decisión, donde una variable más o menos puede cambiar drásticamente la configuración de nodos. En la Figura 2-10 se muestra un diagrama de lo explicado anteriormente.

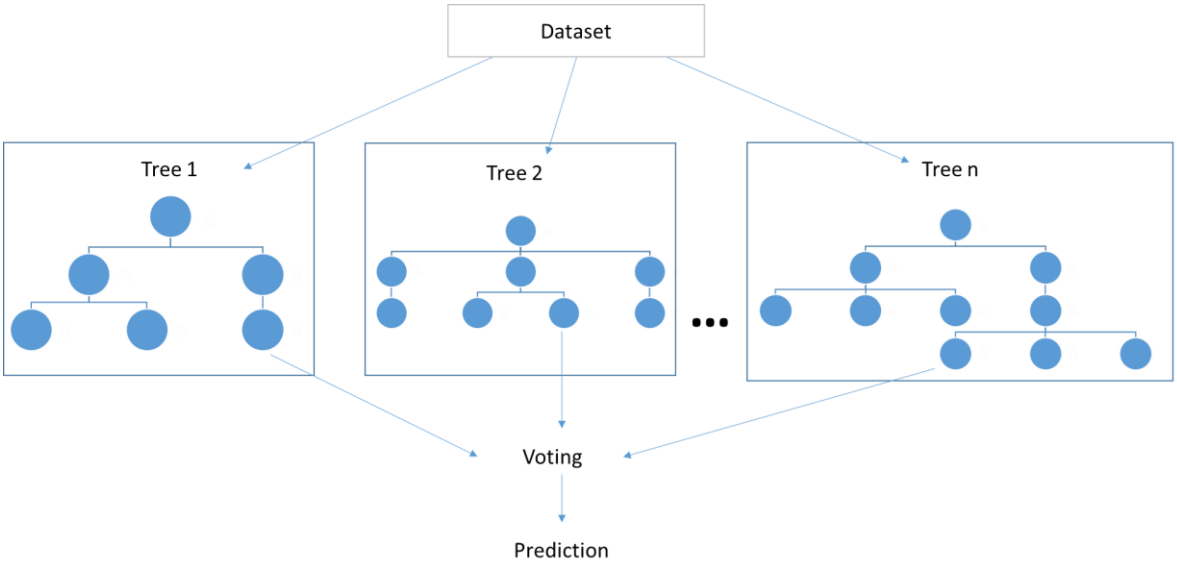

Figura 2-10 Diagrama del entrenamiento y estructura de un random forest. Se muestra que cada árbol se entrena con una porción de los datos y tienen estructuras completamente diferentes entre sí. Cuando una muestra es clasificada por el random forest se realiza algún tipo de votación con la salida de cada árbol.

Al agregar más árboles al random forest se reduce la varianza de la salida, haciendo que el resultado converja a cierto error de generalización. Un ejemplo de esto se muestra en la Figura 2-11, donde se observa que entre 2 y 20 árboles el error decrece, pero al agregar más ya no hay decrecimiento. La demostración de la reducción en el error al aumentar la cantidad de árboles puede encontrarse en el apéndice I de [1], mientras que en [33] se realiza un análisis formal y más específico de algunas propiedades matemáticas del algoritmo.

Figura 2-11 Gráfico de error cuadrático medio de clasificación vs N° de árboles para un bosque aleatorio. Elaboración propia.

Si bien un modelo de random forest no es igual de interpretable que un árbol de decisión puede entregar estimaciones de la importancia de cada variable, lo que puede usarse como selección de características. Para esto existen dos métodos propuestos:

1. Disminución media de la precisión: Una vez que se ha construido el random forest, se realiza una clasificación de datos, obteniéndose una clase de salida y_n para cada vector de entrada x_n. Luego se permuta aleatoriamente el orden de todos los valores de una variable p, con el objetivo de que la variable pase a ser ruido. Con este nuevo conjunto de datos que contiene la variable alterada, se realizan clasificaciones de sus elementos con el random forest entrenado. Luego esto se repite para todas las demás variables y se obtienen nuevas clasificaciones, y se compara cómo cambia la tasa de clasificaciones incorrectas del random forest al alterar cada variable, obteniéndose una tasa de cuán importante es cada una para el modelo.
2. Disminución media de la impureza de Gini: En cada nodo de cada árbol se puede calcular cuánto se reduce la impureza al dividir los datos con la variable seleccionada, luego todos los valores de reducción en la impureza se suman y se determina en promedio cuánto aporta cada variable a la disminución de la impureza [34].

Ambas métricas entregan buenos resultados experimentales, pero tienen algunas desventajas que deben tomarse en cuenta: la primera métrica puede dejar fuera variables con fuerte correlación con otra, pero que entregan información nueva; mientras que la segunda métrica puede ignorar variables que no aparecen en muchos árboles dentro del random forest. Suele preferirse la primera, pero depende de la aplicación.

Random forest es muy poderoso siempre y cuando el problema cuente con una cantidad suficiente de datos, de modo que se pueda introducir suficiente aleatoriedad en el entrenamiento. Tanto así que en los últimos años ha habido publicaciones [35] [36] que plantean si es la mejor familia de clasificadores.

2.9 PRESENTACIÓN DE RESULTADOS

Para presentar los resultados de este trabajo se utilizan dos técnicas, las que se explican a continuación.

2.9.1 Curva ROC

Muchos clasificadores entregan como salida un número que indica la certeza de que la muestra corresponda a una clase, por lo que se suele aplicar un umbral θ para determinar la clase a la que será asignada; si el valor de la salida es mayor que el umbral se le asigna una de las clases, de lo contrario a la otra. Si se define a x como la la muestra para la que se encontró una certeza $p(x)$, su clase se obtiene con la siguiente expresión:

$$si\ p(x) \geq \theta \Rightarrow Clase\ 1 \tag{6}$$

$$si\ p(x) < \theta \Rightarrow Clase\ 2 \tag{7}$$

Si se define como positiva a la clase buscada y la otra como negativa, dada la naturaleza de los problemas con dos clases, existen cuatro resultados posibles para cada muestra clasificada:

- Verdadero positivo (VP): La muestra es positiva y se clasifica como positiva
- Falso negativo (FN): La muestra es positiva y se clasifica como negativa
- Falso positivo (FP): La muestra es negativa y se clasifica como positiva
- Verdadero negativo (VN): La muestra es negativa y se clasifica como negativa

La tasa de verdaderos positivos (TVP) o sensibilidad es la razón de ejemplos de la clase buscada clasificados correctamente, se define de la siguiente manera:

$$TVP = \frac{VP}{VP + FN}, \tag{8}$$

donde la suma $VP + FN$ es igual al total de muestras positivas. Por otro lado, la tasa de falsos positivos (TFP), también conocida como tasa de falsas alarmas, es la razón de veces que ejemplos son clasificados erróneamente como de la clase buscada, se define como:

$$TFP = \frac{FP}{FP + VN}, \tag{9}$$

donde la suma $FP + VN$ es el total de muestras negativas. Una curva ROC es un gráfico de dos dimensiones en que se grafican los valores de TVP y TFP para distintos valores de θ, ubicándose en los ejes de la abscisa y ordenada respectivamente. Se utiliza para determinar qué tan efectivo es un clasificador para distintos umbrales y también sirve para visualizar su desempeño de modo que pueda compararse con otros y así seleccionar el más adecuado. En general se utiliza para evidenciar el *trade-off* entre los aciertos y las falsas alarmas del predictor en problemas de dos clases. En la Figura 2-12 se muestran varios ejemplos de curvas ROC con distintas capacidades de discriminar clases.

Una curva ROC tiene puntos importantes a considerar. El punto de abajo a la izquierda (0,0) representa la estrategia de clasificar todas las muestras como negativas, con lo que nunca se equivoca en predecir los casos negativos, pero a su vez no detecta a los positivos. En el punto (1,1) se da la situación contraria, donde se ubicaría un clasificador cuyo umbral clasifique todo como positivo, haciendo que se equivoque en todos los negativos.

Lo deseable es que la curva tienda a pasar cerca del punto (0,1) pues en ese caso la clasificación será perfecta, lo que en las aplicaciones reales nunca ocurre, por lo que usualmente se elige un umbral que se acerque a este punto. En la Figura 2-12, la recta a representa un clasificador que asigna las clases de la base de datos de manera aleatoria, por lo que un modelo funcional debe pasar por encima de esta.

El punto óptimo a utilizar depende de los costos asociados a los diferentes errores y los beneficios de clasificar correctamente. Por ejemplo, si una clasificación correcta es muy valiosa en relación a no detectarla o a obtener un falso positivo, como en el caso de la detección de tumores, los falsos negativos serán muy costosos, por lo que se priorizará la detección [37].

Una métrica adicional que se define para evaluar el modelo es el valor predictivo positivo (PPV) o precisión en recuperación de la información, que indica qué porcentaje de las predicciones positivas son fueron acertadas. Se calcula con la siguiente expresión:

$$PPV = \frac{VP}{VP + FP}. \tag{10}$$

Una medida que suele utilizarse para comparar modelos es el área bajo la curva ROC [23], que en caso de ser un clasificador perfecto es 1 y en caso de no poder discriminar sería 0,5, por lo que su valor real siempre está entre estos valores.

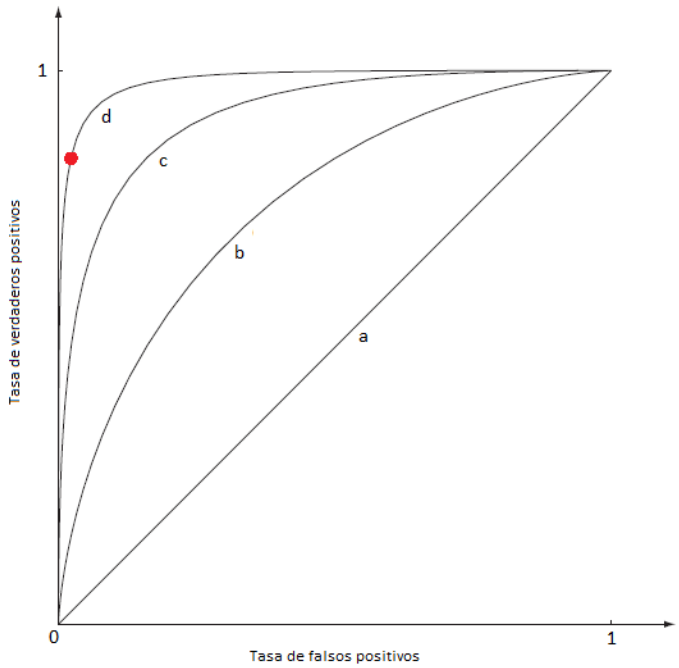

Figura 2-12 Ejemplo de cuatro curvas ROC para clasificadores distintos, donde la que posee mayor poder de clasificación es el ejemplo d, mientras que a corresponde a la recta que se obtiene al clasificar aleatoriamente las muestras, haciendo que la tasa de verdaderos positivos sea igual a la de falsos positivos.

2.9.2 Matriz de confusiones

Una matriz de confusiones también se utiliza para evaluar modelos de clasificación, y puede utilizarse para casos donde hay más de dos clases. Esta matriz es de dimensiones $n \times n$, donde n es la cantidad de clases con las que se está trabajando. Los números en cada columna representan la cantidad de veces que se clasificaron muestras como pertenecientes a una clase, mientras que en el caso de las filas corresponden a la cantidad de muestras que efectivamente pertenecen a cada clase.

En la Tabla 2-5 se muestra la composición de una matriz de confusiones. En la casilla (1,1) se ubica el número total de muestras clasificadas como pertenecientes a la clase 1 y que efectivamente pertenecen a esa clase, mientras que en la casilla (1,2) se muestra el total de muestras clasificadas como 1 pero que en realidad pertenecen a la clase 2, y así sucesivamente.

La mayor utilidad que entrega esta técnica es que permite determinar en qué medida el clasificador está confundiendo dos clases, además de entregar las clasificaciones exitosas en la diagonal.

Tabla 2-5 Diagrama de la composición de una matriz de confusiones

		Predicción: Sí	Predicción: No
Clase real	P	Verdaderos positivos (VP)	Falsos negativos (FN)
	N	Falsos Positivos (FP)	Verdaderos negativos (TN)

En la Tabla 2-6 se muestra un ejemplo de matriz de confusión para tres clases: A, B y C. En la posición $(3,3)$ de la matriz se encuentra la tasa de clasificaciones de datos como C cuando pertenecían a esta clase. La tasa de precisión total se encuentra al sumar todos los elementos de la diagonal y dividirlos por el total de clasificaciones.

Tabla 2-6 Ejemplo de una matriz de confusiones con tres clases

	A	**B**	**C**
A	25	5	10
B	10	20	0
C	5	5	20

$$Precisión = 75\%$$

2.10 ESTADO DEL ARTE DE LA PREDICCIÓN DE DESERCIÓN DE ESTUDIANTES

A continuación, se hace una revisión a grandes rasgos de estudios similares al hecho en este trabajo. En [38] se realizó un estudio sobre los datos de la Universidad de Santander con datos históricos de los años 1986 a 1999, incluyendo información básica del estudiante como facultad, programa de estudio, género y área de origen, entre otras; además de información del rendimiento académico como el promedio de notas acumulado o el resultado de la prueba utilizada para ingresar a la universidad. En este trabajo se aplicó clustering para obtener una primera lectura de los datos y se analizaron los centroides de cada cluster para obtener

descripciones cualitativas. Se clasificó cada cluster de acuerdo a sus características, por ejemplo, *alumnos con buen desempeño académico* o *alumnos con mal desempeño académico y desertores.* Luego, para obtener un modelo de clasificación se entrenó un árbol de decisión con reglas sencillas del tipo:

- Si el alumno tiene entre 20 y 25 años entonces no es desertor
- Si el alumno tiene entre 25 y 30 años y está en primer año es desertor

Un objetivo del trabajo fue que el modelo fuese interpretable por sus usuarios, por lo que se priorizó su simplicidad. El trabajo no entrega matrices de confusión u otras medidas de su desempeño, solo menciona que la universidad para la que se realizó el estudio lo evaluó con una nota 7,94 de 10.

En [3] se realizó un estudio sobre la retención de los alumnos de ingeniería y ciencias de la Virginia Commonwealth University mediante una red neuronal. En este se construyeron dos modelos independientes: uno para predecir la retención de los alumnos nuevos y otro para clasificar al mismo grupo en una de tres categorías: riesgoso, intermedio y avanzado. Una de las variables más importantes de este modelo es el grade point average (GPA), cuyo equivalente sería el promedio de notas. El resultado de este modelo tuvo resultados muy prometedores, pero la base de alumnos utilizada es pequeña, siendo cercana a 300, por lo que es posible que esté sobreentrenado y no pueda generalizar.

Otro estudio [6] se llevó acabo con datos de estudiantes de enseñanza media de Dinamarca. Este modelo tiene una base de datos de gran tamaño, con 72.598 estudiantes de los cuales el 23,8% desertó y utilizaron el 50% para el conjunto de entrenamiento y el otro 50% para el de prueba. El trabajo expone que de acuerdo a investigaciones realizadas por los autores no se deben usar las mismas variables al modelar alumnos que comienzan sus estudios a cuando ya llevan un tiempo haciéndolo, por lo que solo se enfoca en estudiantes que ya llevan al menos 6 meses de enseñanza media. A cada alumno se le calcularon un conjunto de variables obtenido en una fecha determinada, la que se escogió aleatoriamente para los alumnos no desertores y tres meses antes de la deserción para alumnos desertores. El algoritmo que les dio mejores resultados fue Random Forest llegando a un 93,5% de aciertos con un área bajo la curva de la curva ROC de 0,965.

El trabajo citado en [4] tenía como objetivo comparar distintos modelos para predecir retención de alumnos y se realizó con datos de una universidad de estados unidos, considerando los datos de 16.066 estudiantes enrolados entre 2004 y 2008 con un 40% de deserción a lo largo de toda la carrera universitaria. El trabajo comparó varios métodos: SVM, Random Forest, Árboles de decisión y otros. En general todos los algoritmos obtuvieron resultados similares y en particular con Random Forest obtuvieron una precisión del 81,8%. Para evaluar los distintos algoritmos se siguió utilizó validación cruzada con 10 iteraciones.

En el estudio [39] también se comparan distintos modelos predictivos para predecir la retención de estudiantes en la Universidad de Saint Cloud State (SCSU), donde el 29,33% del os alumnos desertaron. Los modelos se entrenaron con variables que según los autores pueden agruparse en qué tan preparado está el alumno para la educación superior, su capacidad académica, situación financiera y resultados de su primer año de estudio. Todas las variables contaban con algún número de datos faltantes, los que fueron tratados de diversas maneras dependiendo de cada una, por ejemplo algunos se reemplazaron por 0, otros por el promedio, otros por el máximo y en el trabajo se detalla lo que se hizo con cada una. Las variables se seleccionaron usando PCA y todas ellas fueron utilizadas en los 6 modelos entrenados, siendo random forest el que obtuvo los mejores resultados, clasificando correctamente al 85,87% de los estudiantes.

El año 2017 se realizó una tesis de magíster para tratar el tema de deserción de estudiantes en la Universidad de Chile [40], que pudo predecir al 74,23% de los estudiantes que desertaron en primer año. El trabajo utilizó la metodología KDD y consideró variables de diversa índole como Nota de Enseñanza Media, puntaje en la PSU de lenguaje, puntaje en la PSU de matemáticas, puntaje ponderado de la PSU, ingreso bruto, becas, género, número de integrantes del grupo familiar y si están vivos sus padres.

Un área de estudio similar a la deserción de estudiantes es la deserción de clientes. En el trabajo [41] publicado en 2018 se hace una revisión de 10 algoritmos de Machine Learning sobre una base de datos de 3333 registros de clientes de una empresa de telecomunicaciones con un 14% de desertores. Las variables utilizadas por los modelos son distintas métricas de la actividad de los usuarios, como por ejemplo número de minutos diarios utilizados, número de minutos nocturnos utilizados, cantidad de mensajes de voz, antigüedad de la cuenta, entre otras. La conclusión del estudio es que random forest y otra técnica similar llamada Ada boost entregaron los mejores resultados con una precisión de 96%.

3 METODOLOGÍA

A continuación se explica el procedimiento seguido en este trabajo, ordenado según la estructura de la metodología KDD.

3.1 ESTUDIO DEL CONTEXTO

Tal como se estableció en la sección 2.3 lo primero que se hizo al seguir la metodología KDD fue definir los objetivos y examinar el contexto en el que está inserto el problema a resolver.

3.1.1 Definición de objetivos

Se definió como objetivo implementar un sistema capaz de predecir qué alumnos de primer año se retirarían a lo largo del año, el que debía ejecutarse a lo largo del año y entregar un índice de riesgo para cada alumno, de modo que los directores de carrera pudieran acceder a los resultados y contactar a los alumnos más riesgosos, para evitar su deserción.

3.1.2 Antecedentes de UDLA

A continuación se describen algunos aspectos del funcionamiento de UDLA que son de interés para realizar el análisis de deserción de estudiantes.

3.1.2.1 Bases de datos

UDLA cuenta con un sistema de bases de datos con información de diversas fuentes de la universidad, como el sistema de planificación de la universidad, el sistema de gestión de bibliotecas, Aulas Virtuales, el sistema de ingreso de notas y otras más, donde cada fuente cuenta una con su propia forma de funcionamento. En vista de disgregada que se encuentra la información UDLA posee un Data Warehouse que consolida datos de los distintos sistemas de la universidad, al que se puede acceder mediante consultas en lenguaje SQL.

Hasta el 2014 la mayor parte de la información se almacenaba de los estudiantes vigentes y se borraba la de los desertores. Por esto hay una gran cantidad de información de alumnos desertores que ya no se encuentra disponible.

3.1.2.2 Docente

En UDLA la gran mayoría de las asignaturas tiene tres evaluaciones principales a lo largo del semestre: cátedra 1, cátedra 2 y examen. Además de estas evaluaciones principales, hay asignaturas que tienen ejercicios realizados a lo largo del semestre, laboratorios u otro tipo de evaluaciones.

Por otro lado, también existe una plataforma online que entrega material de apoyo a los alumnos y en algunos casos se usa para realizar pequeñas evaluaciones en línea, e incluso existen cursos completos que se llevan a cabo remotamente. A esta plataforma se le llama aulas virtuales.

3.1.2.3 Aspectos financieros

Los alumnos de UDLA deben pagar una matrícula y un arancel anual para poder estudiar. Se permiten diversos medios de pago como cheques, pagarés con cuotas mensuales o pago al contado. En el sistema de UDLA queda registrado el RUT de la entidad que paga el arancel del alumno, con lo que se puede diferenciar si es un estudiante con una beca estatal, si paga él o un tercero, como por ejemplo una empresa.

Por otra parte, existe una política de cobranza en la universidad, importante a considerar para este modelo: si un alumno que paga con cuotas mensuales tiene más de cuatro cuotas morosas, no se le permite tomar carga académica al semestre siguiente.

3.1.2.4 Ingreso de estudiantes

Para ingresar a estudiar en UDLA el primer semestre del año, los alumnos se matriculan durante el verano, usualmente desde diciembre hasta abril. A lo largo de este proceso, los alumnos deben seguir una serie de pasos y llenar datos personales como comuna y región de origen, sexo, ingreso familiar, colegio de origen y notas de enseñanza media (NEM), entre otras. Uno de los problemas que tiene esta información es que hay una parte relevante que UDLA no puede corroborar. Por ejemplo, al ser una universidad no asociada al Departamento de Evaluación, Medición y Registro Educacional, no recibe los puntajes PSU ni un listado oficial del NEM de los alumnos que ingresan a la institución. Durante los últimos procesos de admisión se han solicitado más antecedentes de los alumnos, con lo que la información se ha vuelto más confiable, pero los datos históricos no lo son.

Otro aspecto que es necesario destacar es que existen diferentes fuentes de las que llegan alumnos a matricularse UDLA: los que llegan directamente a la universidad, los que contactan mediante el sitio web y los que son contactados telefónicamente por la universidad para ofrecerles sus servicios.

3.1.3 Análisis de información disponible

Una vez fijados los objetivos comenzó a explorarse el Data Warehouse de UDLA para determinar qué datos se encontraban disponibles y con ello el tipo de variables que podrían obtenerse. El servidor contaba con cientos de tablas, cada una con datos de diferente naturaleza, por lo que fue necesario entrevistar a expertos de las distintas áreas de la universidad para entender su estructura.

A partir de la información disponible y de los estudios citados en las secciones 2.1.2 y 2.10, se estableció que la información disponible de los alumnos puede separarse en las siguientes categorías:

- Personal: usualmente se obtiene cuando el alumno ingresa a la institución. Incluye datos como el colegio de origen, NEM, puntaje PSU (en caso de haberla rendido), comuna de residencia, región de origen, edad, sexo y otros similares.
- Matrícula: se obtiene cada vez que el alumno se matricula. Está relacionada a los datos de su programa de estudio, sede de estudio, carrera, si el alumno estudia en régimen diurno o vespertino, etc.
- Aulas virtuales: relacionada a la plataforma de apoyo virtual que posee UDLA y el uso que le da el alumno. Ejemplos de la información que se puede extraer de este tipo de tablas son: si el alumno tiene acceso a la plataforma; cuántas veces ingresó; cómo fue su rendimiento en las evaluaciones en línea, junto a la cantidad de intentos que realizó en cada una; y uso del chat, entre otras.
- Asignaturas: rendimiento del alumno en las diferentes evaluaciones, cantidad de evaluaciones por rendir, cursos aprobados, asistencia (aunque ir a clases sólo es obligatoria para alumnos nuevos), cantidad de asignaturas inscritas y eliminadas, avance de malla y muchas otras variables derivadas.
- Financieras: UDLA tiene tablas en su Data Warehouse con información de los pagos que los alumnos deben hacer y han hecho a la universidad, a partir de lo que se pudo obtener mucha información de relevancia. Con la información de la tabla se calcularon variables como la forma de pago principal del alumno, cuotas morosas, días de mora, cuotas pagadas, saldo moroso, saldo pagado y otras similares.
- Biblioteca: asociada al uso que hace el alumno de los recursos de biblioteca que la universidad tiene disponibles. Aquí hay información de préstamo de libros, de salas, computadores y otros.
- Solicitudes: en esta categoría están las consultas, comentarios favorables, solicitudes o reclamos que se hacen a UDLA por parte de los estudiantes.

Al revisar en detalle los datos históricos de UDLA surgieron dos problemas:

- Hay información cuya evolución en el tiempo no fue almacenada y sólo se tiene su último valor. Por ejemplo, en la base de datos cada alumno tiene asociada una asistencia a cada una de sus asignaturas, la que es actualizada cada semana sumando las sesiones asistidas de esa semana al total anterior, reemplazando su valor. Por tanto, sólo se podía acceder a la asistencia de cada alumno al final del semestre. Lo mismo ocurre con la información financiera, donde se actualiza el estado de cada cuota cuando cambia de estado, por ejemplo, al pasar de al día a morosa o pagada.
- La problemática más grave es que en UDLA se borra toda la información académica de los alumnos que se retiran, por lo que no existe registro de sus notas y otros datos asociados a sus asignaturas.

3.2 SELECCIÓN

En esta parte del proceso KDD se definen las etiquetas y el set de datos a utilizar. A continuación, se explican los aspectos considerados para esto.

3.2.1 Falta de datos históricos

A raíz de que UDLA no guarda información histórica, la única fuente de información completa, la única fuente de información completa corresponde a una tabla creada en 2014. Esta tabla almacena la información académica y de pago de los alumnos de manera semanal, pero sólo tiene datos del segundo semestre de 2014, el primero de 2015 y el primero de 2016. Por esto se decidió acotar el problema: solo se utilizaron los datos disponibles del primer semestre de 2015 y 2016 para en este trabajo, pero se siguió utilizando la definición de deserción descrita en la sección 3.1.1, usando como etiqueta todas las deserciones que se efectuaron por alumnos de primer año entre el inicio de clases y el comienzo del año académico siguiente.

3.2.2 Información al avanzar el semestre

En vista de que el objetivo del sistema es estimar un índice de riesgo de deserción para cada estudiante, de modo que se pueda actuar de manera temprana sobre los más propensos a dejar la institución, es necesario que el modelo entregue resultados lo antes posible: idealmente al principio del año y luego al avanzar el semestre actualice su clasificación a medida que más información se vaya haciendo disponible.

Como existían datos históricos semanales de los primeros semestres de los años 2015 y 2016 fue necesario determinar en qué fecha se realizaría la modelación. Al estudiar las variables que se podían obtener, se observó que la cantidad de información disponible crece a lo largo del semestre, por ejemplo, de evaluaciones, asistencia y préstamos de libros.

A su vez, es deseable predecir las deserciones lo antes posible, debido a que con el paso del tiempo hay alumnos que dejan la institución y por ende quedan menos por predecir, como también porque la detección temprana da más opciones a los directores de carrera de ayudar a sus alumnos. El hecho de que la cantidad de información aumente a medida que avanza el semestre pero que a su vez los alumnos vayan desertando genera una disyuntiva sobre la fecha a considerar para realizar el modelamiento: si la predicción se hace recién iniciado el semestre tendrá menos información, por lo que probablemente será menos precisa, pero es valioso efectuarla lo antes posible. Por el contrario, mientras más tarde en el semestre se haga la predicción se contará con más información, pero un porcentaje importante de deserciones ya se habrán efectuado y habrá menos tiempo para ayudar a los alumnos que aún no desertan.

Según los estudios citados en la sección 2.10, el rendimiento académico suele ser muy valioso para predecir deserción de estudiantes. En la Figura 3-1 y la Figura 3-2 se muestra la cantidad de calificaciones ingresadas en la base de datos a lo largo del semestre, junto con las solicitudes formales de deserción del alumno, puesto que son las deserciones que tienen una fecha precisa asociada. En ambos gráficos la línea amarilla representa la suma de las deserciones con solicitud formal (retiros) que se produjeron después de esa semana durante el semestre, cuyo eje está a la derecha, de color amarillo. La línea azul representa la cantidad de cátedras 1 del primer semestre cuya nota había sido ingresada hasta esa semana, mientras que las líneas naranja y gris representan lo mismo para las cátedras 2 y examen respectivamente. Ambos gráficos parten desde el inicio de clases del primer semestre del año respectivo y terminan cuando el segundo semestre llevaba varias semanas de clases.

En los gráficos mencionados en el párrafo anterior se ve que la información sobre el rendimiento académico crece abruptamente en ciertas fechas, cuando se realiza la corrección de las evaluaciones. Además, al final de cada gráfico quedan alrededor de 500 deserciones con solicitud formal por predecir. Cabe destacar que hay más deserciones producidas en el año pero que no están incluidas en el gráfico porque no se tiene una fecha precisa de cuándo se produjeron, solo se sabe que el alumno no volvió a clases el año siguiente.

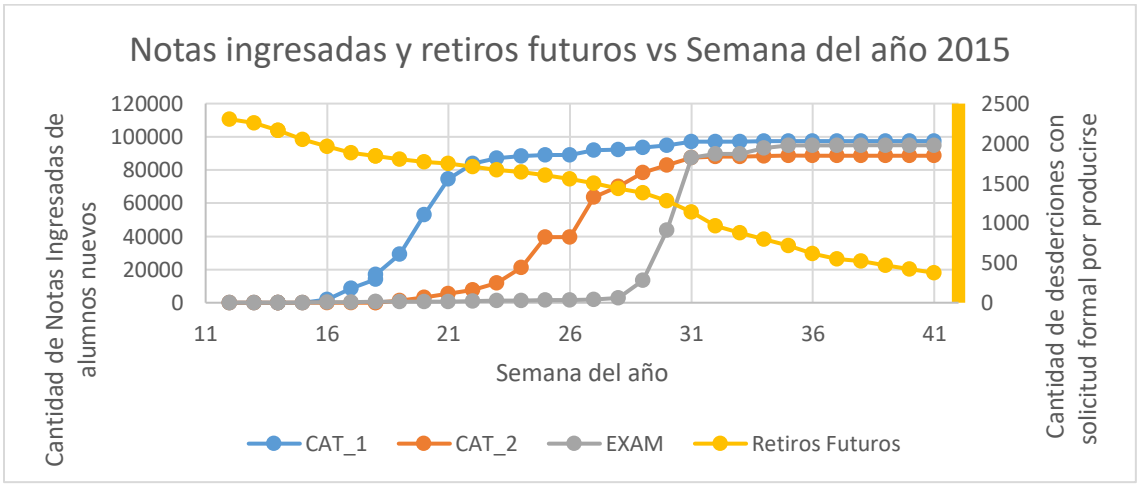

Figura 3-1 Cantidad de notas ingresadas durante el primer semestre de los alumnos nuevos y cantidad de deserciones con solicitud formal (retiros) de alumnos nuevos por producirse vs la semana del año 2015. En amarillo se muestra la cantidad de retiros futuros, en azul las notas de la cátedra 1, en naranjo de la cátedra 2 y en gris del examen.

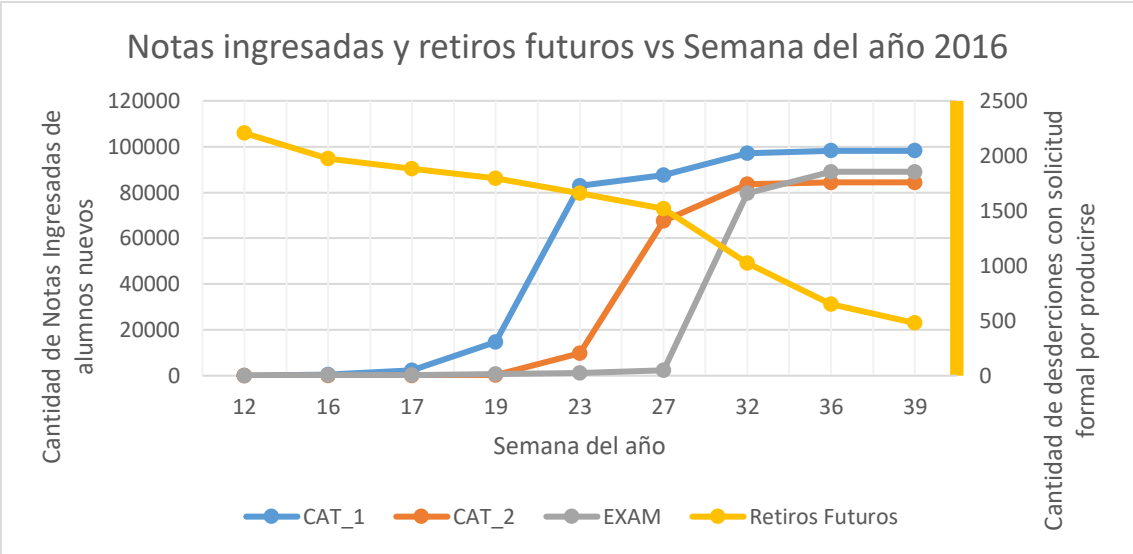

Figura 3-2

Cantidad de notas ingresadas durante el primer semestre de los alumnos nuevos y cantidad de deserciones con solicitud formal (retiros) de alumnos nuevos por producirse vs la semana del año 2015. En amarillo se muestra la cantidad de retiros futuros, en azul las notas de la cátedra 1, en naranjo de la cátedra 2 y en gris del examen.

3.2.3 Definición de la base de datos

A partir de lo descrito en la sección 3.2.1 se decidió lo siguiente para el modelo:

1. Dividir el semestre en tramos de tiempo de cierto número de semanas y entrenar un modelo distinto en cada uno, bajo la suposición de que las variables de importancia cambian al aparecer información nueva a lo largo del año.
2. Concentrarse sólo en el primer semestre del 2015 y 2016 debido a que son los únicos semestres con toda la información disponible. Además, si las pruebas son exitosas, extender el sistema para el segundo semestre cuando haya información suficiente debiese ser simple.
3. No diferenciar entre los distintos regímenes de estudio de UDLA para no complicar en mayor medida la estructura del sistema. Los regímenes se refieren a si el alumno estudia en
    a. Diurno: horario tradicional de clases.
    b. Vespertino: horario nocturno de clases
    c. Executive: horario destinado a profesionales que quieren complementar sus estudios por lo que suele tener más flexibilidad.

A partir de las 3 decisiones mencionadas en la lista anterior, se procedió a definir las fechas en que se realizarían los modelamientos. En el gráfico de la Figura 3-1 se aprecia que en la semana 22 del año 2015 aproximadamente el 90% de las calificaciones de la cátedra 1 ya se encontraban ingresadas en la base de datos. En

el caso de la cátedra 2, el ingreso de notas fue muy cercano a las del examen, por lo que, en lugar de buscar un 90%, se optó por escoger la semana 28, donde la cantidad de ingresos era cercana al 80%. Finalmente, para los exámenes se escogió la semana 31, donde más del 90% ya estaban evaluados, puesto que más adelante comenzó el segundo semestre, que tiene sus propias variables.

Asimismo, en la Figura 3-2 se aprecia que condiciones similares de la cantidad de ingreso de notas se dieron en la semana 23 para la cátedra 1, semana 27 para la cátedra 2 y semana 32 para el examen, todas del año 2016.

Se definieron cuatro bases de datos diferentes, para así entrenar modelos de predicción que se ajusten a los datos al avanzar el semestre:

1. Set de datos del inicio del semestre: datos de la semana 12 de los años 2015 y 2016. Entrega un índice de riesgo temprano, pero con menos información de los alumnos. Incluye información personal y de matrícula.
2. Set de datos de la cátedra 1: datos de la semana 22 del año 2015 y 23 del año 2016. Además de la información del set de datos anterior, se agregó información del uso de biblioteca, las notas de ejercicios y cátedra 1, asistencia y comportamiento de pago.
3. Set de datos de la cátedra 2: datos de la semana 28 del 2015 y 27 del 2016. A las variables anteriores se le agregaron indicadores con información del rendimiento en la cátedra 2.
4. Set de datos del fin del semestre: datos de la semana 31 de 2015 y 32 del 2016, fecha que corresponde al final del semestre. Se incluyen las notas de los exámenes y el cierre de las asignaturas.

En la Tabla 3-1 se muestra un resumen de las fechas correspondientes al día viernes de la semana escogida para la toma de muestras de cada base de datos.

Tabla 3-1 Fechas escogidas para construir los conjuntos de datos correspondientes a los cuatro momentos del semestre donde se realiza la modelación

Dataset	**Fecha 2015**	**Fecha 2016**
Inicio del semestre	20 de marzo	13 de marzo
Cátedra 1	20 de mayo	1 de junio
Cátedra 2	5 de julio	1 de julio
Fin de semestre	7 de agosto	12 de agosto

3.2.4 Definición de las etiquetas

Para el modelamiento se realizó la siguiente definición de etiquetas:

- Desertor: cualquier alumno de primer año que haya dejado la institución entre el inicio de clases y abril del año siguiente, o que no haya vuelto a matricularse.

- Retención: alumno de primer año que no deja la institución entre el inicio de clases y el año siguiente y se vuelve a matricular.

Los alumnos que desertaron antes de la fecha definida para cada conjunto de datos no fueron considerados. Esto se debe a que no tiene sentido predecir si un alumno desertará de la institución si ya lo hizo. Se definió una deserción como 1 y una retención como 0.

En la Tabla 3-2 se muestran ejemplos de seis estudiantes ficticios, donde los ✓ simbolizan que el alumno está incluido en la base de datos para ese modelo y las × que no lo está. Aquí se puede apreciar que todos los retiros posteriores al 20 de marzo aparecen en el modelo de origen, pero alumnos como el 2 y el 3 no se consideran en ninguno de los otros. Los alumnos 4 y 5 también se utilizan en el modelo de la cátedra 1, pero no en los demás y así sucesivamente.

Tabla 3-2 Ejemplos ficticios de las deserciones incluidas en cada set de datos extraído para modelar del año 2015

Alumno	**Fecha retiro**	**Ini sem** **20 de marzo**	**Cat 1** 20 de mayo	**Cat 2** 5 de julio	**Fin sem** 7 de agosto
1	03 de marzo	×	×	×	×
2	21 de marzo	✓	×	×	×
3	15 de abril	✓	×	×	×
4	28 de mayo	✓	✓	×	×
5	03 de junio	✓	✓	×	×
6	07 de julio	✓	✓	✓	×
7	9 de agosto	✓	✓	✓	✓
8	9 de septiembre	✓	✓	✓	✓

En la Tabla 3-3 se muestra un resumen de la cantidad de deserciones incluidas en el conjunto de datos. Al inicio del semestre y en la cátedra 1 hay un 31,8% de deserciones por producirse, lo que se reduce a un 30,7% en la cátedra 2 y a un 30,3% al final del semestre.

Otro punto a considerar es que la cantidad de alumnos aumenta entre el inicio del semestre y la cátedra 1 en 474 alumnos. Esta diferencia se debe a dos razones: hubo algunas matrículas entre el inicio de clases y la cátedra 1, y además no se consideraron para el conjunto de datos algunos alumnos con mucha información faltante o errónea. Por esto la cantidad de desertores y retenciones aumenta entre estas dos fechas en 149 y 325 respectivamente.

Entre la cátedra 1 y 2 se aprecia que se produjeron 167 deserciones y no se consideraron 5 alumnos retenidos por falta de información, mientras que entre la

cátedra 2 y el final del semestre se produjeron solo 70 deserciones y se no se consideraron 6 alumnos retenidos por falta de información.

Tabla 3-3 Tabla resumen de los datos los disponibles considerando los primeros semestres de los años 2015 y 2016

	Cantidad de alumnos	Desertores	Retenciones	% deserciones
Inicio del semestre	10702	3401	7301	31,8%
Cátedra 1	11176	3550	7626	31,8%
Cátedra 2	11004	3383	7621	30,7%
Fin del semestre	10928	3313	7615	30,3%

3.3 PREPROCESAMIENTO DE DATOS

Con las bases de datos y sus etiquetas definidas comenzó el preprocesamiento de los datos, donde la labor principal fue el limpiado de datos erróneos y elección del método de reemplazo para los distintos datos faltantes.

3.3.1 Reparación de datos erróneos

A medida que se fue fueron programando las variables y comprendiendo el funcionamiento de UDLA se fueron evidenciando errores en los datos. Desde personas con su género mal asignado hasta notas que no tenían sentido: mayores a 7,0 o que en lugar de un número contenían letras aleatorias. En algunos casos la reparación de estos valores se realizó mediante un simple reemplazo por el valor correcto, pero en otras ocasiones hubo que realizar consultas a las áreas encargadas de las tablas correspondientes y solicitar que se verificara la información. En general la mayoría de los datos con valores evidentemente erróneos se pudieron reparar.

3.3.2 Tratamiento de datos faltantes

Como ya fue explicado en la sección 2.7, existen tres tipos de datos faltantes, por lo que hubo que estudiar cada variable con cuidado para determinar la razón por la que tenía datos faltantes. En algunos casos fue claro que correspondía a un NMAR, como en notas: el alumno aún no rinde esa evaluación; o en cuotas morosas: el alumno ha pagado todas sus cuotas al día. En otros casos probablemente se tratase de un MAR, en particular para datos de información personal, por ejemplo, alumnos que no informan su ingreso familiar, posiblemente porque en caso de ser mayor a cierto valor no podrían obtener algunos beneficios.

Se encontraron casos de variables con datos faltantes tipo MCAR, como en la plataforma de aulas virtuales: algunos alumnos tenían varios inicios de sesión en la

plataforma sin sus tiempos de permanencia, los que fueron imputados reemplazando por el promedio de la duración del total de los alumnos nuevos.

Los datos faltantes NMAR en general se pudieron deducir e imputar con el valor adecuado, mientras que los MCAR y MAR se rellenaron utilizando códigos numéricos que estén fuera del rango de valores que la variable debiese tomar. Por ejemplo -99 para ingreso familiar, esto debido a que algunas técnicas de modelamiento no permiten valores vacíos. Estos valores se introdujeron para ayudar al modelo a diferenciar que estos datos no estaban presentes porque la implementación utilizada no permite datos vacíos.

3.4 TRANSFORMACIÓN

La base de datos de UDLA contiene muchos datos repartidos en cientos de tablas, cada una con su estructura particular, junto a muchos datos asociados a la operación de UDLA, pero que eran irrelevantes para este trabajo. Por esto se necesitaba seleccionar, combinar y ordenar la información relevante en un conjunto de datos apropiado para trabajar con los algoritmos seleccionados para modelar y con una estructura como la mostrada en la Tabla 3-4.

Tabla 3-4 Diagrama descriptivo de la estructura de datos a la que se llegó para entrenar los modelos

ID Alumnos	Variables	Etiquetas

En base a entrevistas con empleados de diversas áreas de UDLA de como áreas cobranza, gestión de alumnos desertores, profesores y encargados de captación de alumnos nuevos, junto a la experiencia de otros trabajos listados en la sección 2.10, se escogieron las variables que podrían tener algún grado de utilidad para resolver el problema de clasificación. A continuación se describe el proceso transformación para todas las variables basándose en las categorías descritas en la sección 3.1.3. Cabe recordar que algunas sólo tienen sentido para ciertos modelos, por ejemplo, las notas de la cátedra 2 no se utilizan para el modelo de la cátedra 1.

3.4.1.1 Variables personales

Estas variables no requirieron mayores cálculos u operaciones. Lo único que se realizó fue compilar datos de diferentes tablas esparcidas en la base de datos. Ejemplos de estas variables son comuna de origen, promedio de notas de la enseñanza media, ingreso familiar y resultados en la PSU. La información disponible en las bases de datos de UDLA fue complementada con dos fuentes de información:

- Encuesta CASEN del año 2013: Se utilizaron los datos disponibles en el documento *Indicadores comunales Región Metropolitana en base a CASEN 2013* para obtener el ingreso promedio por hogar de las distintas comunas

de Santiago. Esta información se incluyó en el conjunto de datos, asociando su comuna de residencia al ingreso promedio de las viviendas de ella, a modo de complementar lo que el alumno informó a UDLA como sus ingresos familiares [7].

- API de Google Maps: Se incluyeron como variables una aproximación del tiempo y distancia que debe recorrer un alumno para llegar desde su comuna de residencia a la de estudio puesto que se utilizaron en otros trabajos similares citados en la sección 2.10 [42].

En la Tabla 3-5 se presenta el detalle de todas las variables mencionadas en esta sección.

Tabla 3-5 Detalle de las variables personales utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Es Chileno	No	Si	Si	Si	Si	No
Estado Civil	Si	Si	Si	Si	Si	Si, con soltero
Edad	No	Si	Si	Si	Si	No
Notas enseñanza media	No	Si	Si	Si	Si	No
Puntaje PSU lenguaje	No	Si	Si	Si	Si	Si, con 0
Puntaje PSU matemáticas	No	Si	Si	Si	Si	Si, con 0
Promedio en PSU	No	Si	Si	Si	Si	Si, con 0
Indicador si la región de origen es igual a la de estudio	No	Si	Si	Si	Si	No
Comuna de residencia	Si	Si	Si	Si	Si	No
Comuna de trabajo	Si	Si	Si	Si	Si	No
Tramo de renta familiar	No	Si	Si	Si	Si	Si con -99
Máximo del tramo de renta familiar	No	Si	Si	Si	Si	Si con -99
Ingresos promedio por hogar de la comuna de residencia	No	Si	Si	Si	Si	No

Tiempo en segundos de viaje de comuna de residencia a la de estudio	No	Si	Si	Si	Si	No
Distancia de en metros de viaje de comuna de residencia a la de estudio	No	Si	Si	Si	Si	No
Costo de matrícula del colegio	No	Si	Si	Si	Si	Si, con el promedio
Mensualidad del colegio	No	Si	Si	Si	Si	Si, con el promedio
Comuna del colegio	No	Si	Si	Si	Si	Si, con el promedio

3.4.1.2 Variables de matrícula

La mayoría de estos datos estaban almacenados de manera ordenada por lo que se pudieron incluir con facilidad en la tabla final con la estructura mencionada. Ejemplos de estas variables son régimen de estudio, carrera, programa, fecha de matrícula y cantidad de asignaturas a cursar durante el semestre. En la Tabla 3-6 se muestra el detalle de estas variables.

Tabla 3-6 Detalle de las variables de matrícula utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Programa de estudio	Si	Si	Si	Si	Si	No
Sede de estudio	Si	Si	Si	Si	Si	No
Régimen de estudio	Si	Si	Si	Si	Si	No
Facultad de estudio	Si	Si	Si	Si	Si	No
Avance de malla	No	No	No	No	Si	No

Estado académico	No	Si	Si	Si	Si	No
Diferencia de días entre la fecha en que se matriculó y el inicio de clases	No	Si	Si	Si	Si	No

3.4.1.3 Evaluaciones

Como existe una cantidad muy elevada combinaciones de asignaturas que un alumno puede tener y no todos los alumnos tienen las mismas clases. Para simplificar el análisis de las evaluaciones se optó por generar variables de notas agrupadas por la cátedra 1, cátedra 2, exámenes, ejercicios, otras evaluaciones y notas finales. Ejemplos de estos indicadores son el promedio de notas del alumno en la cátedra 1, su desviación estándar, a cuantas no se presentó, en cuantas evaluaciones tuvo un rendimiento mayor a la media de sus compañeros, cuantas reprobó, cuantas aprobó, cuantas tuvieron media sobre 4, cuantas tuvieron un porcentaje de aprobación mayor al 50%, entre otras.

Estas variables se escogieron porque entregan información sobre el rendimiento individual del alumno, su compromiso con las evaluaciones, su rendimiento en comparación a sus compañeros y una estimación de la dificultad que tenían las pruebas. El detalle de estas variables se ubica en la Tabla 3-7.

Tabla 3-7 Detalle de las variables de evaluaciones utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Cantidad de asignaturas inscritas	No	Si	Si	Si	Si	No
Cantidad de notas de ejercicios ingresadas	No	No	Si	Si	Si	No
Cantidad de notas de ejercicios no ingresadas	No	No	Si	Si	Si	No
Cantidad de notas de ejercicios pendientes	No	No	Si	Si	Si	No
Cantidad de ejercicios con inasistencia	No	No	Si	Si	Si	Si, con 0
Razón de ejercicios con inasistencia	No	No	Si	Si	Si	No

Promedio de notas en ejercicios	No	No	Si	Si	Si	No
Desviación estándar en notas de ejercicios	No	No	Si	Si	Si	Si, con 0
% de aprobación en ejercicios	No	No	Si	Si	Si	No
Cantidad de notas ingresadas de la cátedra 1	No	No	Si	Si	Si	No
Cantidad de notas no ingresadas de la cátedra 1	No	No	Si	Si	Si	No
Cantidad de cátedra 1 con inasistencia	No	No	Si	Si	Si	Si, con 0
Razón de inasistencias en la cátedra 1	No	No	Si	Si	Si	No
Cantidad de cátedras 1	No	No	Si	Si	Si	No
Promedio de notas en cátedra	No	No	Si	Si	Si	No
Desviación estándar en notas de la cátedra 1	No	No	Si	Si	Si	No
Porcentaje de aprobación de la cátedra 1	No	No	Si	Si	Si	No
Cantidad de Cátedra 1 sobre la media	No	No	Si	Si	Si	No
Cantidad de cátedra 1 bajo la media	No	No	Si	Si	Si	No
Promedio de cantidad de cátedras 1 de sus compañeros	No	No	Si	Si	Si	No
Promedio en la cátedra 1 de su curso	No	No	Si	Si	Si	No
Cantidad de cátedras 1 con media del curso sobre 4	No	No	Si	Si	Si	No
Cantidad de cátedras 1 con media del curso bajo 4	No	No	Si	Si	Si	No

Cantidad de cátedra 1 con aprobación mayor al 50%	No	No	Si	Si	Si	No
Cantidad de cátedra 1 con aprobación menor al 50%	No	No	Si	Si	Si	No
Cantidad de notas ingresadas de la cátedra 2	No	No	No	Si	Si	No
Cantidad de notas no ingresadas de la cátedra 2	No	No	No	Si	Si	No
Cantidad de cátedra 2 con inasistencia	No	No	No	Si	Si	Si, con 0
Razón de inasistencias en la cátedra 2	No	No	No	Si	Si	No
Cantidad de cátedras 2	No	No	No	Si	Si	No
Promedio de notas en cátedra	No	No	No	Si	Si	No
Desviación estándar en notas de la cátedra 2	No	No	No	Si	Si	No
Porcentaje de aprobación de la cátedra 2	No	No	No	Si	Si	No
Cantidad de Cátedra 2 sobre la media	No	No	No	Si	Si	No
Cantidad de cátedra 2 bajo la media	No	No	No	Si	Si	No
Promedio de cantidad de cátedras 2 de sus compañeros	No	No	No	Si	Si	No
Promedio en la cátedra 2 de su curso	No	No	No	Si	Si	No
Cantidad de cátedras 2 con media del curso sobre 4	No	No	No	Si	Si	No
Cantidad de cátedras 2 con media del curso bajo 4	No	No	No	Si	Si	No
Cantidad de cátedra 2 con aprobación mayor al 50%	No	No	No	Si	Si	No

Cantidad de cátedra 2 con aprobación menor al 50%	No	No	No	Si	Si	No
Cantidad de notas ingresadas del examen	No	No	No	No	Si	No
Cantidad de notas no ingresadas del examen	No	No	No	No	Si	No
Cantidad de examen con inasistencia	No	No	No	No	Si	Si, con 0
Razón de inasistencias en el examen	No	No	No	No	Si	No
Cantidad de cátedras 2	No	No	No	No	Si	No
Promedio de notas en cátedra	No	No	No	No	Si	No
Desviación estándar en notas de examen	No	No	No	No	Si	No
Porcentaje de aprobación de examen	No	No	No	No	Si	No
Cantidad de examen sobre media	No	No	No	No	Si	No
Cantidad de examen bajo la media	No	No	No	No	Si	No
Promedio de cantidad de exámenes de sus compañeros	No	No	No	No	Si	No
Promedio en examen de su curso	No	No	No	No	Si	No
Cantidad de exámenes con media del curso sobre 4	No	No	No	No	Si	No
Cantidad de exámenes con media del curso bajo 4	No	No	No	No	Si	No
Cantidad de examen con aprobación mayor al 50%	No	No	No	No	Si	No
Cantidad de examen con aprobación menor al 50%	No	No	No	No	Si	No

Cantidad de notas ingresadas de otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Cantidad de notas no ingresadas de otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Cantidad de inasistencias a otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Razón de inasistencias a otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Cantidad de otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Promedio de notas en otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Porcentaje de aprobación en otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Desviación estándar en las notas de otras evaluaciones	No	No	Si	Si	Si	Si, con 0
Promedio final en sus asignaturas	No	No	No	No	Si	No
Desviación estándar de notas finales en sus asignaturas	No	No	No	No	Si	No
Asignaturas aprobadas	No	No	No	No	Si	No
Razon de asignaturas aprobadas	No	No	No	No	Si	No
Asignaturas Reprobadas	No	No	No	No	Si	No
Cantidad de asignaturas con nota final sobre la media	No	No	No	No	Si	No
Razón de asignaturas con nota final sobre la media	No	No	No	No	Si	No
Cantidad de asignaturas con nota final bajo la media	No	No	No	No	Si	No
Total de asignaturas	No	No	No	No	Si	No

3.4.1.4 Uso de biblioteca

Las tablas asociadas al uso de biblioteca tienen una estructura de solicitudes y devoluciones de recursos, donde cada recurso pertenece a un tipo. Con esta información se puede realizar medir el uso que los alumnos le dan a la biblioteca mediante algunos indicadores. En la Tabla 3-8 se muestra un ejemplo simplificado de cómo se guarda la información de los préstamos de recursos de la biblioteca.

Tabla 3-8 Ejemplo básico de la forma en que se guarda la información del uso de los recursos de la biblioteca de UDLA

Recurso	Tipo de recurso	Prestado a	Fecha de prestamo	Fecha de devolución esperada	Fecha de devolución real
El Aleph	Libro	Matias Galleguillos	5/3/2015	16/3/2015	18/3/2015
Sala 1	Sala	Matias Galleguillos	8/4/2015	8/4/2015	8/4/2015

Lo que se realizó fue agrupar la cantidad de usos por parte de cada alumno en el mes y los usos acumulados en el semestre en una de tres categorías:

- Préstamo de libros
- Préstamo de salas
- Préstamo de otros recursos

En otros recursos se incluyen DVD, VHS u otros elementos electrónicos que se encuentran en biblioteca. En la Tabla 3-9 está el detalle de estas variables, si se utilizó one-hot encoding, en qué modelos fueron utilizadas y si hubo que realizar imputación de datos faltantes.

Tabla 3-9 Detalle de las variables de uso de biblioteca utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Numero de libros prestados durante el mes	No	No	Si	Si	Si	No
Cantidad de salas de estudio solicitadas durante el mes	No	No	Si	Si	Si	No
Prestamo de otros recursos durante el mes	No	No	Si	Si	Si	No
Préstamo de libros durante la última semana	No	No	Si	Si	Si	No
Solicitud de salas de estudio durante la última semana	No	No	Si	Si	Si	No
Préstamo de otros recursos durante la última semana	No	No	Si	Si	Si	No
Préstamo de libros acumulado	No	No	Si	Si	Si	No
Préstamo de otros recursos acumulado	No	No	Si	Si	Si	No
Solicitud de salas de studio acumulado	No	No	Si	Si	Si	No
Préstamo de libros promedio por semana	No	No	Si	Si	Si	No
Préstamo de otros recursos promedio por semana	No	No	Si	Si	Si	No
Solicitudes de salas de estudio promedio por semana	No	No	Si	Si	Si	No

3.4.1.5 Uso de aulas virtuales

Hay una gran cantidad de tablas de la base de datos asociadas a la plataforma de aulas virtuales. Están relacionadas mediante una estructura compleja, por lo que fue necesario contar con ayuda de un experto en éstas. Para llevar esto a la estructura deseada de variables explicada en la Tabla 3-4 fue necesario transformar los datos en variables. Las variables reflejan si el alumno tiene acceso a aulas virtuales, las asignaturas que tiene con material en aulas virtuales, el uso que le da a la plataforma y su rendimiento en las evaluaciones que se realizan en la plataforma. En la Tabla 3-10 se encuentra el detalle de las variables, si se utilizó one-hot encoding, en qué modelos fueron utilizadas y si hubo que realizar imputación de datos faltantes.

Tabla 3-10 Detalle de las variables del uso de aulas virtuales utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Indicador si el usuario tiene acceso a aulas virtuales	No	No	Si	Si	Si	No
Cantidad de asignaturas con recursos en aulas virtuales	No	No	Si	Si	Si	No
Cantidad de veces que ha hecho login	No	No	Si	Si	Si	No
Cantidad de segundos de permanencia en aulas virtuales	No	No	Si	Si	Si	Si con el promedio
Promedio de segundos de permanencia por login	No	No	Si	Si	Si	No
Cantidad de discusiones creadas en los los foros	No	No	Si	Si	Si	No
Cantidad de respuestas escritas en los foros	No	No	Si	Si	Si	No
Cantidad de ingresos al chat	No	No	Si	Si	Si	No

Cantidad de mensajes enviados por el chat	No	No	Si	Si	Si	No
Promedio en evaluaciones de aulas virtuales	No	No	Si	Si	Si	No
Cantidad de evaluaciones en aulas virtuales	No	No	Si	Si	Si	No
Total de intentos en evaluaciones de aulas virtuales	No	No	Si	Si	Si	No
Promedio de intentos por evaluación de aulas virtuales	No	No	Si	Si	Si	No

3.4.1.6 Información de los pares

Las variables con información de los pares de los alumnos también provienen de varias tablas separadas. El objetivo de estos atributos es representar el entorno en que se encuentra el estudiante, en particular los retiros de sus compañeros de carrera y de programa, bajo la suposición de que podrían afectar cómo el estudiante percibe la carrera que estudia. Ejemplos de estas variables incluyen promedio de asistencia de los pares, deserciones de años anteriores de la carrera del alumno y promedio de notas de los pares. El detalle de las variables se muestra en la Tabla 3-11.

Tabla 3-11 Detalle de las variables de los pares del alumno utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Promedio de compañeros por curso	No	Si	Si	Si	Si	No
Promedio final de notas de su curso	No	No	Si	Si	Si	No
Cantidad de asignaturas aprobadas por su curso	No	No	Si	Si	Si	No
Razón de asignaturas con media sobre 4	No	No	Si	Si	Si	No

Cantidad de asignaturas reprobadas por su curso	No	No	Si	Si	Si	No
Cantidad de asignaturas con aprobación mayor al 50%	No	No	Si	Si	Si	No
Razón de asignaturas con aprobación mayor al 50% del alumno	No	No	Si	Si	Si	No
Cantidad de asignaturas con aprobación menor al 50%	No	No	Si	Si	Si	No
Promedio de compañeros por curso	No	Si	Si	Si	Si	No
Cantidad de alumnos en su programa de estudio	No	Si	Si	Si	Si	No
Cantidad de retiros en el semestre anterior en el programa de estudio	No	Si	Si	Si	Si	No
Porcentaje de retiros en el semestre anterior en el programa de estudios	No	Si	Si	Si	Si	No
Cantidad de retiros en el programa a la misma fecha del año anterior	No	Si	Si	Si	Si	No
% de retiros del año anterior en el programa a la fecha	No	Si	Si	Si	Si	No
Cantidad de alumnos en su carrera	No	Si	Si	Si	Si	No
Cantidad de retiros del semestre anterior en la carrera	No	Si	Si	Si	Si	No

Porcentaje de retiros en el semestre anterior en la carrera	No	Si	Si	Si	Si	No
Cantidad de retiros en la carrera a la misma fecha del año anterior	No	Si	Si	Si	Si	No
% de retiros del año anterior en la carrera a la fecha	No	Si	Si	Si	Si	No

3.4.1.7 Información financiera

La información financiera de los alumnos se obtuvo a partir de una tabla que contenía los pagos hechos y por realizar de los alumnos. El objetivo las variables de tipo financieras es representar tanto el comportamiento de pago del alumno como su capacidad pago, puesto que como ya fue mencionado, muchos estudiantes dejan de estudiar por falta de financiamiento y por la política que existe en la universidad de no dejar inscribir asignaturas a los alumnos con 4 o más cuotas morosas. El detalle de estas variables se encuentra en la Tabla 3-12.

Tabla 3-12 Detalle de las variables financieras del alumno utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Tiene CAE	No	Si	Si	Si	Si	No
Paga empresa	No	Si	Si	Si	Si	No
Tiene beca estatal	No	Si	Si	Si	Si	No
Es alumno enlace	No	Si	Si	Si	Si	No
Tipo de plan de pago	Si	Si	Si	Si	Si	No
Cantidad de cuotas morosas	No	No	Si	Si	Si	No
Cantidad de cuotas vencidas	No	No	Si	Si	Si	No
Razón de cuotas morosas	No	No	Si	Si	Si	No
Días de mora	No	No	Si	Si	Si	No
Máximo de días mora posibles	No	No	Si	Si	Si	No

Razón de días de mora respecto del máximo	No	No	Si	Si	Si	No
Saldo moroso	No	No	Si	Si	Si	No
Número de cuotas por pagar en el año	No	Si	Si	Si	Si	No
Saldo total por pagar en el año	No	Si	Si	Si	Si	No
Cantidad de cuotas pagadas	No	No	Si	Si	Si	No
Pago total realizado	No	Si	Si	Si	Si	No
Cantidad de cuotas agrupadas por mes	No	Si	Si	Si	Si	No
Cantidad de cuotas morosas agrupadas por mes	No	No	Si	Si	Si	No
Cantidad de cuotas agrupadas vencidas	No	No	Si	Si	Si	No
Razón de cuotas morosas agrupadas por mes	No	No	Si	Si	Si	No
Arancel neto	No	Si	Si	Si	Si	No
Matícula neta	No	Si	Si	Si	Si	No
Porcentaje de descuento en el arancel	No	Si	Si	Si	Si	No
Porcentaje de descuento en la matrícula	No	Si	Si	Si	Si	No
Monto en becas	No	Si	Si	Si	Si	No
Monto en becas tipo beca	No	Si	Si	Si	Si	No
Monto en beca tipo CAE	No	Si	Si	Si	Si	No
Monto en beca tipo convenio	No	Si	Si	Si	Si	No
Monto en beca tipo deportiva	No	Si	Si	Si	Si	No
Monto en beca tipo destreza artística	No	Si	Si	Si	Si	No

Monto de beca tipo mérito académico	No	Si	Si	Si	Si	No
Monto de beca tipo otros	No	Si	Si	Si	Si	No
Monto de beca sin clasificacion	No	Si	Si	Si	Si	No

3.4.1.8 Asistencia

Estas variables reflejan las sesiones de clase a las que los alumnos asistieron. Para construir las variables se tomó el total de sesiones asistidas y no asistidas del alumno en el momento de toma de la muestra, sumando las de todas sus asignaturas en un solo valor. El detalle de estas variables se muestra en la Tabla 3-13.

Tabla 3-13 Detalle de las variables de asistencia del alumno utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Total de sesiones asistidas	No	No	Si	Si	Si	No
Total de sesiones programadas	No	No	Si	Si	Si	No
Total de sesiones registradas	No	No	Si	Si	Si	No
Razón de sesiones asistidas	No	No	Si	Si	Si	No
Asignaturas con asistencia mayor a 50%	No	No	Si	Si	Si	No

3.4.1.9 Relación con la universidad

Si bien la relación de un alumno con la universidad es abstracto y difícil de medir, se encontraron algunos datos que permiten reflejar en alguna medida. Para esto se utilizó el sistema de reclamos y solicitudes que pueden utilizar los alumnos para contactarse con el establecimiento, además de una variable que indica si el alumno ha tenido algún rol docente en la universidad. El detalle de estas variables está en la Tabla 3-14.

Tabla 3-14 Detalle de las variables de la relación del alumno con la universidad utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Cantidad de reclamos	No	No	Si	Si	Si	No
Cantidad de solicitudes formales	No	No	Si	Si	Si	No
Cantidad de consultas a UDLA	No	No	Si	Si	Si	No

3.4.1.10 Cambios de comportamiento

Las variables de asistencia, financieras y de uso de la biblioteca tienen registros semanales en una tabla del Data Warehouse de UDLA desde el año 2014, lo que permitió agregar cálculos de cambios de comportamiento de los alumnos para detectar cambios su comportamiento, por ejemplo, un alumno que paga su arancel todos los meses podría dejar de pagar comenzar a acumular cuotas morosas en cierta fecha. Además, se construyeron variables derivadas para medir la variación en el entre una cátedra y otra o en comparación con sus notas en el examen. Como medida de variación se utilizó el valor Z, que es el valor en la fecha que se toma la muestra menos la media del valor de las semanas anteriores dividido en la desviación estándar de las semanas anteriores y entrega una medida de cuántas desviaciones estándar el valor se aleja de la media. El detalle de estas variables está en la Tabla 3-15.

Tabla 3-15 Detalle de las variables de variación de comportamiento del alumno utilizadas en este trabajo

Variable	One-hot encoding	Inicio de semestre	Cátedra 1	Cátedra 2	Fin de semestre	Imputación de datos Faltantes
Diferencia entre el promedio de la cátedra 2 y cátedra 1	No	No	No	Si	Si	No
Diferencia en la cantidad de inasistencias entre la cátedra 2 y la cátedra 1	No	No	No	Si	Si	No
Diferencia en la razón de inasistencias entre la cátedra 2 y la cátedra 1	No	No	No	Si	Si	No

Diferencia entre cantidad de cátedra 2 sobre la media y cátedra 1 sobre la emdia	No	No	No	Si	Si	No
Diferencia en el porcentaje de aprobación de la cátedra 2 y cátedra 1	No	No	No	Si	Si	No
Diferencia en el promedio del examen y la cátedra 1	No	No	No	No	Si	No
Diferencia en cantidad de inasistencias entre examen y cátedra 1	No	No	No	No	Si	No
Diferencia en razón de inasistencias entre examen y cátedra 1	No	No	No	No	Si	No
Diferencia de cantidad de notas sobre la media entre cátedra 1 y examen	No	No	No	No	Si	No
Diferencia en porcentaje de aprobación entre el examen y la cátedra 1	No	No	No	No	Si	No
Diferencia en porcentaje de aprobación entre el examen y la cátedra 2	No	No	No	No	Si	No
Diferencia en cantidad de inasistencias entre examen y cátedra 2	No	No	No	No	Si	No
Diferencia en razón de inasistencias entre examen y cátedra 2	No	No	No	No	Si	No
Diferencia de cantidad de notas sobre la media entre cátedra 2 y examen	No	No	No	No	Si	No
Diferencia en porcentaje de aprobación entre el examen y la cátedra 2	No	No	No	No	Si	No

Diferencia en el promedio de notas entre el examen y el promedio de las cátedras 1 y 2	No	No	No	No	Si	No
Diferencia de inasistencias entre el examen y el promedio de las cátedras 1 y 2	No	No	No	No	Si	No
Diferencia entre la razón de inasistencias entre el examen y el promedio de inasistencias entre la cátedra 1 y 2	No	No	No	No	Si	No
Diferencia entre cantidad de notas sobre la media entre el examen y el promedio de la cátedra 1 y 2	No	No	No	No	Si	No
Diferencia entre entre el porcentaje de aprobación entre el examen y el promedio de la cátedra 1 y 2	No	No	No	No	Si	No
Promedio de la razón de días de mora	No	No	Si	Si	Si	No
Desviación estándar de los días de mora	No	No	Si	Si	Si	No
Promedio de la razón de cuotas morosas en las semanas anteriores	No	No	Si	Si	Si	No
Desviación estándar de la razón de cuotas morosas en las semanas anteriores	No	No	Si	Si	Si	No
Promedio de saldo moroso en las semanas anteriores	No	No	Si	Si	Si	No
Desviación estándar de las semanas anteriores	No	No	Si	Si	Si	No
Promedio de la razón de sesiones asistidas	No	No	Si	Si	Si	No

Desviación estándar de la razón de sesiones asistidas	No	No	Si	Si	Si	No
Variación en la razón de días de mora	No	No	Si	Si	Si	No
Variación en el saldo moroso	No	No	Si	Si	Si	No
Variación en la razón de sesiones asistidas	No	No	Si	Si	Si	No
Diferencia de prestamo de libros entre este mes y el promedio mensual del alumno	No	No	Si	Si	Si	No
Diferencia de solicitud de salas de estudio entre este mes y el promedio mensual del alumno	No	No	Si	Si	Si	No
Diferencia de prestamo de otros recursos entre este mes y el promedio mensual del alumno	No	No	Si	Si	Si	No

3.5 DATA MINING

En esta sección del informe se detalla cuántas variables se obtuvieron para cada modelo y de qué tipo eran, luego se trata la separación de conjuntos de los datos y finalmente se detalla la selección de variables.

3.5.1 Variables

Como ya fue mostrado en la sección 3.2.2, al avanzar en el semestre, la cantidad de información asociada a cada alumno va creciendo, lo que se traduce en que se puede obtener una mayor cantidad de variables. Esto se presenta en el gráfico de barras de la Figura 3-3, donde se aprecia que para el modelo del inicio del semestre se consideraron 78 variables, para el de la primera cátedra 174, para el de la segunda 195 y para el del final del semestre 235 variables.

Es de especial interés destacar que hay un salto muy pronunciado entre el primer modelo y el segundo, llegándose a tener más del doble de variables, mientras que en el resto los aumentos no son tan grandes.

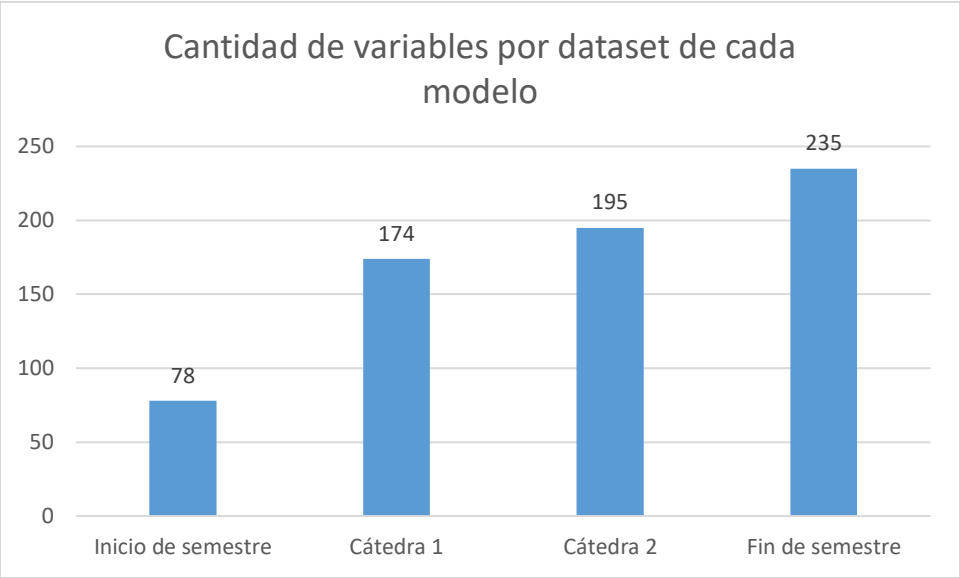

Figura 3-3 Gráfico de barras de la cantidad de variables consideradas para cada modelo del sistema de clasificación de alumnos desertores de primer año.

En el anexo Anexo 1 se muestran gráficos donde está la cantidad de variables por tipo disponibles en cada set de datos. Cabe destacar que algunas de estas variables fueron removidas por no tener muy poca variabilidad y no fueron incluidas en los listados de variables de la sección 3.4.

Las variables categóricas fueron codificadas utilizando one-hot encoding que se explica en la sección 2.6 haciendo que para el modelo la cantidad de variables fuera mayor, pero con varias variables binarias excluyentes.

3.5.2 Separación de conjuntos

Al realizar la separación de los datos en conjunto de entrenamiento, validación y prueba en para la posterior validación cruzada, se puede realizar una selección preliminar de variablesa, pero estudios recientes [43] establecen que esto podría reducir la capacidad de generalización del modelo, llegando a entregar clasificadores con menor poder de clasificación. Por esto se optó por realizar la separación de conjuntos sobre el total de variables y seleccionar variables en la etapa siguiente.

En primera instancia la separación de conjuntos se realizó con el algoritmo SOM Based Stratified Sampling (SBSS) [44], que mediante mapas autoorganizativos busca garantizar que todos los conjuntos sean representativos. Esta separación demoraba varias horas, por lo que se realizaron pruebas con separación aleatoria (SRS) explicada en la sección 2.4 y se observó que los modelos entrenados entregaban resultados prácticamente iguales, por lo que se optó por esta última para ahorrar tiempos de cómputo haciendo que se mantuviera el porcentaje de desertores entre los conjuntos sin considerar decimales.

Las Tabla 3-16 a Tabla 3-19 se muestran información sobre los conjuntos de entrenamiento y prueba para los set de datos construidos para los cuatro modelos propuestos en la sección 3.2.3. Cabe recordar que al usar validación cruzada de k iteraciones no es necesario contar con un conjunto de validación. En las tablas mencionadas se ve que la proporción de desertores se mantuvo entre los distintos conjuntos.

Tabla 3-16 Información en la base de datos para el modelo de inicio del semestre

	Totales	Entrenamiento	Prueba
Alumnos	10702	7491	3211
Retenciones	7301	5100	2201
Desertores	3401	2391	1010
% Desertores	32%	32%	32%

Tabla 3-17 Información base de datos para modelo Cátedra 1

	Totales	Entrenamiento	Prueba
Alumnos	11175	7822	3353
Retenciones	7625	5310	2315
Desertores	3550	2512	1038
% Desertores	32%	32%	31%

Tabla 3-18 Información de base de datos modelo Cátedra 2

	Totales	Entrenamiento	Prueba
Alumnos	11003	7702	3301
Retenciones	7620	5340	2280
Desertores	3383	2362	1021
% Desertores	31%	31%	31%

Tabla 3-19 Información de base de datos modelo Fin de semestre

	Totales	Entrenamiento	Prueba
Alumnos	10927	7648	3279
Retenciones	7614	5369	2245
Desertores	3313	2279	1034
Tasa	30%	30%	31%

3.5.3 Selección de características

Como se explicó en la sección 2.8.2, el método random forest es capaz de entregar una estimación de qué tan importante es cada variable para el modelamiento. El método escogido para esto fue la disminución media de la impureza de Gini, explicada en la misma sección, puesto que es la que se vio utilizada en otros trabajos similares citados en la sección 2.10.

3.5.4 Clasificador

Como se mencionó en la sección anterior, si bien el modelo final a utilizar es random forest, se optó por entrenar una red neuronal para tener otro algoritmo con el que comparar. A continuación, se detallan los parámetros escogidos para cada técnica junto con una breve descripción.

3.5.4.1 Red neuronal

Para comparar el resultado del random forest con otro método de clasificación se entrenaron 4 redes neuronales, una para cada base de datos descrita en la sección 3.2.3, con un conjunto de variables elegido en base al random forest entrenado con la misma base de datos. Como función de activación se utilizó RELU y los parámetros se buscaron utilizando validación cruzada k-fold con 10 iteraciones, entrenando los modelos con 300 épocas. Más detalles de esto se pueden encontrar en el Anexo 3.

3.5.4.2 Random Forest

Se entrenaron 4 random forest utilizando todas las variables disponibles de cada base de datos. Se ajustaron los siguientes parámetros:

- Número de árboles: número de árboles a construir para cada random forest
- Profundidad Máxima: La profundidad máxima que puede llegar a tener cada árbol de decisión del random forest
- Máximo características a probar: el número máximo de variables a probar en cada nodo de cada árbol para encontrar la que separe mejor
- Mínimo de muestras para separar: número mínimo de alumnos que debe haber en un nodo de un árbol para que se busque una separación
- Mínimo de muestras para hoja: mínimo de alumnos que debe haber en un nodo hoja (al final del árbol)
- Impureza de Gini mínima para separar: criterio para detener la búsqueda de mejor separación. Cuando un nodo tiene un valor de impureza de Gini mayor a este se busca mejorarlo separando nuevamente, si tiene un valor menor se detiene esa rama del árbol.

Para encontrar los parámetros óptimos de cada modelo se iteró 2000 veces utilizando rangos de búsqueda cercanos a heurísticas conocidas.

Para calcular la salida de los random forests se utilizó la media de la probabilidad entregada por los árboles del modelo, es decir, del siguiente modo:

- Las variables de un alumno se ingresan al random forest
- Cada árbol de decisión t del random forest llega a un nodo hoja con una proporción p_t de desertores, la que se considera como la probabilidad de que el alumno sea desertor, según el árbol t

- Se obtiene la probabilidad media con

$$p = \frac{\sum_1^T p_t}{T}, \tag{11}$$

donde T corresponde al total de árboles en el random forest.

3.6 INTERPRETACIÓN Y EXPLORACIÓN

A continuación, se explican cómo se realizó la interpretación de resultados, con las técnicas descritas en la sección 2.9.

3.6.1 Evaluación de los modelos

Los random forest entrenados entregan como salida pseudo-probabilidades de que cada alumno deserte, es decir, valores entre 0 y 1. Estos valores se utilizaron para construir curvas ROC, como fueron explicadas en la sección 2.9.1. Para esto se definieron 1000 umbrales θ, equiespaciados por $\Delta = 10^{-3}$. Con las curvas graficadas se calculó el área bajo la curva para tener una medida comparable de las curvas, como así también se compararon visualmente para tener una estimación de su rendimiento y comparar los resultados de las distintas pruebas realizadas.

Si bien las curvas ROC permiten estimar el poder de clasificación de un modelo para distintos umbrales es necesario fijarlo a un valor θ_c para utilizarlo en la práctica y para obtener la matriz de confusiones del conjunto de prueba. El valor de θ_c podría haberse definido buscando el punto de mayor precisión en la curva ROC, pero esto supondría que el costo de clasificar a un desertor como retención fuera equivalente al de clasificar a una retención como potencial desertor, lo que podría no ser cierto. A raíz de esto se debió escoger un θ_c que tomara en cuenta una estimación los costos de los dos tipos de errores en la clasificación de los alumnos.

Para estimar los costos descritos en el párrafo anterior es necesario recordar que los usuarios de las clasificaciones entregadas por los modelos serían los directores de carrera de UDLA, quienes estarían encargados de contactar a los alumnos riesgosos para evitar que dejen la institución. Con esto en mente se realizaron las siguientes suposiciones, debido a que la información exacta no existe:

1. Un director de carrera será capaz de retener al $s = 50\%$ de los alumnos que contacte.
2. El director de carrera invertirá aproximadamente $n = 45$ minutos al mes por alumno contactado durante 10 meses
3. El 50% del arancel de cada alumno cubre los gastos de operación que debe hacer la universidad para mantenerlo estudiando. Se supuso esto porque UDLA no posee un análisis detallado de estos costos.

Además de estas suposiciones se consideraron los siguientes valores:

1. El promedio de sueldo bruto de un director de carrera es de $\alpha = \$1.485.503$ pesos chilenos
2. Cuando un alumno se retira de la institución se le asigna un monto de retiro, que corresponde a la cantidad de dinero que ya no pagará a la institución, porque no seguirá estudiando. A partir de la información disponible en el Data Warehouse de UDLA se encontró que el promedio de este valor es de $m = \$981.347$
3. Un director de carrera trabaja aproximadamente $t = 8{,}5$ horas al día
4. En promedio se trabaja $d = 22{,}5$ días al mes por 12 meses.

Con estos valores se puede calcular el beneficio promedio por detección correcta y el costo de cada alumno clasificado como potencial desertor. En primer lugar se calcula el costo por hora de trabajo de un director de carrera como

$$D = \frac{s}{t \cdot d} \tag{12}$$

$$= \frac{\$1.485.503}{8{,}5 \cdot 22{,}5}$$

$$= \$7{,}766 \left[\frac{clp}{hora}\right].$$

En base a esto se puede calcular el costo anual por gestionar a un alumno G como

$$G = 10 \cdot D \cdot \frac{n}{60} \tag{13}$$

$$= 10 \cdot 7{,}766 \cdot \frac{45}{60}$$

$$= \$58.245 \left[\frac{clp}{alumno}\right]$$

Considerando la suposición 3 se puede estimar el beneficio de UDLA al evitar una deserción equivale a:

$$R = \frac{m}{2} \tag{14}$$

$$= \$490.674\ [clp].$$

Finalmente, se debe tomar en cuenta que se estima que sólo un 50% de los alumnos que el director de carrera gestione serán retenidos, por lo que se puede

calcular que, en promedio, la razón entre el ingreso que da un alumno retenido contra los costos de gestionarlo es:

$$\begin{aligned} I &= 0{,}5 \cdot \frac{R}{G} \\ &= 0{,}5 \cdot \frac{490.674}{58.245} \\ &= 4{,}12 \\ &\approx 4 \end{aligned} \tag{15}$$

Lo que significa que, en promedio un potencial desertor retenido entrega ingresos cuatro veces mayores a sus costos de gestión y operación. Con esto se puede definir la matriz de costo como:

$$C = \begin{bmatrix} 0 & 1 \\ 4 & 0 \end{bmatrix} \cdot 116.510, \tag{16}$$

donde el número 116.510 proviene de que la mitad de los alumnos gestionados son retenidos. La que se utilizó para evaluar los distintos puntos de la curva ROC de acuerdo a sus errores.

4 RESULTADOS

A continuación, se presentan los resultados de este estudio para los random forest entrenados con los datos de cada una de las etapas del semestre descritas previamente. También se entrenaron redes neuronales para tener una comparación, pero estos resultados se dejaron en el anexo de la sección Anexo 3.3.

4.1 PARÁMETROS DE LOS MODELOS PARA CADA ETAPA

Para obtener los mejores resultados posibles se optimizaron los parámetros de cada random forest. Para encontrar los parámetros óptimos se dividió el conjunto de entrenamiento en 10 segmentos y se recorrió una grilla de 2000 combinaciones de parámetros donde para cada combinación se hizo una validación cruzada de 10 iteraciones, explicada en la sección 2.4. En cada iteración de la validación cruzada se entrenó con 9 segmentos del conjunto y se validó con el restante. La combinación de parámetros seleccionada corresponde a la que en promedio entregó mayor precisión.

Como se expresó en la sección 2.8.2, el número de árboles de decisión de un random forest es necesario que sea lo suficientemente grande como para minimizar el error, pero aumentarlo más allá de ese punto no mejora los resultados. Utilizando

el procedimiento del párrafo anterior se determinó que 200 árboles era un número suficientemente grande para este problema.

En las Tabla 4-1 a Tabla 4-4 se muestran los parámetros encontrados para cada uno de los modelos. La explicación de cada parámetro se encuentra en la sección 3.5.4.2. Un punto a destacar es que en todos los casos se encontró que la profundidad máxima de los árboles no se fijaba con un valor predeterminado. Esto se debe a que el número de muestras para separar, el mínimo de muestras para hoja y la impureza mínima para separar, actúan como un límite de acuerdo a la estructura de cada árbol, lo que hace innecesario tener que fijar también la profundidad máxima.

Tabla 4-1 Parámetros óptimos para el random forest que predice si los alumnos desertarán durante el año utilizando el dataset construido con variables disponibles al inicio del semestre

Parámetro	**Valor óptimo**
Profundidad Máxima	Ilimitado
Máximo características a probar	4
Mínimo de muestras para separar	4
Mínimo de muestras para hoja	1
Impureza de Gini mínima para separar	8,60E-07

Tabla 4-2 Parámetros óptimos encontrados para el random forest que predice si los alumnos desertarán durante el año utilizando el dataset construido con variables disponibles luego de la cátedra 1

Parámetro	**Valor óptimo**
Profundidad Máxima	Ilimitado
Máximo características a probar	12
Mínimo de muestras para separar	4
Mínimo de muestras para hoja	2
Impureza de Gini mínima para separar	8.53E-05

Tabla 4-3 Parámetros óptimos encontrados para el random forest que predice si los alumnos desertarán durante el año utilizando el dataset construido con variables disponibles luego de la cátedra 2

Parámetro	**Valor óptimo**
Profundidad Máxima	Ilimitado
Máximo características a probar	13
Mínimo de muestras para separar	2
Mínimo de muestras para hoja	3
Impureza de Gini mínima para separar	6,63E-06

Tabla 4-4 Parámetros óptimos encontrados para el random forest que predice si los alumnos desertarán durante el año utilizando el dataset construido con variables disponibles al final del semestre

Parámetro	Valor óptimo
Profundidad Máxima	Ilimitado
Máximo características a probar	9
Mínimo de muestras para separar	4
Mínimo de muestras para hoja	1
Impureza de Gini mínima para separar	7.30E-05

4.2 DESEMPEÑO DE LOS MODELOS CON LOS DATOS DE LOS CONJUNTOS DE PRUEBA

En esta sección se muestran los resultados obtenidos por los distintos modelos construidos, evaluados con los conjuntos de prueba descritos, en la sección 3.5.2.

4.2.1 Curvas ROC

En esta sección se muestran las curvas ROC obtenidas por modelos entrenados con los conjuntos de entrenamiento descritos previamente. En la Figura 4-1 se muestra la curva ROC del inicio del semestre mostrando su capacidad de clasificación sobre los datos de prueba. Aquí se calculó que el área bajo la curva es de 0,68.

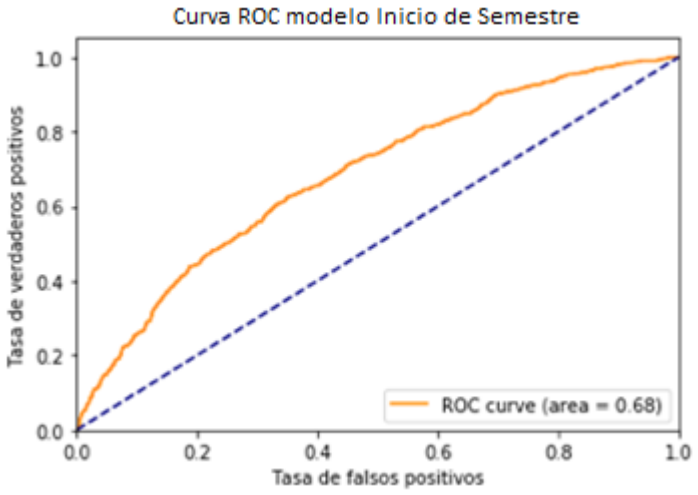

Figura 4-1 Curva ROC del modelo random forest correspondiente al inicio del semestre

Las Figura 4-2 a Figura 4-4 se muestran las curvas ROC de los modelos entrenados con los datos de la cátedra 1, de la cátedra 2 y del fin del semestre respectivamente. En las figuras también se incluyó el resultado del cálculo del área bajo la curva, siendo $0{,}8$, $0{,}83$ y $0{,}86$ respectivamente.

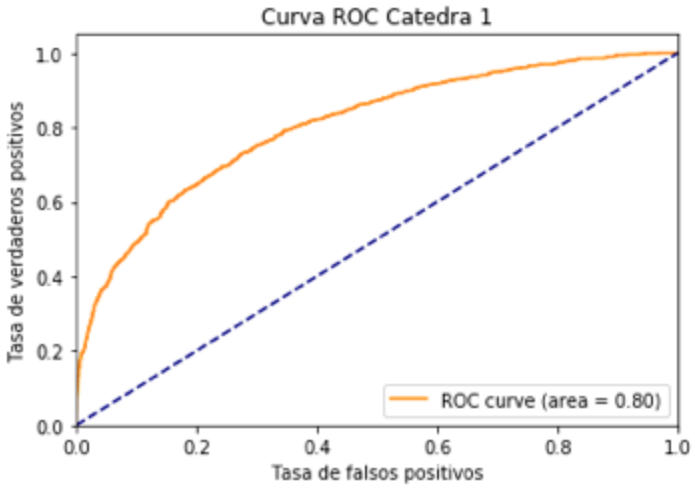

Figura 4-2 Curva ROC del modelo random forest correspondiente a la cátedra 1

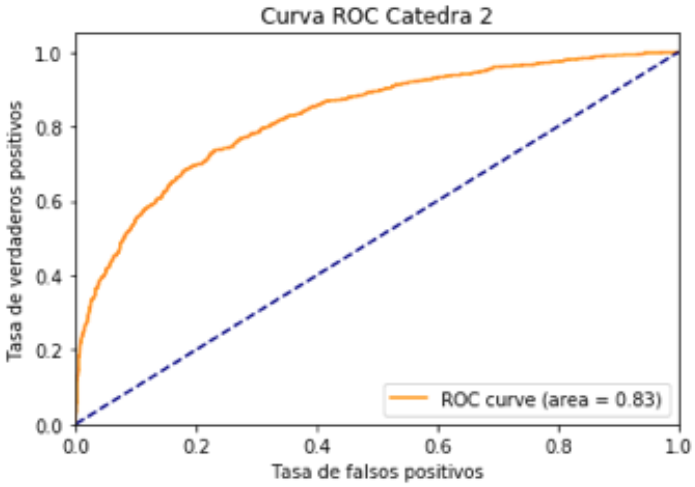

Figura 4-3 Curva ROC del modelo random forest correspondiente a la cátedra 2

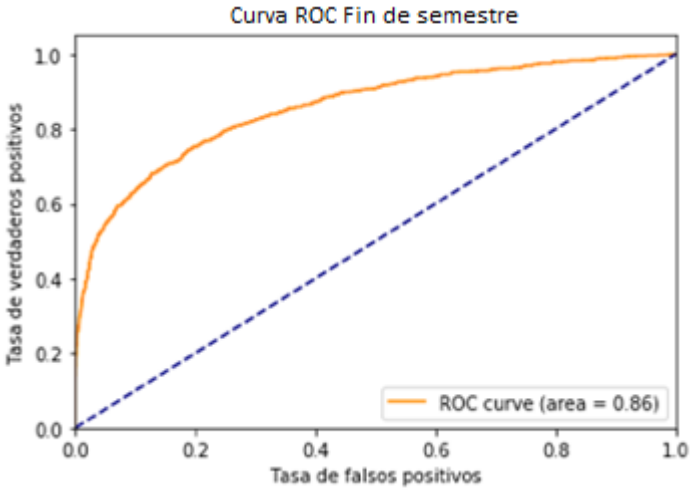

Figura 4-4 Curva ROC del modelo random forest correspondiente al final del semestre

4.2.2 Matrices de Confusión

4.2.2.1 Modelo Inicio de semestre

Las tablas Tabla 4-5 y Tabla 4-6 muestran las matrices de confusión del modelo random forests para el inicio del semestre. Las filas muestran los alumnos realmente retenidos o desertores, mientras que en las columnas se muestra la predicción realizada. A la derecha de la matriz se muestran tres valores de interés: especificidad, sensibilidad y precisión, en ese orden desde arriba hacia abajo.

En la Tabla 4-5 se ve que el modelo predice de manera correcta el 70,2% de las veces. Solo acierta a un 11,6% de los desertores, pero clasifica correctamente al 96% de las retenciones. Esto lo hace un modelo muy sesgado hacia los alumnos que permanecen en la institución.

Tabla 4-5 Matriz de confusión del random forest para el inicio del semestre, maximizando la precisión

		Predicción			
		Retención	Deserción		
Real	Retención	2141	89	Especificidad	96,0%
	Deserción	867	114	Sensibilidad	11,6%
				Precisión	**70,2%**

Si se modifica el umbral de clasificación incluyendo los costos descritos en la sección 3.6.1 los valores de la matriz de confusión cambian considerablemente, disminuyendo significativamente la precisión. En este caso se detecta correctamente al 94,8% de los desertores, pero a costa de clasificar erróneamente el 86,2% de los alumnos retenidos y acertándole solo al 38,5% del total de alumnos.

Tabla 4-6 Matriz de confusión del random forest para el inicio del semestre minimizando los costos

		Predicción			
		Retención	Deserción		
Real	Retención	305	1925	Especificidad	13,7%
	Deserción	51	930	Sensibilidad	94,8%
				Precisión	**38,5%**

4.2.2.2 Modelo Cátedra 1

Las tablas Tabla 4-7 y Tabla 4-8 están las matrices de confusión para el modelo random forest de la cátedra 1. En la primera tabla se muestran los resultados al elegir el umbral que maximiza la precisión, que es de un 76,7%, superior a la obtenida en el modelo anterior. Por otra parte, si bien la tasa de especificidad es menor que la del modelo anterior, se aprecia una sensibilidad mucho mayor, siendo cuatro veces la del modelo inicial mostrado en la Tabla 4-5.

Tabla 4-7 Matriz de confusión del random forest para la cátedra 1 maximizando precisión

		Predicción			
		Retención	Deserción		
Real	Retención	2074	183	Especificidad	91,9%
	Deserción	597	499	Sensibilidad	45,5%
				Precisión	**76,7%**

Al construir la matriz de confusión con el umbral que minimiza los costos, se ve que la precisión decae a un 64,8% y la especificidad baja a cerca de la mitad al compararla con la de la Tabla 4-7, llegando a un 55,4%. Estas bajas se ven acompañadas de una mejora relevante en la sensibilidad, pero el desempeño del modelo sigue siendo bajo.

Tabla 4-8 Matriz de confusión del random forest para la cátedra 1 minimizando los costos

		Predicción			
		Retención	Deserción		
Real	Retención	1251	1006	Especificidad	55,4%
	Deserción	174	922	Sensibilidad	84,1%
				Precisión	**64,8%**

4.2.2.3 Modelo Cátedra 2

En la Tabla 4-9 se muestra la matriz de confusión con el umbral que maximiza la precisión. Se aprecia que los resultados son similares a los de la Tabla 4-7,

llegándose a una precisión levemente mejor que la del modelo de la cátedra 1, con una especificidad levemente inferior y una sensibilidad más elevada, siendo un 8,4% superior.

Tabla 4-9 Matriz de confusión del random forest para la cátedra 2 maximizando precisión

		Predicción			
		Retención	Deserción		
Real	Retención	2040	221	Especificidad	90,2%
	Deserción	480	561	Sensibilidad	53,9%
				Precisión	**78,8%**

Similar al caso descrito previamente, la matriz de confusión de la Tabla 4-10, donde se minimizaron los costos, tiene resultados muy similares a los de la Tabla 4-8. La especificidad, sensibilidad y precisión tienen valores entre un 1 y un 3% superiores.

Tabla 4-10 Matriz de confusión del random forest para la cátedra 2 minimizando los costos

		Predicción			
		Retención	Deserción		
Real	Retención	1326	935	Especificidad	58,6%
	Deserción	153	888	Sensibilidad	85,3%
				Precisión	**67,1%**

4.2.2.4 Modelo Fin de semestre

Con los resultados obtenidos con los datos del fin de semestre se construyeron las matrices de confusión de esta sección. En la Tabla 4-11 se muestran los resultados al maximizar la precisión del modelo, llegando a un 82,5%. La especificidad es de un 95,3%, mientras que la sensibilidad llega al 52,1%.

Tabla 4-11 Matriz de confusión del random forest para el fin del semestre maximizando precisión

		Predicción			
		Retención	Deserción		
Real	Retención	2194	107	Especificidad	95,3%
	Deserción	468	510	Sensibilidad	52,1%
				Precisión	**82,5%**

La Tabla 4-12 muestra los resultados al minimizando los costos según la matriz de costo. En este caso la precisión baja a un 74,2% debido a que la especificidad baja a un 70%, mientras que la sensibilidad sube a un 84,2%.

Tabla 4-12 Matriz de confusión del random forest el fin del semestre minimizando los costos

		Predicción			
		Retención	Deserción		
Real	Retención	1610	691	Especificidad	70,0%
	Deserción	155	823	Sensibilidad	84,2%
				Precisión	**74,2%**

4.2.2.5 Costos de cada clasificador

A partir de las matrices de confusión de la sección anterior, se pueden calcular los costos de los errores de cada clasificador utilizando los dos puntos de operación mencionados. En la Tabla 4-13 se muestran estos valores al maximizar la precisión del modelo y al minimizar los costos. Se aprecia que al maximizar la precisión los costos son considerablemente más elevados, siendo un 67% más grande en el modelo del inicio del semestre, un 51% en el de la cátedra 1, un 38% en el de la cátedra 2 y un 51% al final del semestre.

Tabla 4-13 Costos de los modelos según el punto de operación escogido

Modelo	Maximizando precisión	Minimizando costos
Inicio de semestre	$ 414.426.070	$ 248.049.790
Cátedra 1	$ 299.547.210	$ 198.300.020
Cátedra 2	$ 249.447.910	$ 180.240.970
Fin de Semestre	$ 230.573.290	$ 152.744.610

4.3 Variables Más Relevantes

En esta sección se muestran las 10 variables más importantes de acuerdo a la importancia determinada por los random forest entrenados, ordenadas de manera descendente. En la Tabla 4-14 se muestran las variables más importantes del modelo del inicio del semestre, mientras que en la Tabla 4-15 las correspondientes al de la cátedra 1, en la Tabla 4-16 las de la cátedra 2 y finalmente en la Tabla 4-17 se muestran las 10 variables más importantes del modelo del final del semestre.

Tabla 4-14 Variables más importantes para el modelo del inicio del semestre

Ranking	Variable	Tipo	Importancia
1	Diferencia día matrícula e inicio de clases	Matrícula	0,076
2	Cantidad de alumnos en el programa	Compañeros	0,07
3	Arancel neto	Financiera	0,067
4	Promedio de compañeros por curso	Compañeros	0,066
5	Saldo por pagar	Financiera	0,066
6	Distancia comuna del hogar – comuna de estudio	Personal	0,059
7	Puntaje PSU lenguaje	Personal	0,057
8	Edad	Personal	0,055
9	% de retiros durante el semestre anterior	Compañeros	0,055
10	% de retiros semestre anterior en su programa a la fecha	Compañeros	0,051

Tabla 4-15 Variables más importantes para el modelo de la cátedra 1

Ranking	Variable	Tipo	Importancia
1	Pago total a la fecha	Financiera	0,04
2	Promedio de notas de la cátedra 1	Académica	0,039
3	Promedio de notas de ejercicios	Académica	0,037
4	Cantidad de notas en cátedra 1 sobre la media	Académica	0,036
5	Sesiones asistidas	Académica	0,033
6	Promedio del saldo moroso	Financiera	0,033
7	Días de Mora	Financiera	0,027
8	Cantidad de alumnos en el programa	Compañeros	0,025
9	Arancel neto	Financiera	0,024
10	Promedio de compañeros por curso	Compañeros	0,024

Tabla 4-16 Variables más importantes para el modelo de la cátedra 2

Ranking	Variable	Tipo	Importancia
1	Cantidad de inasistencias a la cátedra 2	Asistencia	0,044
2	Días de mora	Financiera	0,034
3	Promedio de notas en la cátedra 2	Rendimiento Académico	0,033
4	Cantidad de cuotas morosas	Financiera	0,032
5	Promedio de notas en ejercicios	Rendimiento Académico	0,031
6	Total de sesiones asistidas	Asistencia	0,03
7	Pago total realizado	Financiera	0,029
8	Cantidad de cuotas por pagar en el año	Financiera	0,027
9	Promedio de notas en la cátedra 1	Rendimiento Académico	0,026
10	Variación en el saldo moroso del alumno	Financiera	0,025

Tabla 4-17 Variables más importantes para el modelo del fin del semestre

Ranking	Variable	Tipo	Importancia
1	Días de mora	Financiera	0,05
2	Promedio final de asignaturas	Rendimiento Académico	0,04
3	Cantidad de inasistencias a exámenes	Asistencia	0,039
4	Cuotas por pagar	Financiera	0,033
5	Pago Total	Financiera	0,032
6	Promedio de notas en exámenes	Rendimiento Académico	0,031
7	Avance de malla	Rendimiento Académico	0,031
8	Cantidad de cuotas morosas	Variación Financiera	0,027
9	Variación del saldo moroso	Variación Financiera	0,027
10	Promedio de notas en ejercicios	Rendimiento Académico	0,025

4.4 COMPARACIÓN DE PROMEDIO DE VARIABLES MÁS IMPORTANTES EN MATRIZ DE CONFUSIÓN

En esta sección se muestran los promedios de las 10 variables más importantes según los random forest entrenados. Cada columna muestra el valor promedio de la variable de acuerdo con su posición en la matriz de confusión. En la Tabla 4-18 están los valores obtenidos con el modelo entrenado con datos del inicio del semestre. La Tabla 4-19 muestra la misma información para el modelo correspondiente a la cátedra 1, la Tabla 4-20 para la cátedra 2 y finalmente la Tabla 4-21 para el modelo entrenado con datos del final del semestre.

Tabla 4-18 Promedio del valor de las variables más importantes separadas según la matriz de confusiones del modelo del inicio del semestre para los 15 alumnos clasificados con mayor confianza

Variable	Verdaderos Negativos	Falsos Positivos	Verdaderos Positivos	Falsos Negativos
Diferencia día matrícula e inicio de clases	-92,93	-53,93	-47,00	-121,60
Cantidad de alumnos en el programa	385,2	59,33	50,27	435,73
Arancel neto	1.657.031	1.240.149	1.313.080	2.402.905
Promedio de compañeros por curso	36,07	33,20	31,28	35,08
Saldo por pagar	1.350.355	1.263.482	1.320.405	2.179.654
Distancia comuna del hogar comuna de estudio	31.607,93	5.156,40	7.437,00	106.608,3
Puntaje PSU lenguaje	393,53	45,07	72,07	327,8
Edad	23,00	28,00	23,87	25,47
% de retiros durante el semestre anterior	6,28	10,96	12,40	6,65
% de retiros semestre anterior en su programa a la fecha	3,31	4,86	5,23	3,60

Tabla 4-19 Promedio del valor de las variables más importantes separadas según la matriz de confusiones del modelo correspondiente a la cátedra 1 para los 15 alumnos clasificados con mayor confianza

Variable	Verdaderos Negativos	Falsos Positivos	Verdaderos Positivos	Falsos Negativos
Pago total a la fecha	669.703,10	72.781,87	33.751,93	678.496,40
Promedio de notas de la cátedra 1	5,38	3,16	1,15	5,42
Promedio de notas de ejercicios	6,08	4,54	2,88	5,99
Cantidad de notas en cátedra 1 sobre la media	3,87	0,73	0,07	3,93
Sesiones asistidas	158,87	77,33	42,00	133,53
Promedio del saldo moroso	22.015,22	130.958,90	103.266	21.973,17
Días de Mora	0,00	54,27	67,67	0
Cantidad de alumnos en el programa	433,80	88,67	99,07	336,13
Arancel neto	1.936.940	1.492.013	1.234.481	1.884.594
Promedio de compañeros por curso	185,85	190,24	180,47	204,13
Cantidad de cuotas morosas	0	2,73	3,73	0

Tabla 4-20 Promedio del valor de las variables más importantes separadas según la matriz de confusiones del modelo correspondiente a la cátedra 2 para los alumnos 15 clasificados con mayor confianza

Variable	Verdaderos Negativos	Falsos Positivos	Verdaderos Positivos	Falsos Negativos
Cantidad de inasistencias a la cátedra 2	0,00	1,87	3,67	0
Días de Mora	0,00	72,73	101,47	1,87
Promedio de notas de la cátedra 2	5,79	2,85	1,37	5,73
Cantidad de cuotas morosas	0,00	3,33	4,53	0,13
Promedio de ejercicios	6,06	4,06	4,24	6,00
Sesiones Asistidas	212,93	88,00	58,53	168,87
Pago total	870.539,40	198.264,47	33.040,20	637.390,47
Cuotas por pagar	7,67	11,07	12,53	7,67
Promedio de notas en la cátedra 1	5,43	3,58	3,53	5,44
Variación del saldo moroso	-0,28	1,68	1,98	-0,10

Tabla 4-21 Promedio del valor de las variables más importantes separadas según la matriz de confusiones del modelo para los 15 alumnos clasificados con mayor confianza

Variable	Verdaderos Negativos	Falsos Positivos	Verdaderos Positivos	Falsos Negativos
Dias de mora	0,00	98,73	120,53	1,33
Promedio final de asignaturas	5,73	2,40	1,52	5,58
Cantidad de inasistencias a exámenes	0,00	2,00	4,4	0
N° cuotas por pagar	5,47	10,80	12,53	6,67
Pago Total	1.016.492	171.757,67	52.748,6	888.164,47
Promedio de notas en exámenes	5,64	1,74	1,66	5,60
Avance de malla	10,72	8,83	0,52	10,37
Cantidad de cuotas morosas	0,00	4,00	5,8	0,2
Variación del saldo moroso	-0,26	1,73	2,23	0,016
Promedio de notas en ejercicios	6,12	4,26	2,95	6

5 ANÁLISIS DE RESULTADOS

En esta sección realiza el análisis y discusión de los resultados mostrados en la sección 4.

5.1 GENERALES

A partir de lo que se muestra en las tablas de las secciones 4.2.2 y Anexo 3.3 se puede establecer que al tener más información disponible de los alumnos se puede predecir con mayor efectividad si desertarán o no. El random forest entrenado con datos del inicio del semestre, que utiliza un umbral que maximiza la precisión, obtuvo una precisión de apenas 70,2%, en el de la cátedra 1 subió a 76,7%, en el de la cátedra 2 a 78,8% y finalmente el del final del semestre llegó a 82,5%, por lo que cada modelo es mejor que el anterior. Esto se ve confirmado con las curvas ROC mostradas en la sección 4.2.1 donde se ve como el área bajo la curva crece al pasar de los modelos cercanos al inicio del semestre hacia los más cercanos al final.

Este mismo comportamiento se vio en las redes neuronales entrenadas, donde los resultados también mejoran al avanzar en el semestre, pero en general la precisión de esos modelos fue inferior al de los random forest. Esto probablemente se deba a dos motivos principales. El primero es que los random forest son capaces de funcionar con una mayor cantidad de variables, por lo que se tuvo que incluir menos variables en las redes neuronales. El segundo motivo es que las variables se seleccionaron con la importancia de variables entregada por los random forest, lo que probablemente haga que haya algún sesgo hacia este modelo, y podría haber alguna variable que funcione mejor con las redes neuronales. Por lo anterior, y porque las redes neuronales se implementaron solo para tener una comparación, los próximos análisis se centrarán en los random forests.

5.2 MODELO INICIO DE SEMESTRE

En primer lugar, al observar los resultados mostrados en la Tabla 4-5 se aprecia que el random forest entrenado tiene una especificidad muy elevada, puesto que clasifica correctamente el 96% de los alumnos que no desertaron la universidad. El modelo logra ese resultado a costa de su sensibilidad, la que es extremadamente baja, llegando solo a un 11,6%, lo que lo hace muy sesgado hacia los no desertores. Hay que considerar que la tasa de desertores del conjunto de prueba utilizado es de 0,31, por lo que una precisión de 70,2% es levemente superior a clasificar a todos los alumnos como no desertores.

Al utilizar el mismo modelo, pero con un umbral que minimice los costos se llega a los resultados de la Tabla 4-6, donde se ve que la especificidad del modelo es extremadamente baja, llegando a un 13,7% y precisión del modelo llega apenas al

38,5%. Si bien la sensibilidad es de 94,8% la tasa de aciertos es muy baja, por lo que este modelo no es capaz de distinguir efectivamente entre los potenciales desertores y las retenciones.

Lo anterior también se confirma con la curva del mismo modelo, la que se muestra en la Figura 4-1. En la figura se aprecia que el área bajo la curva es de apenas 0,68 pasando relativamente cerca de la línea de identidad.

Cuando se utilizaron los mismos datos para entrenar una red neuronal en lugar de un random forest se obtuvieron los resultados mostrados en la Tabla 8-1, donde se aprecia que el modelo obtenido es inútil si se escoge un umbral que maximice la precisión puesto que clasifica prácticamente todos los datos como retenciones. Al cambiar el umbral para minimizar los costos se obtienen los resultados de la Tabla 8-2. La precisión y especificidad disminuyen en para aumentar la sensibilidad. Los resultados de este modelo son levemente mejores que los de la Tabla 4-6, pero aun así dejan mucho que desear.

Para entender mejor por qué este modelo no tiene un buen rendimiento es conveniente estudiar en más detalle la Tabla 4-14, donde se muestran las variables más importantes del modelo. Se aprecia que entre las ellas hay 3 variables que involucran a los pares del alumno, 3 de la matrícula, 1 financiera y 3 de información personal. Dentro de estas no hay ninguna variable que permita medir comportamiento, puesto que a esta altura del año no hay información que permita medirlo. En particular se observa que la variable más importante es la diferencia de días entre el inicio del semestre y la fecha en que el alumno se matricula, seguida de la cantidad de compañeros que el alumno tiene en el programa y en tercer lugar el arancel neto del alumno.

A modo de complementar lo anterior, al analizar la Tabla 4-18, se pueden estudiar en mayor detalle los comportamientos identificados por el modelo. Aquí se aprecia que los alumnos que el modelo tiende a clasificar como desertores se matricularon más cerca del inicio de clases que los detectados como retenciones. Los alumnos clasificados correctamente como desertores se matricularon en promedio menos de dos meses antes del inicio de clases, mientras que los detectados como retenciones lo hicieron más de 3 meses antes. La diferencia en días podría deberse a que los alumnos decididos por estudiar en UDLA tienden a matricularse el año anterior, mientras que los que lo hacen porque se quedaron sin más opciones tienden a ser más riesgosos en términos de deserción. La segunda variable más importante refleja que el modelo tiende a clasificar a los programas más pequeños como más riesgosos.

El resto de las variables no parecen ser capaces de discriminar de manera efectiva a los alumnos. Por ejemplo, en la tercera no parece haber un patrón claro, los alumnos que el modelo cree que son más propensos a desertar parecen tener en promedio un arancel más bajo, pero esto no parece tener mayor sentido. Las variables asociadas a la PSU tenían una gran cantidad de datos faltantes, cuyos

valores se imputaron con 0 para indicar que no había información, lo que afectó la calidad del dato y probablemente contribuyó a que no se observe que el modelo pueda utilizarla para discriminar entre desertores y retenciones. La distancia desde la comuna donde el alumno vive a la de estudio ni el porcentaje de retiros del semestre anterior en el programa tenían datos faltantes, pero tampoco parecen contribuir con información útil para el modelo. Esta falta de variables relevantes para el sistema explica el bajo rendimiento del modelo.

5.3 MODELO CÁTEDRA 1

Como ya fue dicho, este modelo tiene un rendimiento superior al del inicio del semestre. En la Tabla 4-7 se ve que al escoger un umbral que maximice la precisión, esta llega a un 76,7%, pero incluso más importante que eso es que la sensibilidad sube a 45,5%, siendo casi cuatro veces la del modelo analizado en la sección anterior, clasificando correctamente a 499 desertores en lugar de solo 114, donde cabe destacar que el conjunto de prueba de este modelo tiene un mayor número de desertores por a predecir que el del inicio del semestre. A pesar de lo anterior la especificidad decae desde un 96% a un 91,9%, lo que se traduce en que se clasificaron mal 183 alumnos como desertores, un alza de los 89 en el modelo anterior.

Si se considera la Tabla 4-8, donde se muestran los resultados obtenidos al minimizar los costos de las clasificaciones, se aprecia que la precisión baja considerablemente, a un 64,8%, pero que es muy superior al 38,5% obtenido por el modelo anterior, mostrado en la Tabla 4-6. En este caso se logra clasificar correctamente al 84,1% de los desertores y al 55,4% de las retenciones, lo que se traduce en que el modelo solo falla en detectar a 174 alumnos que dejaron la institución, pero se pasa de clasificar correctamente al 91,9% de las retenciones al considerar el umbral que optimiza la precisión a un 55,4%, pasando de 183 retenciones mal clasificadas a 1006.

La curva ROC de este modelo, mostrada en la Figura 4-2 también muestra resultados considerablemente superiores a los del modelo anterior. Tiene un área bajo la curva de 0,8 comparada a la de 0,68 de la sección anterior.

Luego revisamos la Tabla 4-15, donde se observan las 10 variables más importantes para este modelo. En la tabla se observa que la variable más importante es de tipo financiero y las 4 siguientes tienen que ver con el comportamiento académico del alumno, ya sea de asistencia o rendimiento. Como se observa en la Tabla 4-19 en promedio los alumnos que el modelo identifica como retenciones han pagado cerca de 670.000 pesos a la fecha en que se tienen los resultados de la cátedra 1, mientras que los alumnos clasificados como desertores han pagado cerca del 10% de este valor. Este comportamiento se condice con la experiencia de personal de UDLA encargada de recibir solicitudes de deserción, con quienes se habló en las entrevistas mencionadas en la sección 3.4.

La segunda variable más importante corresponde al rendimiento en la cátedra 1, donde se aprecia que los alumnos clasificados como desertores tienen promedios bajo 4 en sus cursos, mientras que las retenciones tienen un promedio superior a 5. Este rendimiento puede ser un indicador de que el potencial desertor alumno comienza a perder interés en su carrera.

La tercera variable es el promedio de notas en ejercicios, y tiene el mismo comportamiento que la anterior, pero las retenciones mal clasificadas tienen notas superiores a los desertores, mientras que los alumnos clasificados como retenciones tienen en promedio notas sobre 6.

La cuarta variable es la cantidad de notas que tuvo el alumno en la cátedra 1 sobre el promedio de su curso en todos los cursos que está tomando. Esta característica intenta medir el rendimiento del alumno en relación con sus pares, y se ve que los alumnos clasificados como desertores tienen en promedio menos de 1 nota sobre la media, mientras que las retenciones tienen sobre 3. Similar a las variables mencionadas en los párrafos anteriores, esta variable puede indicar que el alumno está perdiendo en su carrera.

En la quinta variable, que corresponde a las sesiones asistidas, se aprecia que en promedio los alumnos retenidos asisten a más del doble de sesiones de clases que los desertores, lo que es un indicador de la dedicación que el alumno tiene por su carrera. Algo de interés es que los alumnos retenidos detectados como deserciones en promedio asisten más que los verdaderos desertores.

La séptima variable más importante está muy relacionada a la variable más importante, donde se ve que los alumnos que han realizado menos pagos tienen días de mora y por ello son clasificados como potenciales desertores.

El resto de las variables tienen un comportamiento similar a lo analizado previamente o son difíciles de descifrar por qué son importantes para el modelo, como el caso de los compañeros por curso, donde no parece haber una tendencia distinguible.

Al compararlo con el modelo inicio del semestre, el modelo cátedra 1 tiene una mayor cantidad de variables que se puede entenderse por qué el modelo las consideró como importantes. Además, se ve que la información que no estaba disponible antes, como el rendimiento de la cátedra 1, la asistencia o la morosidad son de gran relevancia, lo que explica el mayor rendimiento de este modelo.

5.4 MODELO CÁTEDRA 2

Similar al caso anterior, este modelo tiene una mayor precisión que el de la cátedra 1. Al analizar la Tabla 4-9 se ve que la precisión alcanza un 78,8%, con una especificidad de 90,2% y una sensibilidad de 53,9%. Esto se traduce en que se

clasifican correctamente a 561 desertores del total de 1041 del conjunto de prueba. A su vez, 221 alumnos retenidos se clasifican como desertores. A diferencia de lo que ocurre entre el inicio de semestre y cátedra 1, la mejora entre el modelo de la cátedra 1 y este es menor: hay un aumento en los desertores detectados correctamente de un 18,5% pasando de un 45,5% a un 53,9%.

Al minimizar los costos la precisión disminuye a un 67,1%, pero logrando clasificar correctamente al 85,3%, equivocándose solo en 153 desertores, pero acertando solo al 58,6% de los alumnos retenidos, equivalente a 1326. Esto significa que el 51% de las veces que el modelo detecta a un desertor se estará equivocando. La Figura 4-3 también muestra una mejora en el área bajo la curva de la curva ROC, la que aumentó en 0,03 respecto de la del modelo de la cátedra 1 confirmando que posee un mayor poder de clasificación.

Respecto a las variables más importantes mostradas en la Tabla 4-16 se aprecia que 2 de las 3 primeras están relacionadas a comportamiento del alumno en la cátedra 2. La más importante corresponde a las inasistencias del alumno a esta evaluación, lo que indica que el alumno ya podría haber perdido interés en su carrera o está teniendo problemas para seguir estudiando. Complementando esto con la Tabla 4-20 se ve que las retenciones detectadas correctamente con mucha confianza no tienen inasistencias a la cátedra 2, mientras que las deserciones detectadas en promedio tienen más de 3. Los alumnos clasificados erróneamente como desertores tienen algunas inasistencias, mientras que los desertores clasificados como retenciones no las tienen.

En segundo lugar, se encuentran los días de mora, que también estaba presente entre las 10 variables más importantes del modelo anterior, pero estaba en la séptima posición, por lo que subió 5 posiciones. Esto puede deberse a que el alumno ya no es capaz de pagar su arancel o a que el alumno ha perdido interés en sus estudios por lo que no quiere seguir pagando. Complementando esto con la Tabla 4-20 se ve que los alumnos clasificados como retenciones tienen muy pocos días de mora, mientras que los desertores tienen valores muy altos, superiores a dos meses.

La tercera variable corresponde al promedio de notas del alumno en la cátedra 2 y tiene un comportamiento muy similar a la cátedra 1, donde se ve que los alumnos clasificados como desertores tienen un rendimiento considerablemente más bajo. Los alumnos clasificados correctamente como desertores con alta probabilidad tienen en promedio notas cercanas al 1.

El resto de las variables tienen un comportamiento similar a lo ya establecido, pero es de interés destacar la variable en la décima posición, donde se ve que los alumnos clasificados como retenciones tienden a haber reducido su saldo moroso, mientras que los clasificados como desertores tienden a haber aumentado su morosidad cerca de dos desviaciones estándar.

5.5 MODELO FIN DE SEMESTRE

El modelo del final del semestre es el que obtuvo mejores resultados porque utiliza el set de datos con mayor cantidad de información, teniendo disponible un total de 235 variables. En la Tabla 4-11 se aprecia que al maximizar la precisión del modelo se llega logra detectar correctamente al 52,1% de los desertores, y sólo se clasifican 107 retenciones como potenciales desertores. Esto significa que el 83% de las veces que el modelo clasifica a un alumno como potencial desertor está en lo correcto. Cabe destacar que el conjunto de prueba de este modelo tiene solo 13 desertores menos por predecir que el conjunto de prueba del modelo cátedra 2. Asimismo, detecta correctamente el 95,3% de las retenciones.

Al cambiar el punto de operación por el que minimiza los costos se llega a los resultados de la Tabla 4-12, donde se obtiene a una precisión de 74,2%, logrando clasificar correctamente al 84,2% de los desertores, pero bajando el porcentaje de detección de las retenciones a un 70%. Esto se traduce en que para captar a un 61% más de desertores es necesario contactar con un 145% más de alumnos.

Los resultados de este modelo llegan a una precisión similar a la obtenida por [4], que es el trabajo cuyo conjunto de datos se asemeja más al mostrado en este informe, puesto que los demás tenían proporciones de desertores muy diferentes o definieron los conjuntos de datos de otra manera porque tenían una mayor cantidad de información disponible. En particular esto se da en el caso de [6], donde se obtuvieron resultados superiores pero el volumen de datos con que contaban era mucho mayor y contaba con datos de todo el año, a diferencia de este trabajo donde actualmente solo hay información disponible de un semestre por año.

En vista de que este modelo obtuvo la mejor precisión, los resultados de la Tabla 4-13 se explican fácilmente, y se ve que este modelo obtiene los costos más bajos, siendo, siendo casi $200.000.000 más bajos que los del inicio del semestre al maximizar la precisión, y cerca de $100.000.000 más bajos al minimizar los costos.

Las 10 variables más importantes del modelo, mostradas en la Tabla 4-17, evidencian que a esta altura del año la información disponible al inicio del semestre ya no es relevante para el modelo. Todas las variables listadas son de comportamiento del alumno, habiendo 5 relacionadas al comportamiento de pago del alumno, 4 de rendimiento académica y 1 de asistencia. Esto se debe a que hay más información y de mayor confianza puesto proviene del Data Warehouse de UDLA.

La variable más importante corresponde a los días de mora del alumno, que en el modelo de la cátedra 2 estaba en segunda posición y en de la cátedra 1 en la séptima. Se ve una progresión clara de que al avanzar en el semestre se va

transformando en una variable determinante en la detección de desertores. Esto puede a que hubo un cambio en su situación económica o porque es un síntoma de que el alumno no va a seguir estudiando por lo que deja de preocuparse por pagar. En la Tabla 4-21 se muestra que en promedio los alumnos clasificados como desertores tienen valores de más de 100 días de mora, lo que es muy cercano al límite de morosidad fijado por UDLA para que sus alumnos tengan permitido seguir estudiando.

La segunda variable más importante es el promedio final de las asignaturas, donde se ve que los alumnos clasificados como desertores tienen un promedio de notas considerablemente más bajo, en el rango de reprobación, mientras que los alumnos clasificados como retenidos tienen promedios sobre 5. Similar a lo anterior, esto puede denotar falta de interés por parte del alumno en continuar sus estudios o dificultad en comprender los contenidos que se le enseñan, situación en la que los directores de carrera podrían ayudar a sus alumnos con mayor efectividad.

La tercera variable más importante corresponde a la cantidad de inasistencias a exámenes, donde se ve que los alumnos clasificados como desertores con alta probabilidad tienen 2 o más inasistencias, mientras que los clasificados como retenciones no tienen. Esta variable también es un indicador del interés del alumno por seguir estudiando.

Entre las variables restantes hay varias de comportamiento de pago, donde todas demuestran comportamientos similares: los alumnos clasificados como desertores tienden a no haber pagado sus cuotas del arancel y por ende estar morosos. Esto se ve reflejado en la cantidad de cuotas por pagar, el monto total pagado a la fecha, la cantidad de cuotas morosas y la variación del saldo moroso de los estudiantes.

Las variables de promedio de notas en ejercicios y exámenes también demuestran que los desertores tienden a obtener peores calificaciones que los alumnos retenidos. Lo anterior influye la última variable por mencionar, el avance de malla de los alumnos, donde se ve que al reprobar asignaturas no crece.

6 CONCLUSIONES

En primer lugar, se puede concluir que se cumplió el objetivo de construir un modelo capaz de predecir qué alumnos tienen mayor probabilidad de desertar. Si bien al inicio del semestre no se logró una precisión similar a la de [4], el resto de los modelos si llegaron a valores cercanos y en particular el modelo del final del semestre posee una precisión.

En segundo lugar, fue confirmada la suposición de que al tener una mayor cantidad de datos y de mayor confiabilidad la predicción es más certera y se vio que en general las variables nuevas incorporadas en cada modelo, que no estaban disponibles previamente, tuvieron más importancia que las usadas en los anteriores. Por ejemplo, al inicio del semestre la fecha de la matrícula en relación con el inicio de clases fue la más importante, pero en el resto de los modelos no tuvo mayor importancia. Algo similar ocurrió al pasar de la cátedra 1 a la cátedra 2, las notas de la primera ya no eran tan importantes como las de la segunda.

En tercer lugar, y relacionado a lo anterior, se vio que los tipos de variables importantes también fueron cambiando a lo largo del semestre. En un principio la información personal y de matrícula del alumno era muy importante, pero luego la académica, financiera y de asistencia tomaron mayor relevancia en los modelos siguientes. También, dentro de los atributos más importantes se observò una variable de variación en el comportamiento del alumno en los dos últimos modelos.

En cuarto lugar, en base a los resultados se puede establecer que, con la información actual, la predicción al iniciar el semestre no entrega mayor valor y para mejorarla probablemente sería necesario incluir otras fuentes de información. Ejemplos de esto podría ser información sobre el alumno en el colegio, no solo el NEM, sino información más detallada de su asistencia, si reprobó cursos o ranking en la PSU o SIMCE del colegio de origen del alumno.

En quinto lugar, si bien la estimación de costos en los errores de clasificación de los modelos fue construida con suposiciones importantes puesto que actualmente no hay información más precisa, sirvió para obtener una aproximación del impacto que tendría el modelo sobre UDLA y sus alumnos. Estos costos por errores de clasificación son de cientos de millones, habiendo más de $200.000.000 de diferencia entre el mayor y el menor: la baja precisión del primer modelo se tradujo en costos casi un 80% superiores a los del modelo del final del semestre al maximizar la precisión y un 60% más altos al minimizar los costos.

Para mejorar los resultados obtenidos en este trabajo se recomienda lo siguiente como trabajo futuro:

- Incluir datos del segundo semestre en el modelamiento. Al haber más datos es probable que la precisión de estos modelos sea aún más alta que la actual.

Para esto es necesario esperar a que se cuente con información del segundo semestre de los años que vienen.

- Incluir más variables de otras fuentes de datos e idealmente buscar información si es posible obtener información sobre el comportamiento de los alumnos en el colegio, variables como asistencia, asignaturas reprobadas o si repitió algún año pordían mejorar la clasificación.

- Volver a entrenar el modelo con datos del año 2017. Al momento de crear los conjuntos de datos para este trabajo todavía no se contaba con toda la información de los desertores que no volvieron el año 2018, por lo que se optó por dejar fuera esta información para comenzar el análisis de datos. Al agregar estos casos al dataset no solo se contará con un mayor volumen de datos, sino que también se incluirán datos más actualizados.

- Replicar el conjunto de datos utilizado por el trabajo citado de investigación sobre colegios daneses [6], donde la información de los desertores se obtuvo 3 meses antes de que desertaran y la de los no desertores en una fecha aleatoria. Esta modelación podría entregar mejores resultados, pero es necesario contar con datos del año completo para poder crear el set de datos, cosa que actualmente no está disponible.

- Estudiar el impacto de crear modelos separados por carrera o área de estudio, para lo que habría que esperar a que hubiera datos de más años. Este enfoque permitiría estudiar en más detalle el comportamiento del alumno por asignatura e identificar cuáles son críticas para su éxito en la universidad.

- Extender el modelo a alumnos antiguos, siendo de particular interés los alumnos de segundo año puesto que tienen una tasa de deserción levemente menor a la de los alumnos de primer año. Si bien la tasa de deserción de estos alumnos es más baja que la de los de primer año, también se cuenta con una mayor cantidad de información, pudiéndose comparar las variables de su año de estudios actual con la de los años anteriores.

- Separar a los alumnos que no pueden seguir estudiando por morosidad en una etiqueta diferente puesto que los directores de carrera no tienen la habilidad de ayudar a estos alumnos directamente.

7 GLOSARIO DE TÉRMINOS Y ABREVIACIONES

1. **Cátedra**: evaluación principal intermedia en los cursos de la Universidad de las Américas. Suelen realizarse 3 a lo largo del semestre
2. **Confianza**: qué tan probable es, según el modelo, que un alumno sea desertor o retención. Un alumno clasificado con probabilidad 1 de ser desertor sería tendría la máxima confianza posible.
3. **Desertor**: alumno que deja la institución de educación superior en algún momento entre el momento que inician las clases y abril del año siguiente. Tiene subtipos:
    - Retiro: el alumno realiza una solicitud formal de retiro a la institución, llevando una carta y documentos que respalden sus motivos, en caso de ser necesario.
    - Desertor espontáneo: son los alumnos que debían tomar carga académica pero no lo hacen, por ejemplo, un estudiante que debía pasar a segundo año, pero nunca se presenta a firmar los documentos para matricularse al principio del semestre.
    - Alumnos vigentes sin carga: son alumnos que están matriculados y pueden tomar asignaturas, pero no lo hacen.
    - Deserción por política: no se le permite tomar carga académica alumnos con más de 4 meses morosos y si esta situación se mantiene por un semestre los alumnos son dados de baja, como si se hubiesen retirado de la institución.
    - Resciliación: Alumno que se matricula peor se retira antes del inicio de clases
4. **KDD**: Knowledge Discovery in Databases
5. **Precisión**: Tasa de aciertos del modelo para un punto de operación
6. **Retención**: alumno que permanece estudiando en UDLA
7. **Conjunto de Datos**: conjunto de datos utilizados para analizar y modelar el problema
8. **SIES**: Servicio de Información de Educación Superior, institución estatal dependiente del Ministerio de Educación encargada de recopilar información de las instituciones de educación superior de Chile e informar
9. **UDLA**: Universidad de las Américas

8 BIBLIOGRAFÍA

[1] L. Breiman, «Random Forests,» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[2] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From Data Mining to Knowledge Discovery in Databases,» *AI Magazine,* vol. 17, nº 3, pp. 37-54, 1996.

[3] R. Alkhasawneh y R. Hobson, «Modeling student retention in science and engineering disciplines using neural networks,» de *Global Engineering Education Conference (EDUCON), 2011 IEEE*, Richmond, USA, 2011.

[4] D. Delen, «A comparative analysis of machine learning techniques for student retention management,» *Decision Support Systems,* vol. 49, nº 4, pp. 498-506, 2010.

[5] Stratton, Ottole y Wetzel, «A Multinomial Logit Model of College Stopout and Dropout Behavior,» *Economics of Education Review,* vol. 27, nº 3, pp. 319-331, 2008.

[6] N.-B. Sara, R. Halland, C. Igel y S. Alstrup, «High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study,» de *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, Bruges, Belgium, 2015.

[7] Insituto Nacional de Estadística, «Compendio Estadístico,» 2017. [En línea]. Available: http://www.ine.cl/docs/default-source/publicaciones/2017/compendio-estadistico-2017.pdf?sfvrsn=6.

[8] Ministerio de Desarrollo Social, «Archivo histórico CASEN,» [En línea]. Available: http://observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_usuarios.php.

[9] E. Himmel, «Modelos de Análisis de la Deserción Estudiantil en la Educación Superior,» 2002. [En línea]. Available: http://www.alfaguia.org/alfaguia/files/1318958524Modelo%20de%20analisis%20de%20la%20desercion%20estudiantil%20en%20la%20educacion%20superior.pdf. [Último acceso: 17 06 2014].

[10] Centro de Estudios MINEDUC, «Serie Evidencias: Deserción en la Educación Superior En Chile,» [En línea]. Available: http://www.mineduc.cl/usuarios/bmineduc/doc/201209281737360.EVIDENCIASCEM9.pdf.

[11] SIES, «Informe Retención de 1er Año de Pregrado Cohortes 2012 - 2016,» Ministerio de Educación, Santiago, Chile, 2018.

[12] González y Uribe, «Estimaciones sobre la repitencia y deserción en la educación superior chilena. Consideraciones sobre sus implicancias,» *Revista de la Calidad de la Educación,* vol. 17, pp. 75-90, 2002.

[13] Fishbein y Ajzen, Belief, attitude, intention and behavior: An Introduction to Theory and Research, MA, EE.UU.: Addison - Wesley, 1975.

[14] Attinasi, «Getting in: Mexican American Students' perceptions of their college-going behavior with implications for their freshman year persistance in the University,» *The Journal of Higher Education,* vol. 60, nº 3 , pp. 247-277, 1986.

[15] C. A. Ethington, «A psychological model of student persistence,» *Research in Higher Education,* vol. 31, nº 3, pp. 266-269, 1990.

[16] G. Becker, Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Chicago, USA: University of Chicago, 1964.

[17] W. Spady, «Dropouts from higher education: An interdisciplinary review and synthesis,» *nterchange,* vol. 1, nº 1, pp. 64-85, 1970.

[18] J. Braxton, J. Milem y A. Shaw-Sullivan, «The Influence of Active Learning on the College Student Departure Process,» *The Journal of Higher Education,* vol. 71, nº 5, pp. 569-590, 2000.

[19] S. Tillman, «Barriers to student persistence in higher education,» 2002. [En línea]. Available: https://www.whdl.org/sites/default/files/v2n1_Tillman.pdf. [Último acceso: 2014 06 17].

[20] J. Braxton, R. Johnson y A. Shaw-Sullivan, «Appraising Tinto's theory of college student departure,» *Higher education: Handbook of theory and research,* vol. 12, nº 1, pp. 107-164, 1997.

[21] *Ley N°20.129 "Establece un sistema nacional de asuguramiento de la calidad de la educación superior". Diario Oficial de la República de Chile, Santiago, Chile.,* 17 de Noviembre de 2006.

[22] A. P. Engelbrecht, «Introduction to Computational Intelligence,» de *Computational Intelligence: An Introduction*, Pretoria, South Africa, Wiley, 2003, pp. 3-13.

[23] Duda, Hart y Stork, Pattern Clasification, 2nd Ed, New York, USA: J. Wiley & Sons, Inc, 2001.

[24] T. Hastie, R. Tibshirani y J. Friedman, «Chapter 7: Model Assessment and Selection,» de *The Elements of Statistical Learning: Data Mining, Inference and Prediction 2nd Edition*, California, USA, Springer, 2009, pp. 220-223.

[25] H. L. Surendra K. Singhi, «Feature Subset Selection Bias for Classification Learning,» de *ICML '06 Proceedings of the 23rd international conference on Machine learning* , Pittsburgh, Pennsylvania, USA, 2006.

[26] B. Lebaron y A. S. Weigend, «A bootstrap evaluation of the effect of data splitting on financial time series,» *Neural Networks, IEEE Transactions on,* vol. 9, nº 1, pp. 213-220, 1998.

[27] A. Janecek, W. Gansterer, M. Demel y G. Ecker, «On the relationship between feature selection and classification accuracy,» de *FSDM'08 Proceedings of the 2008 International Conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery - Volume 4*, Antwerp, Belgium, 2008.

[28] S. V. R. Kumar, «Analysis of Feature Selection Algorithms on Classification: A Survey,» *International Journal of Computer Applications,* vol. 96, nº 17, pp. 29-35, 2014.

[29] I. Guyon y A. Elisseeff, «An Introduction to Variable and Feature Selection,» *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[30] B. Kumari y T. Swarnka, «Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review,» *International Journal of Computer Science and Information Technologies,* vol. 2, pp. 1048-1053, 2011.

[31] K. Potdar, T. S. Pardawala y C. D. Pai, «A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers,» *International Journal of Computer Applications,* vol. 175, nº 4, pp. 7-9, 2017.

[32] Little y Rubin., Statistical Analysis with Missing Data 2nd Edition, Wiley-Interscience, 2002, pp. 3-19.

[33] G. Biau, «Analysis of a Random Forests Model,» *Journal of Machine Learning Research,* vol. 13, nº 1, pp. 1063-1095, 2012.

[34] T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, New York: Springer, 2009.

[35] M. Fernández-Delgado, E. Cernadas, S. Barro y D. Amorim, «Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?,» *Journal of Machine Learning Research,* vol. 15, nº 1, pp. 3133-3181, 2014.

[36] M. Wainberg, B. Alipanahi y B. J. Frey, «Are Random Forests Truly the Best Classifiers?,» *Journal of Machine Learning Research,* vol. 17, nº 110, pp. 1-5, 2016.

[37] T. Fawcett, «An introduction to ROC analysis,» *Pattern Recognition Letters,* vol. 27, nº 8, pp. 861-874, 2006.

[38] A. Salazar, J. Gosalbez, I. Bosch, R. Miralles y L. Vergara, «A case study of knowledge discovery on academic achievement, student desertion and student retention,» de *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on*, Valencia, España, 2004.

[39] H. Dissanayake, D. Robinson y O. Al-Azzam, «Predictive Modeling for Student Retention at St. Cloud State University,» de *International Conference on Data Mining*, Las Vegas, 2016.

[40] M. G. S. Vergara, «Diseño de una estrategia para disminuir las deserciones tempranas de los estudiantes del pregrado de la universidad de chile, proyecto de grado para optar al grado de magíster en ingeniería de negocios con tecnologías de información,» Universidad de Chile, Santiago, Chile, 2017.

[41] S. F. Sabbeh, «Machine-Learning Techniques for Customer Retention: A Comparative Study,» *International Journal of Advanced Computer Science and Applications,* vol. 9, nº 2, pp. 273-281, 2018.

[42] Google, «Web Services Directions API - Directions,» [En línea]. Available: https://developers.google.com/maps/documentation/directions/start.

[43] X. Zhu y Y. Yang, «Variable selection after screening: with or without data splitting?,» *Computational Statistics,* vol. 30, nº 1, p. 191–203, 2014.

[44] R. May, H. Maier y G. Dandy, «Data splitting for artificial neural networks using SOM-based stratified sampling,» *Neural Networks,* vol. 23, nº 2, pp. 283-294, 2010.

[45] A. Ultsch y H. P. Siemon, «Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis,» de *In Proceedings of International Neural Networks Conference (INNC), pp 305-308*, Dortmund, Germany, 1990.

[46] A. Ultsch, «Clustering With SOM: U*C,» de *Workshop on Self-Organizing Maps. Paris, pp 75-82*, Paris, France, 2005.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique,» *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[48] S. Haykin, Neural Networks and Learning Machines, 3rd Edition, Ontario, Canada: Pearson, 2009.

[49] I. Basheer y M. Hajmeer, «Artificial neural networks: fundamentals, computing, design, and application,» *Journal of Microbiological Methods,* vol. 43, pp. 3-31, 2000.

[50] E. Wan, «Finite Impulse Response Neural Networks with Applications in Time Series Prediction,» de *Thesis for the degree of Doctor of Philosophy*, Palo Alto, USA, 1993.

Anexo 1. Tipos de Variable disponibles para cada modelo

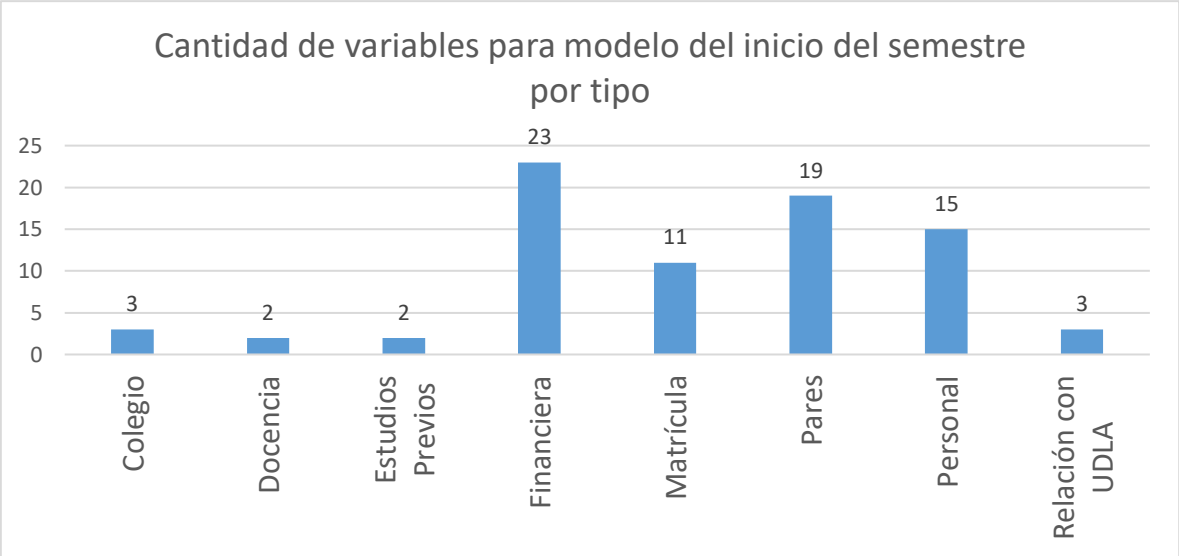

Figura 8-1 Cantidad de variables por tipo para el modelo del inicio del semestre

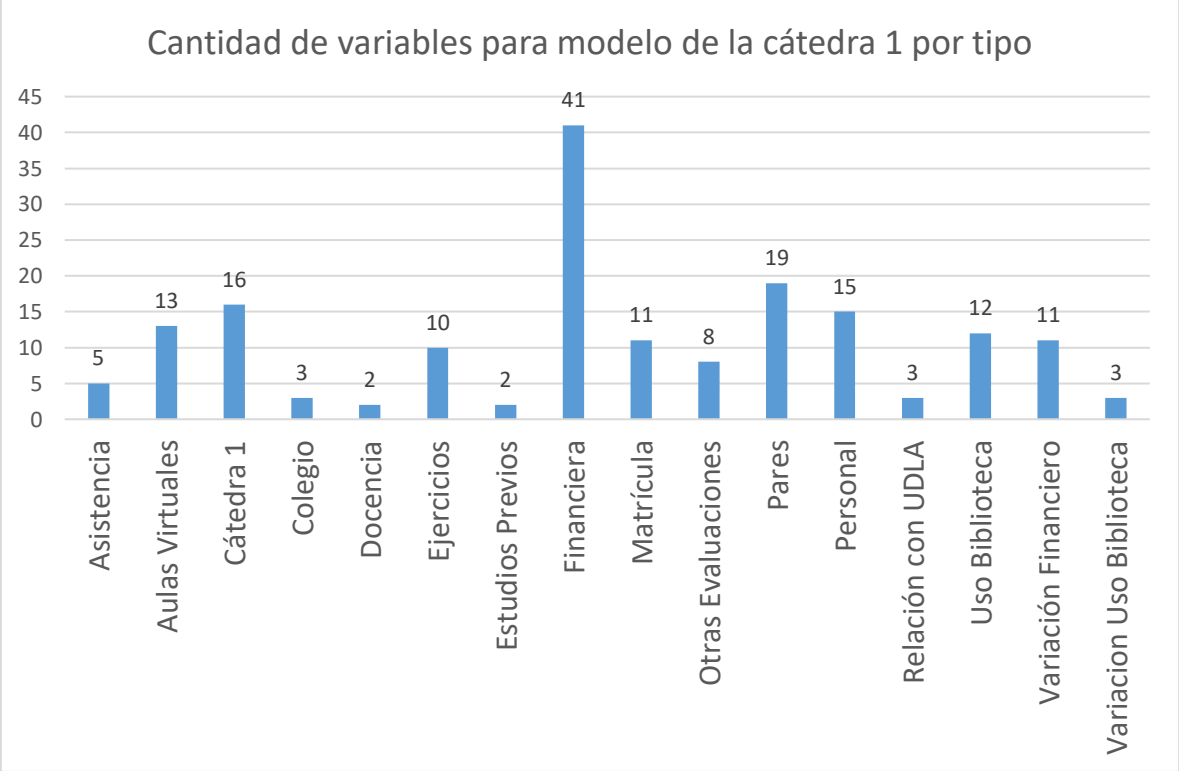

Figura 8-2 Cantidad de variables por tipo para el modelo del inicio de la cátedra 1

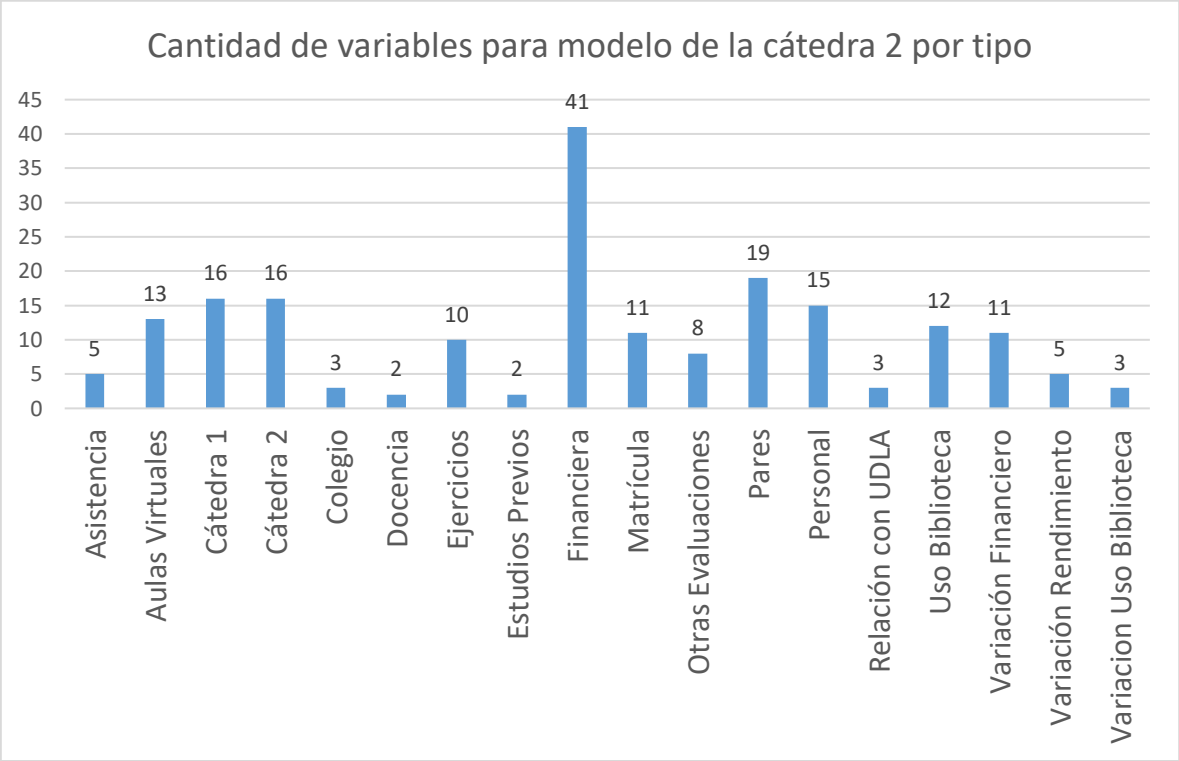

Figura 8-3 Cantidad de variables por tipo para el modelo de la cátedra 2

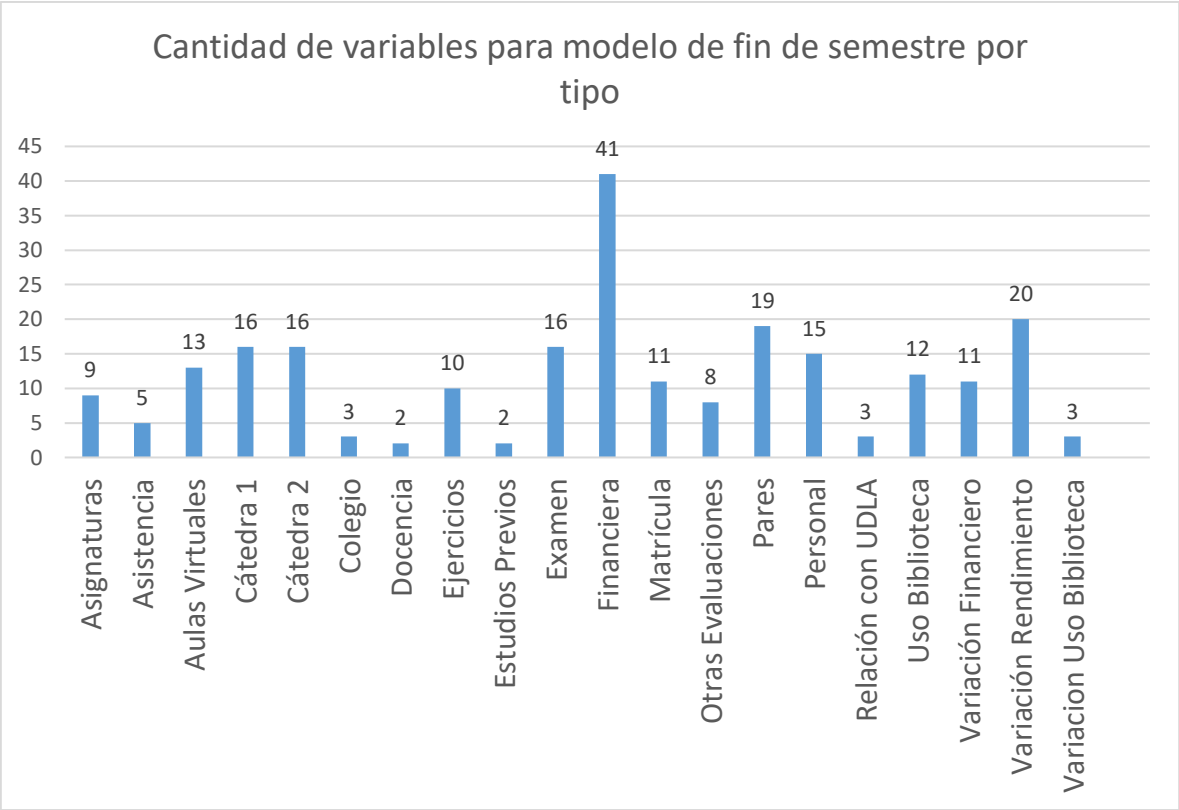

Figura 8-4 Cantidad de variables por tipo para el modelo del final del semestre

Anexo 2. METODOLOGÍAS UTILIZADAS QUE NO ENTREGARON MEJORAS

Anexo 2.1 SOM BASED STRATIFIED SAMPLING

Anexo 2.1.1 MAPAS AUTOORGANIZATIVOS

El propósito de SOM es representar puntos desde un espacio de características a otro, usualmente reduciendo el número de dimensiones y manteniendo las relaciones de distancia y proximidad dentro de lo posible. El mapeo se aprende con una red neuronal simple de dos capas con N entradas, donde N es la cantidad de dimensiones del espacio de características, conectada a $P \times Q$ salidas, donde P es el alto y Q el ancho deseado del mapa.

Para cada vector de características $\vec{x}$ se calcula la activación de la red en cada nodo k de la salida con la ecuación:

$$net_k = \sum_i \vec{x}_i w_{ki},$$

donde cada elemento w_{ki} de la ecuación es el peso correspondiente al i-ésimo elemento de $\vec{x}$. Al nodo con el valor máximo de activación y^* junto a sus vecinos se les actualizan los pesos de acuerdo a:

$$w_{ki}(t+1) = w_{ki}(t) + \eta(\mathrm{t})\Lambda(|y - y^*|)\vec{x}_i,$$

con t el número de iteración del algoritmo, y cada uno de los nodos vecinos de y^* cuyos pesos se están alterando, $\eta(t)$ un factor de aprendizaje que depende de t y la función $\Lambda(|y - y^*|)$ se denomina función ventana. Este vale 1 para $y = y^*$, mientras que tiene valores menores para argumentos más grandes, haciendo que sea vital para que el algoritmo tenga éxito, puesto que asegura que los nodos vecinos del espacio objetivo tengan pesos similares, emulando la similitud de los puntos vecinos en el espacio de origen.

El aprendizaje de los mapas SOM es muy general y puede aplicarse prácticamente a cualquier espacio de origen, objetivo y mapeo no lineal. Una limitante de esta técnica proviene de la inicialización de los pesos para la primera iteración, ya que la elección de este número puede producir torceduras en el mapa. Una condición inicial particular podría generar que un extremo del mapa identifique un patrón y otro extremo también lo identifique, perdiéndose la relación entre ambos [23].

Una regla heurística para determinar el tamaño del mapa es $P \times Q = \beta K^{0.54}$, donde $\beta \in (0{,}2\,;1\,;5)$, K es el número de muestras en la base de datos y P y Q son el alto y ancho del mapa.

Hay dos formas clásicas de visualizar los mapas SOM:

- U-Matrix: es una matriz que muestra la relación entre neuronas vecinas del mapa. Los elementos que aparecen en el mapa corresponden a distancias promedio entre cada nodo y sus vecinos inmediatos. Usualmente, para obtener una visualización simple de la similitud entre nodos, se usa un código de colores en base a los valores calculados [45].
- P-Matrix: en esta matriz cada elemento representa la densidad de puntos en la vecindad de cada nodo de la grilla. Cada uno de éstos representa una hipérbola de radio R que subdivide el espacio, donde R se elige para el conjunto de datos según su función de probabilidad [46]. Con esto se tiene una visualización de la distribución de puntos en el mapa.

Anexo 2.1.2 SOM BASED STRATIFIED SAMPLING

El algoritmo sigue los siguientes pasos [44]:

1. Establece el tamaño del mapa
2. Entrena
3. Divide los datos en cada uno de los nodos, en la proporción escogida para cada uno de los conjuntos de entrenamiento validación y prueba.

De este modo se escoge la misma proporción de datos desde cada uno de los nodos, lo que ayuda a que los conjuntos sean representativos no solo en la proporción de las etiquetas, sino también en la distribución del espacio de características.

Anexo 2.2 BALANCEO DE BASES DE DATOS

Una base de datos está balanceada si sus clases están representadas en una proporción igual o al menos comparable, condición que comúnmente no se cumple en aplicaciones reales. Por ejemplo, en detección de fraude la mayoría de los clientes no los cometen, llegando a ser 100 veces más prevalentes que los casos de interés, mientras que en otros estudios esta razón puede llegar a ser de 100.000 a 1.

Un desbalance de clases puede traer problemas graves de entrenar un modelo, existiendo la posibilidad de llegar a un clasificador trivial cuya predicción siempre la clase dominante. Un ejemplo de esto son las mamografías, donde usualmente un 98% de sus pixeles son normales, mientras que sólo el 2% restante son anómalos. Un clasificador que etiquete todos los pixeles como normales tendría un 98% de aciertos, pero no estaría cumpliendo su función: detectar problemas de salud.

Típicamente se utiliza una de las siguientes técnicas para tratar con bases de datos desbalanceadas:

1. Asignando un costo a los diferentes tipos de errores que produzca el clasificador, de modo que se prioricen los casos de interés.
2. Alterando las cantidades de cada clase en la base de datos, ya sea submuestreando la clase dominante o sobremuestreando la clase minoritaria (repitiendo ejemplos).

En este trabajo se utilizó una metodología que combina ambas: Synthetic Minority Over-sampling Technique (SMOTE), que mezcla el submuestreo con una forma particular de sobremuestero. Considerando cada elemento del conjunto de datos como un vector, el algoritmo de SMOTE es el siguiente:

1. Se buscan los k vecinos más cercanos a cada una de las muestras de la clase minoritaria
2. La diferencia entre cada muestra y sus vecinos se multiplica por un número aleatorio en 0 y 1
3. Lo anterior se suma al valor de la muestra inicial, tomándose un punto entre ella y cada uno de sus vecinos, lo que se muestra en la Figura 8-5, estando en azul el ejemplo nuevo.

Así se crean ejemplos sintéticos de la clase minoritaria en el espacio de características, hasta llegar a un sobremuestreo fijado por el usuario, por ejemplo el algoritmo puede duplicar o triplicar la cantidad de muestras de la clase minoritaria.

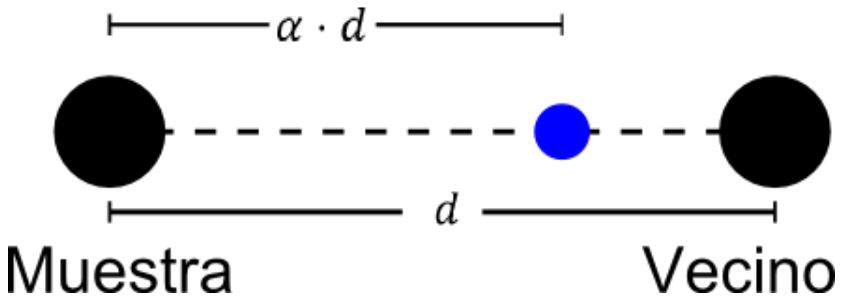

Figura 8-5 Diagrama de la creación de ejemplos sintéticos con el algoritmo de balanceo de bases de datos Synthetic Minority Over Sampling Technique (SMOTE)

Además de esto, SMOTE submuestrea algunos casos de la clase dominante de manera aleatoria hasta que se cumpla una razón fijada por el usuario. Cabe destacar que todo esto se realiza sólo sobre el conjunto de entrenamiento, de lo contrario se podrían incluir sesgos indeseados o ruido en las bases de validación y prueba, alterando la evaluación del modelo.

En general este método se comporta mejor que submuestreo o sobre-muestreo por sí solas [47].

Anexo 2.2.1 RESULTADOS CON SMOTE

Se realizaron pruebas para determinar si la normalización y balanceo de datos efectivamente mejorarían el desempeño de los modelos. El conjunto de entrenamiento fue balanceado con SMOTE. Esto se hizo tanto para redes neuronales como para random forests.

En el caso de la normalización se probó estandarización y escalamiento, ambos descritos en la sección Anexo 3.2, aplicándose tanto al conjunto de entrenamiento como de validación. De este modo, para la red neuronal se probaron las siguientes combinaciones:

- Datos en bruto, sin normalizar ni balancear
- Entrenamiento balanceado
- Entrenamiento y validación estandarizados
- Entrenamiento y validación escalados
- Entrenamiento y validación estandarizados, además de entrenamiento balanceado
- Entrenamiento y validación escalados, además de entrenamiento balanceado

Se entrenaron 2000 redes neuronales para las diferentes combinaciones de normalización y balanceo sobre las 4 bases de datos disponibles. Los parámetros escogidos para esta prueba corresponden a los recomendados en [47]: 5 vecinos y un 400% de porcentaje de SMOTE. La precisión y puntajes máximos promedio obtenidos fueron prácticamente iguales para todos los casos, pero al estudiar la varianza de estos valores se notaron diferencias considerables.

En la Figura 8-6 se muestran gráficos de barra para la varianza de la precisión máxima, aquí se aprecia que SMOTE con escalamiento y escalamiento sin balanceo muestran las varianzas más bajas. Aun más, el balanceo no parece tener efecto alguno sobre la varianza, puesto que SMOTE con normalización, sólo normalización y los datos en bruto tienen prácticamente la misma varianza.

A raíz de lo expuesto en el párrafo anterior, se optó por utilizar escalamiento sin balanceo para la red neuronal, mientras que Random Forest no necesita escalamiento porque la diferencia en la magnitud de las variables no afecta su entrenamiento debido a que en cada nodo prueba de a una variable.

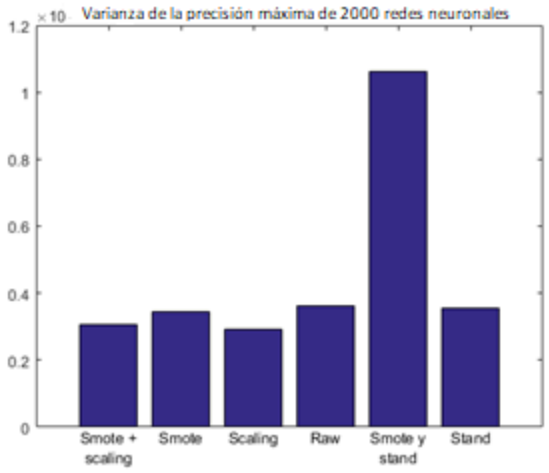

Figura 8-6 Comparación de la varianza de la precisión y puntaje máximo

Al aplicarse SMOTE sobre los datos y utilizar random forest, los resultados no variaron, lo que es de esperarse dada la naturaleza del algoritmo.

Anexo 3. RESULTADOS CON REDES NEURONALES

Anexo 3.1 GENERALIDADES

Las redes neuronales artificiales son una forma de aprendizaje y procesamiento automático inspirado en el funcionamiento del cerebro. Constituyen un modelo muy aproximado de la estructura paralela de este y tienen la capacidad de adquirir conocimiento del ambiente y almacenar el conocimiento adquirido en conexiones sinápticas.

Como se explica en [48] funcionan mediante la interconexión de varias neuronas, cuyo modelo elemental se muestra en la Figura 8-7, asumiendo que se está observando la neurona k dentro de una red. $\{x_1, x_2, \dots, x_n\}$ las entradas de la neurona, $\{w_{k1}, w_{k2}, \dots, w_{kn}\}$ los pesos sinápticos por los que se ponderan las entradas, Σ es un sumador lineal, b_k es un sesgo (*bias)* que se agrega a la suma ponderada, φ es una función de activación no lineal que limita la amplitud de la salida e y_k es la salida de la neurona.

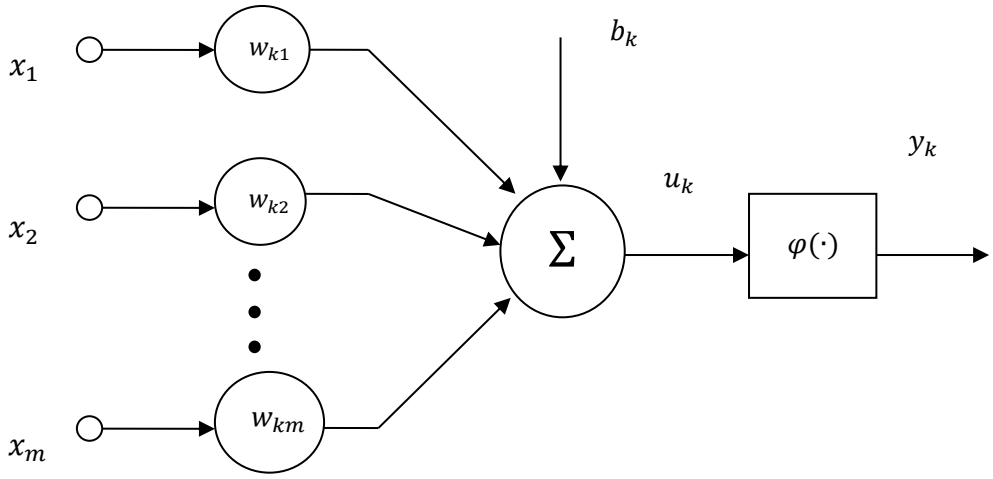

Figura 8-7 Modelo elemental de una neurona utilizada en una red neuronal artificial tipo MLP

Considerando lo anterior, la ecuación de una neurona queda siendo

$$u_k = b_k + \sum_{j=1}^{n} w_{kj} x_j$$

$$y_k = \varphi(u_k).$$

Para aplicar una red neuronal a un problema de clasificación es necesario entrenarla con los conjuntos ya mencionados en la sección 2.4. El algoritmo de aprendizaje toma las características del conjunto de entrenamiento como entradas y va

modificando los pesos w_{kj} comparando el valor conocido de salida obtenido con la clase conocida del conjunto de entrenamiento.

Existen varias posibilidades para elegir la función no lineal φ. Anteriormente se utilizaban funciones sigmoídeas, pero hoy en día lo más común es utilizar rectificadores, principalmente un rectificador lineal (RELU), que se define como:

$$f(x) = \max(0, x)$$

Las Redes Neuronales pueden ser de una o varias capas, las que a su vez poseerán un número variable de neuronas, las capas entre las entradas y las salidas son denominadas capas ocultas. Agregar estas capas puede darle la capacidad de resolver problemas más complejos, pero a su vez hace que requiera más datos para el entrenamiento. Las redes que poseen capas ocultas se denominan Perceptrones Múltiples (MLP). En la Figura 8-8 e Figura 8-9 se muestran representaciones de redes sin y con capa oculta respectivamente.

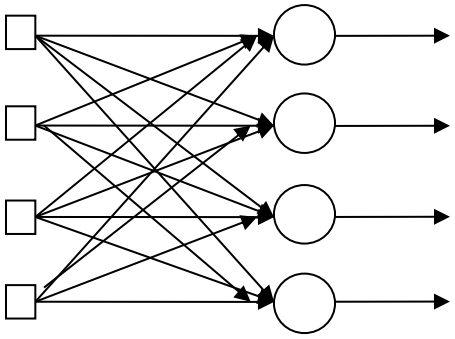

Figura 8-8 Diagrama de una red neuronal artificial sin capa oculta

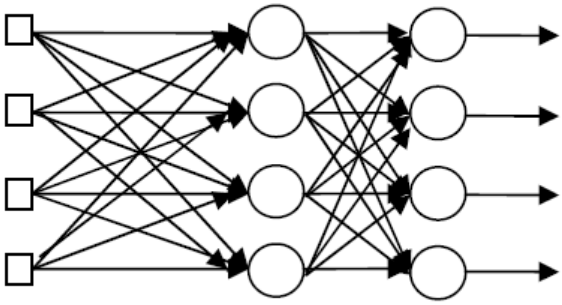

Figura 8-9 Diagrama de una red neuronal artificial con capa oculta

Anexo 3.2 NORMALIZACIÓN

Si las variables tienen rangos muy dispares puede traer problemas en el entrenamiento de una red neuronal, haciéndolo más ineficiente. Si el máximo de una entrada es miles de veces más grande que el de otra, afectará el cambio en los pesos en mayor medida, por lo que es recomendable normalizar las variables antes de utilizarlas en la red neuronal. Para esto existen varios métodos, pero los probados en este trabajo fueron:

- Estandarización: esta técnica busca que los datos tengan media 0 y varianza unitaria, para lo cual se realiza la siguiente operación:

$$x_i = \frac{z_i - \bar{z}_i}{\sigma_i}$$

  Donde $\bar{z}_i$ es el promedio de la variable y σ_i su varianza.

- Re-escalamiento, pero el más recomendable para una red neuronal es el min-máx. [49]. Para lograr esto se utiliza la siguiente ecuación:

$$x_i = \lambda_1 + \frac{(\lambda_2 - \lambda_1)\left(z_i - z_i^{min}\right)}{z_i^{max} - z_i^{min}},$$

donde x_i es el valor normalizado de z_i, z_i^{min} es el valor mínimo de la variable, z_i^{max} el máximo y λ_1 junto a λ_2 conforman el rango al que se quiere llegar; en este caso 0 y 1 respectivamente.

Anexo 3.3 MATRICES DE CONFUSIÓN DE LAS REDES NEURONALES

Anexo 3.3.1 INICIO DE SEMESTRE

En la Tabla 8-1 se muestran los resultados de la red neuronal entrenada con los datos disponibles al inicio del semestre y utilizando un umbral que maximice la precisión. Se ve que la especificidad del modelo es casi del 100%, fallando solo en una muestra, pero esto trae consigo que la sensibilidad sea 0.1% y con ello que la precisión sea apenas un 69,4%.

Tabla 8-1 Matriz de confusión de la red neuronal del inicio del semestre maximizando precisión

Real	**Predicción**

	Retención	Deserción	
Retención	2229	1	100.0%
Deserción	980	1	0.1%
			69.4%

La Tabla 8-2 muestra los resultados del mismo modelo al utiliza un umbral que maximice que minimice los costos. En este caso se aprecia que el modelo tiene una sensibilidad muy alta de 95,5%, pero su especificidad es de 17.3% y su precisión es de 46,1%.

Tabla 8-2 Matriz de confusión de la red neuronal del inicio del semestre minimizando los costos

		Predicción		
		Retención	Deserción	
Real	Retención	291	1393	17.3%
	Deserción	44	937	95.5%
				46.1%

Anexo 3.3.2 CÁTEDRA 1

En la Tabla 8-3 se muestran los resultados obtenidos por la red neuronal entrenada con los datos disponibles luego de la cátedra 1 y utilizando un umbral que maximice la precisión. Se ve que la especificidad alcanza un 89,7%, la sensibilidad un 47,3% y la precisión un 75,8%.

Tabla 8-3 Matriz de confusión de la red neuronal de la cátedra 1 maximizando precisión

		Predicción		
		Retención	Deserción	
Real	Retención	2025	232	89.7%
	Deserción	578	518	47.3%
				75.8%

En la Tabla 8-4 está la matriz de confusión para la misma red neuronal, pero utilizando un umbral que minimice los costos. En este caso la especificidad baja a un 48,4%, mientras que la sensibilidad sube a 86,7% y la precisión baja a un 60,9%.

Tabla 8-4 Matriz de confusión de la red neuronal de la cátedra 1 minimizando los costos

		Predicción		
		Retención	Deserción	
Real	Retención	1092	1165	48.4%

Deserción	146	950	86.7%
			60.9%

Anexo 3.3.3 CÁTEDRA 2

La Tabla 8-5 muestra los resultados de la red neuronal entrenada con los datos disponibles luego de la cátedra 2, utilizando un umbral que maximice la precisión del modelo. En este caso la especificidad llega a un 93,2%, mientras que la sensibilidad a un 46,2% y la precisión a un 78,4%.

Tabla 8-5 Matriz de confusión de la red neuronal de la cátedra 2 maximizando precisión

		Predicción		
		Retención	Deserción	
Real	Retención	2107	154	93.2%
	Deserción	560	481	46.2%
				78.4%

Al considerar un umbral que minimice los costos obtenido al utilizar la función de costos calculada para el problema, se llegaron a los resultados mostrados en la Tabla 8-6. Se puede observar que la especificidad baja a un 53,8%, mientras que la sensibilidad sube a un 85,5% y la precisión llega a 63,8%.

Tabla 8-6 Matriz de confusión de la red neuronal de la cátedra 2 minimizando los costos

		Predicción		
		Retención	Deserción	
Real	Retención	1217	1044	53.8%
	Deserción	151	890	85.5%
				63.8%

Anexo 3.3.4 FIN DE SEMESTRE

En la Tabla 8-7 está la matriz de confusión para la red neuronal entrenada con los datos disponibles al final del semestre y utilizando un umbral que maximice la precisión. Se aprecia que la especificidad es de un 92,9%, la sensibilidad de un 54,1% y la precisión de un 81,3%.

Tabla 8-7 Matriz de confusión de la red neuronal del final del semestre maximizando precisión

		Predicción		
		Retención	Deserción	
Real	Retención	2137	164	92.9%
	Deserción	449	529	54.1%
				81.3%

Cuando se toma el mismo modelo anterior, pero se le aplica un umbral que minimice los costos entregado al aplicar la función de costos, se llegan a los resultados mostrados en la Tabla 4-14. Aquí se ve que la especificidad llega a un 64,4%, la sensibilidad a un 85,7% y la precisión a un 70,8%.

Tabla 8- Matriz de confusión de la red neuronal del final del semestre minimizando los costos

		Predicción		
		Retención	Deserción	
Real	Retención	1482	819	64.4%
	Deserción	140	838	85.7%
				70.8%

Anexo 4. ENTRENAMIENTO CON RETROPROPAGACIÓN DEL ERROR

Este es uno de los algoritmos que existe para entrenar una red neuronal, que sirve para ajustar los pesos w_{ij} ya explicados, de modo que la red se ajuste a los datos, haciendo que las salidas de la red tomen valores cercanos a los deseados. Por ende, con un entrenamiento supervisado se va iterando sobre la red, ajustando los pesos cada vez más.

El algoritmo comienza inicializando los pesos aleatoriamente, para luego calcular la salida de la red con los vectores de características de cada uno de los elementos del conjunto de entrenamiento. Como las etiquetas de este conjunto son conocidas, se registra la diferencia entre el resultado de la RNA y el valor real, la que se denomina error. Como el objetivo es que esta diferencia sea lo más pequeña posible, lo que se hace es resolver el problema de optimización, donde las variables del problema son los valores de los pesos de la red. El algoritmo es una forma de resolver este problema.

Para explicar el algoritmo se define como n_o la cantidad de neuronas de entrada, n_s la cantidad de neuronas de salida, t el número de iteración del algoritmo y $\overrightarrow{x(t)}$ un vector de características de dimensión n_o que está asociado a un vector de etiquetas $\vec{d}(t)$ de dimensión n_c. Al alimentar la RNA con $\overrightarrow{x(t)}$ se producirá una

salida $\overrightarrow{y^c}(t)$, donde cada uno de sus componentes corresponden a las n_c salidas de la red. Por tanto, cada neurona de salida k se tendrá una salida $y_k^c(t)$ y un objetivo $d_k(t)$, donde estos son componentes de los vectores $\overrightarrow{y^c}(t)$ y $\vec{d}$(t) respectivamente. Con esto se puede calcular el error para cada neurona k de salida:

$$\begin{aligned}\varepsilon_k^2(t) &= \left(d_k(t) - y_k^c(t)\right)^2 \\ &= \left(d_k(t) - \varphi\left(s_k^c(t)\right)\right)^2 \\ &= \left(d_k(t) - \varphi\left(\sum_{j=0}^{n_c-1} w_{kj}^c(t)\cdot y_j^{c-1}(t)\right)\right)^2\end{aligned}$$

El error total por minimizar sería la suma de todos estos errores:

$$\varepsilon_T(t) = \sum_{k=1}^{n_c} \varepsilon_k^2(t)$$

Este se puede minimizar mediante el algoritmo del gradiente descendiente tomando como variables de minimización todos los pesos $w_{kj}^l(t)$ de todas las capas de la red. Para esto se parte por obtener las derivadas parciales del error con respecto a las neuronas que están en la capa de salida:

$$\frac{\partial \varepsilon_T(t)}{\partial w_{kj}^c(t)} = \frac{\partial \varepsilon_k^2(t)}{\partial w_{kj}^c(t)} = 2\varepsilon_k(t)\frac{\partial \varepsilon_k(t)}{\partial w_{kj}^c(t)} = 2\varepsilon_k(t)\left\{-\varphi'\left(s_k^c(t)\right)\cdot y_j^{c-1}(t)\right\}$$

Con esto los valores de los pesos se actualizan según las siguientes ecuaciones:

$$\Delta w_{kj}^c(t) = -\mu\frac{\partial \varepsilon_k^2(t)}{\partial w_{kj}^c(t)} = 2\mu\varepsilon_k(t)\cdot\varphi'\left(s_k^c(t)\right)\cdot y_j^{c-1}(t),$$

$$w_{kj}^c(t+1) = w_{kj}^c(t) + \Delta w_{kj}^c(t) = w_{kj}^c(t) + 2\mu\varepsilon_k(t)\cdot\varphi'\left(s_k^c(t)\right)\cdot y_j^{c-1}(t).$$

El parámetro μ que aparece en las dos ecuaciones anteriores corresponde a la tasa de aprendizaje de la red. Su función es controlar la estabilidad y velocidad de la convergencia del algoritmo. Al aumentarse el proceso de entrenamiento se vuelve más rápido puesto que las variaciones son más grandes, pero podría no converger si se eleva demasiado.

Las ecuaciones anteriores son sólo para las neuronas de salida, por lo que es necesario extender este cálculo para cualquier neurona de la red. Para esto se define

$$\delta_k^c(t) = 2\mu\varepsilon_k(t)\cdot\varphi'\left(s_k^c(t)\right)$$

Reemplazando en la primera ecuación para actualizar los pesos se tiene que

$$\Delta w_{kj}^{c}(t) = \mu \delta_{k}^{c}(t) \cdot y_{j}^{c-1}(t),$$

lo que se conoce como la regla delta. Cabe notar que

$$\delta_{k}^{c}(t) = -\frac{\partial \varepsilon_{T}(t)}{\partial w_{kj}^{c}(t)},$$

con lo que se puede extender la regla a las capas ocultas y a la de entrada. Sea la neurona l, se tiene que:

$$\Delta w_{kj}^{l}(t) = \mu \delta_{k}^{l}(t) \cdot y_{j}^{l-1}(t)$$

y

$$\delta_{k}^{l} = \frac{\partial \varepsilon_{T}(t)}{\partial s_{k}^{l}(t)} = -\sum_{j=1}^{n_{l+1}} \frac{\partial \varepsilon_{T}(t)}{\partial s_{j}^{l+1}(t)} \cdot \frac{\partial s_{j}^{l+1}(t)}{\partial s_{k}^{l}(t)}$$

Notando que

$$\frac{\partial \varepsilon_{T}(t)}{\partial s_{j}^{l+1}(t)} = \delta_{j}^{l+1}(t)$$

Y

$$\frac{\partial s_{j}^{l+1}(t)}{\partial s_{k}^{l}(t)} = \frac{\partial \left(\sum_{p=1}^{n_{l}} w_{jp}^{l+1}(t) \varphi \left(s_{p}^{l}(t) \right) \right)}{\partial s_{k}^{l}(t)} = w_{jk}^{l+1}(t) \varphi' \left(s_{k}^{l}(t) \right)$$

Con lo que finalmente se obtiene que el δ de cualquier neurona entre las capas 1 y $c-1$ se puede expresar como

$$\delta_{k}^{l}(t) = -\varphi' \left(s_{k}^{l}(t) \right) \sum_{j=1}^{n_{l+1}} \delta_{j}^{l+1}(t) w_{jk}^{l+1}(t).$$

En resumen, la regla de actualización para cada peso w_{jk}^{l} sólo depende del parámetro μ que es fijado por el usuario; la salida de la neurona de la capa anterior $y_{j}^{l-1}(t)$, que es conocido; y del valor de $\delta_{k}^{l}(t)$, que debe ser calculado. Cabe destacar que para obtener este último valor es necesario conocer todos los δ_{j}^{l+1} de la capa siguiente, por lo que es una ecuación de recurrencia donde el caso base es el da la capa de salida. Como el algoritmo parte de la última capa hacia la entrada utilizando el error sobre el conjunto de entrenamiento el algoritmo recibe el nombre

de retropropagación del error. Además, todo lo anterior depende de la función no lineal escogida [50].